



Inteligência Artificial na Saúde Obstétrica: Transformando o ObsCare com Modelos de Linguagem de Grande Escala

ÂNGELA PATRÍCIA ARAÚJO CARVALHO
julho de 2026

Inteligência Artificial na Saúde Obstétrica: Transformando o ObsCare com Modelos de Linguagem de Grande Escala

Ângela Patrícia Araújo Carvalho

**Dissertação para obtenção do Grau de Mestre em
Engenharia Biomédica**

**Orientadora: Maria Goreti Carvalho Marreiros, Professora Coordenadora Principal do
Instituto Superior de Engenharia do Porto do Instituto Politécnico do Porto**

**Supervisor: Vítor Rafael Palmeiro Crista, Assistente Convidado do Instituto Superior
de Engenharia do Porto do Instituto Politécnico do Porto**

Júri:

Presidente:

Joaquim Fernando Almeida Alves, Professor Adjunto do Instituto Superior de Engenharia do
Porto do Instituto Politécnico do Porto

Arguente:

Júlio César Botelho de Souza, Investigador Auxiliar do Instituto Superior de Engenharia do Porto
do Instituto Politécnico do Porto

Orientadora:

Maria Goreti Carvalho Marreiros, Professora Coordenadora Principal do Instituto Superior de
Engenharia do Porto do Instituto Politécnico do Porto

Resumo

A literatura científica na área da obstetrícia encontra-se em constante evolução. Consequentemente, os profissionais de saúde enfrentam o desafio de acompanhar e integrar continuamente as evidências científicas mais recentes na prática clínica. Neste contexto, torna-se essencial dispor de ferramentas que facilitem o acesso à informação científica, garantindo que esta seja apresentada de forma rápida, fiável e atualizada. Os avanços na área da Inteligência Artificial, em particular os Modelos de Linguagem de Grande Escala (LLMs), demonstram capacidade para melhorar o acesso, a pesquisa e a interpretação da literatura científica.

Esta tese explora a aplicação de LLMs na área da obstetrícia através do desenvolvimento de um *chatbot* destinado a apoiar os médicos no acesso à informação clínica atualizada. O sistema foi concebido para compreender questões formuladas em linguagem natural, pesquisar informação em fontes científicas credíveis e apresentar respostas fundamentadas, facilitando a atualização contínua dos profissionais de saúde. Para além disso, este trabalho avalia diferentes estratégias de extração e segmentação da informação presente em artigos científicos, como forma de otimizar o processo de recuperação de informação e, consequentemente, melhorar a qualidade, a relevância e a fundamentação das respostas geradas.

O trabalho inclui o desenvolvimento da arquitetura do sistema, composta por três módulos dedicados à extração e ao processamento dos diferentes tipos de informação presentes em artigos científicos, nomeadamente texto, tabelas e imagens. Esta abordagem permite responder aos desafios associados a cada tipo de conteúdo, recorrendo a técnicas de processamento que se adequam aos diferentes tipos de dados.

Para avaliar o desempenho do sistema, foi realizado um estudo comparativo entre diferentes abordagens de recuperação de informação e diferentes LLMs. A avaliação da recuperação de informação clínica recorreu a métricas quantitativas, enquanto a avaliação das respostas geradas recorreu à metodologia *LLM-as-a-Judge*. Os resultados obtidos sugerem que a solução proposta apresenta um desempenho promissor, alcançando valores de 0,901 para a relevância do contexto recuperado, 0,960 para a fidelidade das respostas à informação recuperada e 0,922 para a relevância das respostas, utilizando o modelo Qwen 3 32B.

Desta forma, este trabalho contribui para o desenvolvimento de *chatbots* para apoiar os profissionais de saúde que trabalham na área da obstetrícia, propondo uma arquitetura para recuperação e consulta de literatura científica que facilita o acesso a informação clínica relevante e atualizada.

Palavras-chave: Apoio à Decisão Clínica, Chatbot, Inteligência Artificial, Modelos de Linguagem de Grande Escala, Obstetrícia, Processamento de Linguagem Natural.

Abstract

Scientific literature in the field of obstetrics is continuously evolving. Consequently, healthcare professionals face the ongoing challenge of keeping up with and incorporating the latest scientific evidence into clinical practice. In this context, there is a growing need for tools that facilitate access to scientific information while ensuring that it is delivered in a timely, reliable, and up-to-date manner. Recent advances in Artificial Intelligence, particularly Large Language Models (LLMs), have demonstrated significant potential to improve the access, retrieval, and interpretation of scientific literature.

This thesis explores the application of LLMs in the field of obstetrics through the development of a chatbot designed to support physicians in accessing up-to-date clinical information. The proposed system was designed to understand natural language queries, retrieve information from reliable scientific sources, and generate evidence-based responses, thereby supporting the continuous updating of medical knowledge. Furthermore, this work investigates different strategies for extracting and segmenting information from scientific articles in order to optimize the information retrieval process and, consequently, improve the quality, relevance, and grounding of the generated responses.

The proposed solution includes a system architecture composed of three modules dedicated to extracting and processing the different types of information contained in scientific articles, namely text, tables, and images. This modular approach addresses the specific challenges associated with each content type by applying processing techniques tailored to their individual characteristics.

To evaluate the proposed system, a comparative study was conducted using different information retrieval approaches and different LLMs. Information retrieval performance was assessed using quantitative metrics, while the quality of the generated responses was evaluated using the LLM-as-a-Judge methodology. The results suggest that the proposed solution achieves promising performance, obtaining scores of 0.901 for context relevance, 0.960 for response faithfulness, and 0.922 for response relevance when using the Qwen3-32B model.

This work contributes to the development of LLM-based chatbots for supporting healthcare professionals in the field of obstetrics by proposing an architecture for scientific literature retrieval and consultation that facilitates access to relevant and up-to-date clinical information.

Keywords: Artificial Intelligence, Chatbot, Clinical Decision Support, Large Language Models, Natural Language Processing, Obstetrics.

Agradecimentos

Gostaria de expressar o meu sincero agradecimento a todos aqueles que, de forma direta ou indireta, contribuíram para a concretização desta dissertação, através do seu incentivo e disponibilidade para me ajudar nesta caminhada.

Em primeiro lugar, agradeço ao ISEP por toda a formação e pelas oportunidades proporcionadas durante o meu percurso académico. Agradeço igualmente ao GECAD pela oportunidade de estágio, pela forma como fui recebida e integrada na equipa e por todo o apoio prestado.

Deixo também um agradecimento especial à minha orientadora, Professora Goreti Marreiros, e ao meu supervisor, Vítor Crista, pela orientação prestada, pelos valiosos conselhos e pelo acompanhamento ao longo de todo o desenvolvimento deste trabalho.

À minha família e amigos, agradeço por terem sido um suporte constante durante esta etapa da minha vida.

A todos, o meu mais sincero obrigado.

Índice

1	Introdução	1
1.1	Contexto e Motivação.....	1
1.1.1	Fundamentos da Obstetrícia na Saúde Materna	2
1.1.2	ObsCare	4
1.2	Problema	5
1.3	Objetivos	6
1.4	Estrutura da tese	7
2	Inteligência Artificial Generativa e Modelos de Linguagem	9
2.1	Processamento de Linguagem Natural	10
2.2	Arquitetura Transformer	12
2.3	Modelos de Linguagem de Grande Escala	13
2.4	Prompt Engineering	15
2.5	Retrieval-Augmented Generation	15
2.5.1	Componentes Básicos de RAG	16
2.5.2	Pré-processamento e document parsing de documentos científicos	16
2.5.3	Indexação de documentos	20
2.5.4	Recuperação de textos	21
2.5.5	Geração da Resposta	22
2.5.6	Diferentes abordagens de RAG	23
2.6	Métricas de Avaliação para Modelos de Embeddings e Modelos de Linguagem	24
2.6.1	LLM-as-a-Judge	26
2.6.2	Métricas específicas para RAG	27
3	Revisão de Literatura	29
3.1	Metodologia	29
3.1.1	Perguntas de Pesquisa.....	29
3.1.2	Definições de Estratégia de Pesquisa	31
3.2	Resultados	34
3.2.1	Questão 1: Quais sistemas ou soluções baseadas em tecnologias digitais têm sido aplicados na área da saúde, com o objetivo de apoiar a atualização do conhecimento médico e acelerar a análise de literatura científica?.....	34
3.2.2	Questão 2: “Quais tecnologias e técnicas de Inteligência Artificial têm sido aplicadas na análise de literatura científica em contextos clínicos?”.....	39

3.2.3	Questão 3: “Quais são os principais desafios identificados na aplicação dessas tecnologias, particularmente no contexto da saúde?”	45
3.3	Discussão de Resultados.....	47
3.3.1	Questão 1: Quais sistemas ou soluções baseadas em tecnologias digitais têm sido aplicados na área da saúde, com o objetivo de apoiar a atualização do conhecimento médico e acelerar a análise de literatura científica?.....	47
3.3.2	Questão 2: “Quais tecnologias e técnicas de Inteligência Artificial têm sido aplicadas na análise de literatura científica em contextos clínicos?”	48
3.3.3	Questão 3: “Quais são os principais desafios identificados na aplicação dessas tecnologias, particularmente no contexto da saúde?”	50
3.4	Conclusão	51
4	Solução Proposta	53
4.1	Arquitetura do Sistema.....	53
4.2	Módulo de processamento do texto	55
4.3	Módulo de processamento das imagens	56
4.4	Módulo de Processamento de Tabelas	57
5	Implementação	59
5.1	Ferramentas e tecnologias utilizadas	59
5.2	Modelos de Embeddings	60
5.3	Modelos de Linguagem de Grande Escala	61
5.3.1	Infraestrutura Computacional	61
5.3.2	Modelos utilizados	61
5.3.3	Modelos utilizados no Pré-Processamento	62
5.4	Pipeline de extração de texto.....	63
5.4.1	Etapa 1: Extrair título principal do Documento	64
5.4.2	Etapa 2: Extrair subtítulos do Documento	69
5.4.3	Etapa 3: Limpeza e preparação final das seções de texto.....	74
5.5	Pipeline de Extração de figuras dos documentos.....	75
5.5.1	Classificação do tipo de figura.....	76
5.5.2	Estratégias para descrição de figuras.....	77
5.6	Pipeline de Extração de Tabelas.....	80
5.7	Interface do Utilizador	82
6	Experimentação	86
6.1	Avaliação de ferramentas para o pré-processamento de documentos	86
6.1.1	Resultados	89
6.1.2	Discussão de Resultados	96
6.2	Avaliação de estratégias de descrição de figuras	98

6.2.1	Resultados:.....	100
6.3	Avaliação da classificação automática de figuras	101
6.3.1	Resultados.....	102
6.4	Avaliação dos modelos RAG	103
6.4.1	Avaliação dos modelos de embeddings	106
6.4.2	Modelos considerados na avaliação.....	108
6.4.3	Métricas de Avaliação	109
6.4.4	Desenvolvimento de prompts para os LLMs-as-a-Judge	109
6.4.5	Resultados da avaliação de modelos de embeddings	112
6.4.6	Resultados da Avaliação dos LLMs	115
6.4.7	Discussão de resultados	116
7	Conclusão	120
7.1	Objetivos da tese.....	121
7.2	Limitações e Trabalho Futuro	123
	Referências.....	125
	Anexo A: Resultados da descrição do Conteúdo de fluxogramas	137
	Anexos B: Resultados da comparação das duas estratégias de descrição de figuras 143	
	Anexo C: Lista dos artigos científicos e normas clínicas utilizadas na base de dados do chatbot	148
	Anexo D: Resultados obtidos na avaliação da recuperação de chunks e respetivos metadados	152

Lista de Figuras

Figura 1: Arquitetura do <i>Transformer</i> . Fonte [80]	13
Figura 2: Esquema de um sistema de RAG. Fonte: [98]	16
Figura 3: a) Arquitetura de <i>Naive RAG</i> . b) Arquitetura de <i>Graph RAG</i> . Fonte: [97].....	24
Figura 4: Fluxograma PRISMA do processo de identificação, seleção, elegibilidade e inclusão dos estudos.	33
Figura 5: Mapa semântico de artigos científicos da aplicação <i>COVID-19 Explorer</i> . Adaptado de: [136].	35
Figura 6: Gráfico de bolhas e lista dos artigos mais relevantes apresentados pela aplicação. Adaptado de: [139].....	36
Figura 7: (a) Grafo de conhecimento gerado com restrição à entidade de entrada, apresentando ligação a apenas uma entidade biomédica; (b) subgrafo de conhecimento construído com base na relação “ <i>associated with</i> ”. Adaptado de [122].	37
Figura 8: Agente Conversacional AlzheimerRAG. Adaptado de [148].	38
Figura 9: Arquitetura do sistema HoPE para processamento da pergunta do utilizador. Adaptado de [150].	45
Figura 10: Arquitetura geral do sistema proposto	54
Figura 11: Etapas do módulo de processamento do texto	56
Figura 12: Módulo de processamento das imagens	57
Figura 13: Módulo de Extração de Tabelas	58
Figura 14: Diagrama de processamento das secções de texto	64
Figura 15: Fluxograma para descrever a extração de imagens de documentos.....	76
Figura 16: Fluxograma relativo à extração de tabelas	81
Figura 17: Interface principal do sistema de <i>Chatbot</i> baseado em RAG.....	82
Figura 18: Interface de carregamento de documentos para ingestão na base de conhecimento	83
Figura 19: Interface do <i>chatbot</i> utilizada para interação com o sistema.....	84
Figura 20: Interface do sistema de <i>chatbot</i> demonstrando uma interação com o utilizador, incluindo a submissão de uma pergunta e a respetiva resposta gerada pelo sistema.....	84
Figura 21: a) Exemplo de referência do tipo texto; b) Exemplo de referência do tipo tabela;..	85
Figura 22: Exemplo de tabela listrada; a) Imagem de página do documento. Adaptado de: [191]; b) e c) Resultado da biblioteca PyMuPDF na extração da tabela, renderizado novamente para pdf;	90
Figura 23: Exemplo de imagem retirada de documento médico. a) Imagem de uma página do documento. Adaptado de [192]; b) Resultados da biblioteca PyMuPDF na extração de imagens; c) Figura retirada do documento, utilizando a biblioteca PyMuPDF	90
Figura 25: Exemplo de imagem retirada de documento médico. Adaptado de: [191]; a) Resultado do modelo de deteção de tabelas do TATR; b) Resultado do modelo de identificação de estrutura interna da tabela do TATR.....	91
Figura 24: Exemplo de tabela listrada; a) Imagem de página do documento. Adaptado de: [193]; b) Resultado do modelo de deteção de tabelas do TATR, com as tabelas identificadas a vermelho;	

c) Resultado do modelo de reconhecimento da estrutura interna de tabelas do TATR, com a estrutura identificada a verde.....	92
Figura 26: Exemplo de imagem de documento com tabela. Adaptado de [194] a) Resultados da biblioteca PaddleOCR na identificação de elementos narrativos; b) Resultados da biblioteca PaddleOCR para o reconhecimento da estrutura interna da tabela;	93
Figura 27: Exemplo de imagem de documento com duas colunas. Retirado de [195] e [196]; a) Resultado da biblioteca PaddleOCR na identificação de elementos narrativos. Retirado de [195]; b) Resultado da biblioteca PaddleOCR na identificação de elementos narrativos.....	93
Figura 28: Exemplo de imagem de documento com duas colunas. Retirado de [197]; a) Resultado do modelo PaddleOCR na numeração dos elementos narrativos identificados; b) Resultado do modelo PaddleOCR na identificação de elementos narrativos	94
Figura 29: Exemplo de uma imagem com tabela e dividido em duas colunas. a) Imagem de uma página do documento. Adaptado de [198]; b) e c) Resultado do modelo Nougat, convertido para html e renderizado novamente para pdf	95
Figura 30: Exemplo de uma imagem com vários títulos de diferentes formatos. a) Imagem de uma página do documento. Adaptado de: [199]; b) e c) Resultado do modelo Nougat, convertido para html e renderizado novamente para pdf;.....	95
Figura 31: Exemplo de uma figura de um fluxograma. Retirado do documento: [192]	99
Figura 32: Exemplos de imagens usadas para testar os modelos com figuras clínicas, retirada de [196].	99
Figura 33: Exemplo de imagem, com legenda, mas sem referências no texto. Retirado de: [192]	100
Figura 34: Exemplos de imagens utilizadas na avaliação dos modelos multimodais	102
Figura 35: Gráfico de Pontuações de Recall@10, para recuperação de texto (a), imagens (b) e tabelas (c).....	114
Figura 36: Comparação das métricas Recall e nDCG antes e após a aplicação de <i>Reranking</i> .	115
Figura 37: Resultados da avaliação dos LLMs	116

Lista de Tabelas

Tabela 1: Bases de Dados Online	31
Tabela 2: Termos de Pesquisa.....	31
Tabela 3: Critérios de inclusão de artigos	32
Tabela 4: Critérios de exclusão de artigos	32
Tabela 5: Modelos de <i>Embeddings</i>	61
Tabela 6: Configurações dos LLMs	62
Tabela 7: Exemplos de aplicação do <i>prompt</i> de validação de títulos.	68
Tabela 8: Exemplo de classificação hierárquica obtida pela abordagem baseada exclusivamente na informação textual.....	70
Tabela 9: Comparação dos níveis hierárquicos atribuídos pelos modelos LLaMA 3.1 8B e LLaMA 3.3 70B na classificação de subtítulos.....	73
Tabela 10: Ferramentas de <i>parsing</i> e extração de informação em documentos científicos.....	89
Tabela 11: Avaliação das Bibliotecas para estruturação de Tabelas	96
Tabela 12: Avaliação das Bibliotecas para extração de títulos e secções de texto.....	97
Tabela 13: Avaliação das bibliotecas para extração de imagens	97
Tabela 14: Métricas de avaliação do desempenho dos modelos de <i>embeddings</i>	109
Tabela 15: Pontuações de Recall@10 e nDCG@10 para os modelos de <i>embeddings</i> MiniLM, MpNet, MedEmbed e PubMedBERT e a combinação destes modelos com o modelo BM25.	113

Lista de Códigos

Código 1: Estrutura do <i>prompt</i> para a criação dos resumos parciais do documento.....	65
Código 2: Estrutura do <i>prompt</i> para criar o resumo global do documento	66
Código 3: Estrutura do <i>prompt</i> para a validação do título principal do documento	67
Código 4: Estrutura do <i>prompt</i> aplicado ao modelo de linguagem, para classificar os subtítulos por hierarquias.....	69
Código 5: Exemplo de segmentos textuais com informação tipográfica extraída com PyMuPDF	71
Código 6: Estrutura do <i>prompt</i> para a definição dos níveis hierárquicos dos subtítulos	73
Código 7: Estrutura do <i>prompt</i> aplicado ao modelo multimodal para descrever as figuras	77
Código 8: Estrutura do <i>prompt</i> aplicado ao modelo multimodal para descrever as figuras.	77
Código 9: Estrutura do <i>prompt</i> modelo multimodal LLaMA 4 Scout 17B 16E para descrever imagens	78
Código 10: Estrutura do <i>prompt</i> aplicado ao modelo multimodal para descrever as figuras ...	80
Código 11: Estrutura do <i>prompt</i> para o LLM LLaMA 3.3 70B encontrar notas em rodapé relativas à tabela.....	81
Código 12: Resultados obtidos com a utilização do LLM para a recuperação das notas de rodapé das tabelas.	82
Código 13: Estrutura do <i>prompt</i> aplicado ao modelo multimodal para classificar automaticamente as figuras	102
Código 14: Estrutura do <i>prompt</i> para o <i>LLM-as-a-Judge</i> , para avaliar a relevância do contexto	110
Código 15: Estrutura do <i>prompt</i> para o <i>LLM-as-a-Judge</i> , para avaliar a confiança na resposta gerada pelo modelo	111
Código 16: Estrutura do <i>prompt</i> para o <i>LLM-as-a-Judge</i> , para avaliar a relevância da resposta relativamente à pergunta	112

Acrónimos

Lista de Acrónimos

AI	<i>Artificial Intelligence</i>
BM25	<i>Best Matching 25</i>
DL	<i>Deep Learning</i>
LLM	<i>Large Language Models</i>
ML	<i>Machine Learning</i>
MLLM	<i>Multimodal Large Language Model</i>
nDCG	<i>Normalized Discounted Cumulative Gain</i>
NER	<i>Named Entity Recognition</i>
NLP	<i>Natural Language Processing</i>
OCR	<i>Optical Character Recognition</i>
RAG	<i>Retrieval-Augmented Generation</i>
TF-IDF	<i>Term Frequency–Inverse Document Frequency</i>
VCE	Versão Cefálica Externa

1 Introdução

Neste capítulo encontra-se o enquadramento e a motivação para o desenvolvimento da presente tese. Inicialmente, é realizada uma contextualização da área da Obstetrícia no âmbito da saúde materna em Portugal, seguida da identificação dos principais desafios relacionados com a atualização contínua dos profissionais de saúde nesta área. Por fim, são definidos os objetivos do presente trabalho.

1.1 Contexto e Motivação

Os avanços mais recentes na área da Inteligência Artificial (AI), em especial no desenvolvimento de Modelos de Linguagem de Grande Escala (LLMs), demonstram uma grande capacidade para responder a desafios que persistem no setor da saúde. Entre esses desafios, destaca-se a dificuldade de acompanhar os novos estudos clínicos publicados em bases de dados científicas e revistas médicas especializadas [1], [2], [3].

Se, por um lado, a crescente procura por artigos científicos reflete um progresso significativo na área da medicina, marcado pelo aumento do acesso a informações novas sobre doenças, diagnósticos e terapias. Por outro lado, o volume excessivo de informação, aliado à falta de tempo e ao excesso de trabalho dos profissionais de saúde, impossibilita que estes se mantenham a par da informação clínica mais recente [4], [5].

A título de exemplo, a presente tese desenvolve-se no âmbito da Obstetrícia. Como área central da saúde materna e perinatal, desempenha um papel fundamental na saúde e no bem-estar da mulher e do recém-nascido. Trata-se, contudo, de uma especialidade também afetada pela sobrecarga de informação clínica [6].

Para situar esta tese, começa-se com a seção “Fundamentos da Obstetrícia na Saúde Materna”, na qual se enquadra a especialidade, e se explica o papel do obstetra na vigilância clínica e no

apoio à grávida. Nesta secção apresenta-se também um panorama global e nacional dos principais desafios na prática obstétrica, nomeadamente o aumento da complexidade desta área e a lacuna entre o conhecimento científico e a prática clínica. Termina com uma discussão das necessidades da atualização contínua dos profissionais na era digital, como condição para integrar a evidência científica na prática clínica e melhorar os desfechos.

1.1.1 Fundamentos da Obstetrícia na Saúde Materna

A obstetrícia é a área da medicina que cuida da saúde da grávida e do feto durante todas as fases da gestação. O médico especialista em obstetrícia acompanha a gestante desde a primeira consulta de pré-natal até ao puerpério, garantindo que a grávida tenha uma recuperação segura.

Ao longo da gestação, o obstetra é responsável pela vigilância da saúde materno-fetal, através de consultas periódicas, avaliação clínica e monitorização ecográfica e laboratorial, de forma a identificar precocemente eventuais complicações e assegurar um crescimento e desenvolvimento fetal saudável [7], [8]. Esta vigilância inclui a classificação do risco obstétrico, a deteção de patologias maternas emergentes ou pré-existentes e a implementação atempada de medidas preventivas e terapêuticas [9], [10].

Para além disso, o acompanhamento pré-natal integra uma vertente educativa e psicossocial. O médico obstetra precisa esclarecer as dúvidas da grávida, promover comportamentos saudáveis e ajudar na tomada de decisões informadas. Considerando que a gravidez é um período de profundas transformações físicas, hormonais e emocionais, um acompanhamento próximo e individualizado ajuda a diminuir a ansiedade materna e contribui para uma gestação mais segura e positiva [11], [12], [13].

Contudo, ao longo da última década, a prestação de cuidados obstétricos tem-se tornado cada vez mais exigente e complexa [14], [15]. Esta mudança é resultado do aumento da prevalência de obesidade e de doenças crónicas na população em idade reprodutiva, incluindo hipertensão arterial, diabetes, doenças autoimunes e patologias cardiovasculares. Para além disso, observa-se um adiamento da maternidade para idades mais avançadas, aumentando a possibilidade de risco obstétrico e de aparecimento de comorbidades pré-existentes [15], [16], [17].

Durante a gravidez, ocorrem diversas adaptações fisiológicas, como alterações hemodinâmicas, respiratórias e imunológicas, que podem agravar doenças crónicas pré-existentes, modificar a sua expressão clínica e interferir com a resposta terapêutica [18], [19], [14]. Estas modificações podem piorar patologias pré-existentes e aumentar o risco de complicações obstétricas, como pré-eclâmpsia, parto prematuro, restrição do crescimento fetal, malformações congénitas e, nos casos mais graves, mortalidade perinatal [20],[21],[22].

Face a esta realidade, é importante a atualização contínua dos profissionais de saúde. A produção de artigos científicos na área da Obstetrícia tem crescido exponencialmente nas últimas décadas, impulsionada por avanços na investigação da fisiopatologia materno-fetal, do diagnóstico pré-natal, da terapêutica farmacológica na gravidez e das estratégias de prevenção

de complicações obstétricas. Num contexto em que as gestações tornam-se progressivamente mais complexas, integrar a evidência científica mais atualizada na prática clínica é fundamental para haver uma prestação de cuidados seguros e de qualidade [23], [24].

Este avanço científico tem permitido uma compreensão mais profunda das principais patologias relacionadas com a gestação, contribuindo para a identificação precoce de grupos de risco e para a implementação de novas medidas preventivas. Contudo, o facto de existirem estudos e trabalhos publicados com novas intervenções clínicas não garante, por si só, a sua implementação na prática clínica [25], [26]. Existe um desfasamento entre as recomendações baseadas na evidência científica e a realidade praticada nos hospitais e maternidades atualmente [25], [26].

Esta discrepância também aparece em casos em que existem intervenções preventivas amplamente validadas, mas que não são adotadas na prática clínica. Um exemplo é a utilização de ácido acetilsalicílico em baixas doses para prevenção de pré-eclâmpsia em gestantes com fatores de risco para síndromes hipertensivas. Apesar de ser uma medida comprovadamente eficaz, capaz de reduzir em mais de 60% os casos da doença, o estudo realizado por Tavares *et al.* [27] demonstrou que apenas 16% das pacientes receberam esta recomendação durante a gestação. Este resultado poderá estar associado à falta de atualização científica dos médicos.

Outro exemplo é o facto de a evidência científica atual e as *guidelines* internacionais [28] sustentarem que, durante a gravidez, as pacientes devem continuar a tomar corticosteroides para evitar exacerbações de asma e complicações obstétricas. No entanto, dados provenientes de um estudo observacional italiano, evidenciam que uma proporção significativa de médicos continua a descontinuar este tratamento por terem receio de efeitos adversos. Este contraste expõe uma lacuna entre o conhecimento científico consolidado e a prática clínica real, com potenciais repercussões negativas no controlo da doença e nos desfechos materno-fetais [29].

Perante este desfasamento entre a evidência científica e a prática clínica, torna-se evidente que a tomada de decisão terapêutica na gravidez exige conhecimento atualizado. No entanto, estudos relatam as dificuldades dos médicos em aceder rapidamente à informação [30] e em prescrever medicamentos durante a gravidez, devido à escassez de dados relativos à segurança dos fármacos [31].

Entre os diversos recursos disponíveis para colmatar estas limitações, a literatura de artigos científicos mantém-se como um dos principais instrumentos de atualização dos profissionais de saúde. Destacam-se também as normas clínicas, que consistem em recomendações sistematicamente desenvolvidas com o objetivo de apoiar a tomada de decisão de profissionais em circunstâncias clínicas específicas, com base na melhor evidência disponível [32]. As revistas médicas com revisão de pares são também uma fonte essencial, pois garantem um controlo mínimo de qualidade científica [33]. Para além disso, há as revisões sistemáticas e meta-análises disponíveis em bases de dados, como a PubMed, que resumem grandes volumes de informação com rigor metodológico [34], [35].

Face às dificuldades previamente descritas, a tecnologia vem para transformar a forma como os profissionais de saúde acedem e aplicam o conhecimento científico. Neste contexto, a utilização de aplicações móveis constitui um exemplo concreto dessa evolução. Atualmente, destacam-se algumas que são usadas por médicos e estudantes de medicina como, por exemplo, a UpToDate [36], a DynaMed [37] e a Epocrates [38]. Estas ferramentas oferecem várias vantagens, nomeadamente o acesso rápido à informação atualizada, o apoio ao diagnóstico e a verificação das doses de medicação administradas aos pacientes [39], [40].

Outra forma de aprendizagem é a troca de informações entre médicos, facilitada por várias plataformas online como Youtube e Facebook [41], [42]. Um exemplo relevante é a SERMO [43], uma rede social anónima e privada, exclusiva para médicos nos EUA. Nesta plataforma, os profissionais de saúde partilham experiências, discutem casos clínicos e recebem conselhos de colegas da mesma área. No entanto, surgem questões éticas relativas à privacidade e à confidencialidade dos pacientes, especialmente quando os médicos partilham casos clínicos e imagens de exames [44]. Outro exemplo é o OrthoMind [45], uma plataforma social direcionada especialmente para cirurgiões ortopédicos. Esta ferramenta oferece diversas vantagens, como a criação de comunidades entre cirurgiões ortopédicos de diferentes regiões, para a troca de ideias e conhecimentos [45].

Desta forma, a tecnologia é extremamente importante para a atualização contínua dos médicos, pois facilita o acesso à informação científica mais relevante e atualizada. Em muitos casos, é um apoio importante ao processo de diagnóstico dos pacientes. Para além disso, permite melhorar a comunicação entre profissionais de saúde de diferentes regiões, através da partilha de conhecimento e experiência [46].

1.1.2 ObsCare

Face a este enquadramento, a ObsCare contribui para a integração da tecnologia na área da Obstetrícia em Portugal. O *software* foi criado em 2003 para dar resposta à necessidade de um registo de saúde eletrónico específico para obstetrícia e está atualmente em utilização nos serviços de obstetrícia de nove hospitais portugueses. Este é um sistema desenvolvido pela VirtualCare, para ser utilizado por profissionais de saúde, onde se regista toda a jornada das grávidas desde a consulta ao puerpério, cobrindo o internamento, urgência, sala de partos, berçário e consultas.

Este sistema foi concebido para operar com sistemas de diferentes hospitais, permitindo a integração e partilha de informação clínica proveniente de diferentes fontes, nomeadamente cardiotocografia, ecografias e resultados de análises laboratoriais. Atualmente, a ObsCare encontra-se amplamente disseminada no Serviço Nacional de Saúde, estando implementada em 20 hospitais públicos [47], [48].

1.2 Problema

Atualmente, os médicos obstetras enfrentam vários desafios para acompanhar a evolução científica na área da Medicina. O facto de serem constantemente publicadas novas recomendações e atualizações clínicas, impossibilita que os médicos se mantenham a par da informação mais relevante [49], [50]. Existem, no entanto, várias tecnologias que procuram solucionar este problema, como por exemplo, a aplicação de filtros nas bases de dados online, como a PubMed, ou as aplicações com resumos de artigos científicos, como a UpToDate [51]. Embora estas ferramentas facilitem o acesso à informação, podem apresentar desafios relacionados com o tempo necessário para pesquisa e a seleção da evidência mais adequada.

Face a este problema, começaram a surgir soluções que utilizam a AI generativa, como os LLMs, para facilitar e acelerar o acesso à informação que os profissionais de saúde procuram [52]. Estes modelos, baseados em técnicas de Processamento de Linguagem Natural (NLP), são capazes de processar grandes volumes de texto, compreender o contexto da informação e produzir respostas coerentes e ajustadas ao conteúdo da questão, em tempo real. A sua aplicação através de *chatbots* clínicos pode ajudar os médicos a aceder à informação de forma mais fácil, sugerindo hipóteses diagnósticas e auxiliando na gestão de condições específicas, funcionando como um complemento, e não substituto, do raciocínio clínico [53], [54].

Mas estas soluções tecnológicas podem ter outras vantagens, como por exemplo, ajudar a melhorar a comunicação com os pacientes, nomeadamente através da simplificação de terminologia complexa, tradução de conceitos médicos e adaptação da informação ao nível de literacia em saúde do paciente [55].

Existem vários exemplos do ChatGPT, o *chatbot* da OpenAI, a ajudar os médicos a acederem de forma rápida e fácil às informações mais recentes da área da obstetrícia [56],[57]. Quando são considerados os dados clínicos da paciente, o ChatGPT também permite que os médicos recebam sugestões personalizadas, melhorando a experiência do médico e contribuindo para a sua tomada de decisão [58].

Lee *et al.* [57] destacam o potencial de utilizar AI para o diagnóstico e apoio clínico, especialmente para a formulação de diagnósticos diferenciais e para orientar as equipas médicas em condições raras. Enquanto Allahqoli *et al.*[59] demonstraram o desempenho do ChatGPT, num estudo com 30 casos clínicos retirados da obra “*100 Cases in Obstetrics and Gynecology*”, onde o modelo teve uma precisão diagnóstica de 90 %.

Gilson *et al.*[60] realizaram um estudo para avaliar o desempenho do ChatGPT ao responder a questões dos exames de Licenciamento Médico dos Estados Unidos. Os resultados indicaram que as respostas geradas pelo modelo apresentavam uma lógica e um contexto informativo adequados, com um desempenho equivalente ao de um estudante de medicina no terceiro ano. No entanto, o desempenho do modelo diminuiu significativamente à medida que a dificuldade das questões aumentava.

Cadiente *et al.*[61] destacam que os *chatbots* não são perfeitos e devem usar-se como ferramentas complementares, e não como substitutos de uma opinião clínica ou das ferramentas educacionais tradicionais. Mais uma vez, é reforçada a ideia de que os *chatbots* apresentam dificuldades em lidar com diagnósticos complexos, pois ignoram a sequência de exames habitual, dando diretamente o diagnóstico do paciente [61]. Este erro é importante, porque em muitas situações médicas, uma avaliação cuidadosa e gradual é importante para um diagnóstico preciso, mas os *chatbots* tendem a basear-se apenas nos sintomas dos pacientes para sugerir um diagnóstico. Além disso, os *chatbots* demonstraram problemas relacionados com a confidencialidade e privacidade do paciente. Embora entendam o básico sobre questões éticas, como a importância da confidencialidade, esses modelos falham na aplicação desses conceitos em cenários mais complexos, especialmente aqueles que envolvem menores de idade [62].

Hake *et al.*[62] realizaram um estudo com o objetivo de avaliar a capacidade do ChatGPT para resumir dados de artigos científicos, sendo os resumos posteriormente avaliados por médicos. Os resultados demonstraram que o ChatGPT produziu resumos cerca de 70% mais curtos do que os *abstracts* originais dos artigos, mas mantendo elevados níveis de qualidade e precisão. Dos 140 resumos analisados, apenas quatro apresentaram alucinações, que é o termo usado quando um modelo de grande escala, como o ChatGPT, gera conteúdo que não corresponde à realidade.

Desta forma, o problema centra-se na falta de tempo dos médicos para realizar pesquisas e no crescente volume de evidência científica publicada, dificultando a identificação da informação mais relevante para a prática clínica. Neste contexto, os LLMs surgem como soluções promissoras para facilitar o acesso à informação mais pertinente e recente, como novos tratamentos e métodos de diagnóstico. Contudo, por se tratar de uma tecnologia recente e ainda com várias limitações, a sua utilização deve ser cuidadosamente avaliada e integrada de forma responsável na prática clínica.

1.3 Objetivos

Como mencionado anteriormente, os médicos enfrentam alguns desafios para se manterem atualizados face ao volume crescente de avanços científicos e à complexidade dos cuidados clínicos. Esta dificuldade pode contribuir para erros de diagnóstico ou decisões clínicas menos adequadas.

Nesse contexto, a utilização de tecnologias avançadas, em particular as que usam a AI generativa para interagir com os profissionais de saúde, pode ajudar os médicos a manterem-se informados acerca dos mais recentes avanços científicos e tecnológicos na obstetrícia.

O principal objetivo será criar um *chatbot* baseado em RAG, capaz de fornecer informação atualizada, obtida a partir de fontes científicas confiáveis. O sistema pretende permitir que o médico coloque questões em linguagem natural e obtenha respostas atualizadas, personalizadas e adequadas ao contexto clínico mais recente.

A hipótese da pesquisa é a seguinte:

A utilização de um chatbot baseado em Retrieval-Augmented Generation (RAG), alimentado por fontes científicas confiáveis e atualizadas, melhora a capacidade dos obstetras para aceder e utilizar evidência científica recente no apoio à tomada de decisão clínica, quando comparada com os métodos tradicionais de pesquisa de informação.

Com base nisso, a seguir encontram-se os objetivos que se pretende alcançar ao longo do desenvolvimento da dissertação de mestrado.

- **OB1:** Realizar uma revisão da literatura sobre os principais avanços tecnológicos que suportam a atualização contínua dos profissionais de saúde.
- **OB2:** Construir uma base de dados bibliográfica composta por artigos científicos e orientações clínicas relevantes para a prática da obstetrícia em Portugal.
- **OB3:** Selecionar e integrar modelos de *embeddings* e LLMs para suportar a recuperação de informação e a geração de respostas em contexto clínico.
- **OB4:** Desenvolver, testar e validar um agente conversacional baseado em RAG suportado pela base de dados construída.

1.4 Estrutura da tese

Esta tese encontra-se organizada em sete capítulos, cuja estrutura acompanha o desenvolvimento do trabalho.

O capítulo um, denominado “Introdução”, apresenta o contexto e a motivação da investigação descrevendo a relevância da atualização do conhecimento médico no domínio da obstetrícia e os desafios associados à análise de documentação clínica. São ainda definidos o problema de investigação e os objetivos que orientam o desenvolvimento deste trabalho.

O capítulo dois, apresenta os principais conceitos teóricos relacionados com a “Inteligência Artificial Generativa e Modelos de Linguagem”, importantes para a solução proposta.

O capítulo três apresenta a “Revisão Literária” descrevendo a metodologia de pesquisa adotada, as questões de investigação formuladas, a estratégia de pesquisa utilizada e a análise dos estudos selecionados. São igualmente discutidos os principais resultados identificados e as limitações e desafios reportados na literatura.

O capítulo quatro descreve a “Solução Proposta”, apresentando a arquitetura geral do sistema desenvolvido e uma explicação breve dos módulos utilizados no pré-processamento do sistema.

O capítulo cinco descreve a “Implementação” da solução. São apresentadas as tecnologias e os modelos utilizados, bem como os diferentes componentes desenvolvidos para o pré-processamento dos documentos. Por fim, apresenta a interface do utilizador desenvolvida para o sistema RAG.

O capítulo seis apresenta a “Experimentação”, onde é explicado o procedimento utilizado para avaliar a solução desenvolvida. São descritos o conjunto de dados utilizado, a configuração experimental, os modelos avaliados e as métricas consideradas. São igualmente apresentados e discutidos os resultados obtidos na avaliação dos modelos de *embeddings* e dos modelos de linguagem.

Por fim, o capítulo sete apresenta as principais conclusões deste trabalho, discutindo os objetivos definidos inicialmente e as principais limitações e possíveis trabalhos futuros.

2 Inteligência Artificial Generativa e Modelos de Linguagem

Inteligência artificial (AI) refere-se à capacidade que uma máquina tem para reproduzir funções cognitivas humanas, tais como a aprendizagem, o raciocínio e a aptidão de resolver problemas e tomar decisões. Estes sistemas assentam em modelos estatísticos e técnicas de *machine learning* (ML) utilizados para analisar dados e apoiar processos de decisão automatizada [63], [64].

O desenvolvimento mais recente da AI é impulsionado pelos avanços na capacidade computacional e pelo acesso a grandes volumes de dados, fatores que possibilitaram a criação de modelos progressivamente mais sofisticados. Neste contexto, destaca-se a Inteligência Artificial Generativa, uma vertente da AI orientada para a criação de novos conteúdos, como texto, áudio e código, com base nos padrões assimilados durante o treino [63].

Entre as aplicações mais relevantes da Inteligência Artificial Generativa encontra-se a geração automática de texto em linguagem natural, área na qual modelos como o GPT têm demonstrado um desempenho notável [65]. Estes sistemas são capazes de gerar respostas coerentes, contextualizadas e semanticamente estruturadas, aproximando-se em muitos casos, da linguagem humana.

A capacidade de um modelo gerar texto em linguagem natural enquadra-se no domínio do NLP, uma subárea da AI dedicada ao desenvolvimento de métodos que permitem aos sistemas computacionais compreender, interpretar e produzir linguagem humana. Nos últimos anos, os avanços nesta área têm sido determinantes para o surgimento de LLM, cujo funcionamento e as principais características serão explorados nas secções seguintes [66].

2.1 Processamento de Linguagem Natural

Processamento de Linguagem Natural (NLP) é um subcampo da AI que se dedica a analisar, modelar e gerar linguagem humana. O seu principal objetivo consiste em transformar dados não estruturados, como texto ou discurso, em representações estruturadas que possam ser processadas por sistemas computacionais [67], [68].

A linguagem natural apresenta diversos desafios que dificultam o desenvolvimento dos modelos computacionais que funcionem automaticamente. Entre estes destacam-se ambiguidade lexical, na qual uma palavra pode corresponder a diferentes interpretações, e a dependência contextual, em que o significado de uma palavra, expressão ou frase apenas pode ser corretamente inferido quando é considerado o contexto em que ocorre. Isto compromete o desempenho dos modelos de NLP, conduzindo a interpretações incorretas do texto [67], [68].

Para responder a estes desafios, as abordagens de NLP evoluíram ao longo do tempo. Inicialmente, predominavam sistemas baseados em regras linguísticas, que foram posteriormente substituídos por métodos estatísticos treinados em grandes coleções de texto. Atualmente, a *deep learning* (DL) representa a abordagem dominante. Estes são modelos que aprendem automaticamente a identificar padrões linguísticos, inferir relações semânticas e captar dependências contextuais entre palavras e conceitos. Assim, em vez de analisarem cada palavra isoladamente, consideram o contexto em que ela aparece para inferir o seu significado. Para além disso, podem ser posteriormente especializados para domínios particulares, como a medicina, o direito ou a engenharia, através do *fine-tuning*, onde o modelo aprende os conceitos e as terminologias utilizadas nesses domínios [68], [69].

Fundamentalmente, o NLP organiza-se em duas grandes vertentes complementares: *Natural Language Understanding* e *Natural Language Generation*. A *Natural Language Understanding* foca-se na interpretação automática da linguagem, incluindo tarefas como classificação de texto, análise de sentimentos, extração de entidades nomeadas e resposta automática a perguntas. Já a *Natural Language Generation* dedica-se à produção automática de texto compreensível para humanos, sendo aplicada, por exemplo, para criar relatórios, resumir textos automaticamente e gerar conteúdos a partir de dados estruturados ou multimodais [69], [70].

O NLP inclui um conjunto de tarefas que permitem aos sistemas computacionais compreender, analisar, transformar e gerar linguagem humana. Entre as mais relevantes destacam-se:

- Tokenização é o processo de segmentar um texto em unidades menores, chamadas *tokens*, que podem ser palavras, subpalavras, caracteres ou até *bytes*. No início, os *tokens* correspondiam a palavras separadas por espaços e pontuações. Mas com a introdução dos modelos neurais mais modernos, a tokenização passou a utilizar *subword units* (como WordPiece), que dividem palavras em partes menores. Este processo é importante para converter um texto em representações numéricas (*embeddings*) que possam ser processadas por modelos de NLP [71].

- *Stemming*: é uma técnica que reduz as palavras às suas formas base (*stems*) através da remoção de sufixos e terminações, sem garantir que o resultado seja uma palavra válida no dicionário. Por exemplo, *better*, pode ser reduzido para *bett*, desta forma agrupam-se variações morfológicas para melhorar a recuperação de informação [72].
- Lematização: é uma técnica que reduz as palavras, mas utiliza uma análise linguística e vocabulário para converter as palavras na sua forma canónica correta. Diferente do *stemming*, a lematização considera o contexto gramatical e produz palavras corretas, como por exemplo, *better* pode ser reduzido para *good* [72].
- Análise semântica é a área do NLP que procura interpretar o significado das palavras, frases e textos, indo além da sua estrutura gramatical. Desta forma, a análise semântica permite compreender a intensão, as emoções e o contexto expressos pelo utilizador, tendo em conta vários fatores como por exemplo, ambiguidade, cultura, ironia e variações linguísticas que podem alterar o sentido de uma mensagem. Esta análise é importante para tarefas como análise de sentimentos e compreensão automática de texto [73].
- *Clustering* é uma técnica de aprendizagem não supervisionada que agrupa os dados semelhantes em *clusters*, com base numa medida de similaridade ou distância. Desta forma, o *clustering* permite organizar automaticamente documentos em grupos coerentes. Para isso, o *clustering* baseia-se em métricas como a distância Euclidiana ou similaridade do cosseno, para medir a proximidade entre *embeddings*. Em NLP, utiliza-se a técnica de *clustering* para organizar documentos segundo a sua semelhança semântica [74].
- Extração de Entidades Nomeadas, ou *Named Entity Recognition* (NER), é uma tarefa de NLP que consiste em identificar e classificar entidades nomeadas num texto, como pessoas, organizações e locais. O seu objetivo é extrair automaticamente informação relevante, facilitando a transformação de texto não estruturado em dados estruturados. O NER localiza automaticamente essas expressões e atribui-lhes uma categoria semântica adequada. É útil para aplicações como extração de informação, *question-answering* e tradução automática, pois ajuda a preservar e estruturar informação essencial no texto [75].
- Extração de Relações é uma tarefa de NLP que identifica automaticamente entidades num texto e determina as relações semânticas entre elas, normalmente representadas por tripletos. A extração de entidades transforma um texto não estruturado em informação estruturada, útil para aplicações como grafos de conhecimento, *question answering* e recuperação de informação [76].
- Modelação de Tópicos, ou *Topic Modelling*, são técnicas não supervisionadas que identificam automaticamente padrões semânticos em grandes coleções de documentos. Representam cada documento como uma distribuição de tópicos e cada

tópico como uma distribuição de palavras, permitindo uma interpretação humana muito mais fácil, dos conteúdos de um conjunto de documentos. São amplamente utilizados em mineração de texto, NLP e noutras áreas para organizar, resumir e explorar grandes volumes de dados não estruturados [77].

Inicialmente, as tarefas de NLP eram tratadas de forma isolada, através de modelos específicos desenvolvidos para cada aplicação. Contudo, com a evolução do DL, os modelos baseados em redes neurais profundas passaram a aprender representações linguísticas mais expressivas e generalizáveis, permitindo adaptar um mesmo modelo a múltiplas tarefas [78], [79]. Esta mudança permitiu o desenvolvimento de arquiteturas cada vez mais escaláveis, capazes de capturar as dependências contextuais que existem na linguagem natural.

Destaca-se a arquitetura *Transformer*, que introduziu mecanismos de atenção capazes de modelar relações de longo alcance no texto, tornando-se mais eficaz do que as abordagens anteriores. A secção seguinte aprofunda os princípios fundamentais desta arquitetura e o seu papel na evolução mais recente do NLP [78], [79].

2.2 Arquitetura Transformer

Os *Transformers* foram introduzidos em 2017, no trabalho intitulado *Attention Is All You Need*, que apresentou uma nova abordagem para tarefas de NLP. Em contraste com modelos sequenciais, como as *Recurrent Neural Networks*, que processam os *tokens* de forma ordenada e sequencial, o *Transformer* usa um mecanismo de *attention* capaz de processar todos os *tokens* em paralelo [80].

Para representar a posição de cada *token* na sequência de entrada, o *Transformer* adota o método denominado *positional encoding*, que adiciona ao *embedding* de cada *token* um vetor que codifica sua posição relativa no texto. Esses vetores são gerados por funções seno e cosseno de diferentes frequências, o que possibilita ao modelo distinguir a ordem e a distância entre *tokens*, mesmo sem recorrer a estruturas recorrentes [80].

Outra diferença está no núcleo do processamento do *Transformer*, o mecanismo de *self-attention*. Neste processo, cada *token* é transformado em vetores, que são usados para calcular a relevância de cada palavra em relação às outras da mesma sequência. Essa operação é executada em paralelo em várias “cabeças” de atenção (*multi-head attention*), permitindo que diferentes aspetos das relações entre *tokens* sejam capturados simultaneamente, como dependências gramaticais e ligações semânticas [80]. A arquitetura completa, composta pelos módulos de *encoder* e *decoder*, é apresentada na Figura 1.

A arquitetura *Transformer* criou as condições necessárias para o desenvolvimento de modelos cada vez maiores e mais generalistas, ao mudar a forma como é feita o NLP. A capacidade de captar dependências complexas em larga escala contribuiu para o treino de redes com bilhões

de parâmetros, capazes de aprender padrões linguísticos a partir de volumes massivos de dados [81].

O potencial desta arquitetura tornou-se evidente com o destaque o lançamento do BERT, da Google, que, poucos anos após o aparecimento da arquitetura *Transformer*, demonstrou o potencial desta arquitetura. O BERT mostrou que, ao utilizar múltiplas camadas de *Transformers* e o mecanismo de *self-attention*, era possível captar as relações que cada termo estabelece com os restantes elementos da frase. Essa capacidade permitiu superar limitações dos modelos anteriores baseados em *word embeddings* estáticos, que não conseguiam distinguir adequadamente palavras com múltiplos sentidos [81].

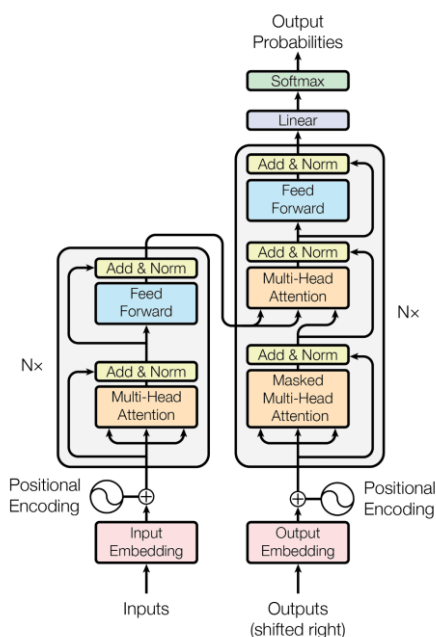


Figura 1: Arquitetura do *Transformer*. Fonte [80]

A evolução destas abordagens conduziu ao aparecimento de uma nova geração de modelos capazes de compreender e gerar linguagem natural com elevado desempenho, nomeadamente os LLMs.

2.3 Modelos de Linguagem de Grande Escala

Os Modelos de Linguagem de Grande Escala (LLMs) são modelos probabilísticos criados para estimar a distribuição conjunta de probabilidade de *tokens*, numa determinada sequência de texto. Dessa forma, estes modelos são capazes de aprender padrões linguísticos a partir do contexto anterior e prever a probabilidade de ocorrência do *token* seguinte ou, alternativamente, gerar novas sequências de texto de forma autónoma [82], [83].

Os LLMs enquadram-se nos modelos de NLP e são treinados na arquitetura de rede neural *transformer*. Esta arquitetura envolve o treino de bilhões ou centenas de bilhões de parâmetros, o que justifica a designação de modelos de “grande escala” [84]. Estes modelos têm

demonstrado um desempenho notável em diversas tarefas, como a geração de texto, tradução automática e sumarização. Para além disso, são capazes de manter conversas de forma semelhante à interação humana e de produzir diferentes tipos de conteúdo, incluindo e-mails, artigos e código computacional. Entre os exemplos mais conhecidos de LLMs encontram-se o PaLM, o LLaMA e a série GPT [84],[85].

Mais recentemente, surgiram os Modelos de Linguagem Multimodal de Grande Escala (MLLMs), que são sistemas de inteligência artificial avançados capazes de processar, compreender e gerar informações a partir de diferentes tipos de dados em simultâneo. Ao contrário dos LLMs, que utilizam apenas texto, os MLLMs utilizam vários tipos de dados, como texto, imagens, áudio e vídeo. Esta capacidade de integrar vários dados diferentes, permite que o modelo realize tarefas como descrever o conteúdo de uma imagem ou responder a perguntas relacionadas com um vídeo [86].

No domínio da medicina, a aplicação dos LLMs tem sido amplamente investigada, demonstrando potencial para apoiar tarefas clínicas e educativas. Entre as suas utilizações destacam-se a geração automática de relatórios médicos, o apoio à formulação de hipóteses diagnósticas e a resposta a questões clínicas [87], [88]. Zhou et al. [82] avaliaram o desempenho destes modelos através do *United States Medical Licensing Examination*, um exame de licenciamento médico nos Estados Unidos que avalia os conhecimentos clínicos e a capacidade de raciocínio médico dos candidatos. Neste estudo, modelos como o Med-PaLM 2 (baseado no PaLM 2) e o MedPrompt (baseado no GPT-4) alcançaram taxas de acerto de 86,5% e 90,2%, respetivamente, valores comparáveis ou superiores ao desempenho médio de médicos especialistas humanos (87,0%). Estes resultados demonstram o potencial destes modelos no apoio à prática clínica.

Apesar do desempenho em exames padronizados, a integração dos LLMs em contextos clínicos reais levanta algumas questões, principalmente relacionadas com a fiabilidade e segurança das respostas geradas pelos modelos. A sua aplicação prática ainda requer mecanismos de validação rigorosos e supervisão humana. Para além disso, os modelos precisam de estar em conformidade com as normas éticas e regulamentares aplicáveis [84], [89].

Os LLMs apresentam algumas limitações importantes, incluindo a geração de alucinações, a existência de enviesamentos e estereótipos e questões éticas relacionadas com a sua utilização [90]. Estas preocupações têm motivado o desenvolvimento de novas técnicas para aumentar a qualidade e segurança das respostas.

Uma dessas técnicas é o *prompt engineering*, que consiste na conceção e otimização das instruções fornecidas ao modelo, de forma a orientar o processo de geração de respostas mais precisas, claras e adequadas ao contexto [91], [92].

Outra abordagem passa pela implementação de RAG, onde o conhecimento do modelo para responder a questões vem de fontes externas, atualizadas e fidedignas. Ou seja, em vez de depender exclusivamente do conhecimento adquirido durante o pré-treino, o modelo utiliza essa informação como contexto para gerar a resposta. Na área da saúde, esta abordagem é

particularmente relevante, uma vez que permite fundamentar as respostas em evidência verificável, contribuindo para reduzir a ocorrência de alucinações e aumentar a fiabilidade das respostas [84].

2.4 Prompt Engineering

A prática de *prompt engineering* transforma requisitos informais em instruções melhoradas para LLMs. Ao explicitar o que deve ser feito, com que informação e em que formato, as instruções orientam de forma clara o comportamento do modelo, produzindo respostas mais alinhadas com a tarefa e menos suscetíveis a ambiguidades [93].

Assim, o *prompt* influencia diretamente o conteúdo, o estilo e a estrutura das respostas. Consoante a tarefa, pode variar de uma frase breve a um conjunto articulado de instruções, que inclui contexto do problema, exemplos da resposta e até pode referir como o modelo se deve comportar. Esta estruturação do *prompt* transforma o modelo num “especialista”, por exemplo, para resumir documentos, escrever código, apoiar decisões, gerar ideias ou responder a questões técnicas. Um design cuidadosamente preparado torna as respostas mais claras e diretas, reduzindo os erros [84],[93].

Entre as estratégias baseadas em exemplos, existem as variações zero-shot e *few-shot*. Na abordagem de zero-shot, o modelo é orientado apenas pelas instruções definidas no *prompt*, sem qualquer exemplo. Já na abordagem de *few-shot*, acrescentam-se ao *prompt* alguns exemplos de resposta que funcionam como demonstrações do comportamento esperado, sem atualização de pesos. Esta inclusão tende a reduzir ambiguidades e a estabilizar o estilo e o formato das respostas [84], [94].

Outra estratégia é a *Chain of Thought*, que incentiva a decomposição de problemas em sub-perguntas encadeadas e a explicitação de passos intermédios de raciocínio. Ao dividir a pergunta principal em etapas, a *Chain of Thought* melhora a capacidade de LLMs responderem a questões mais complexas [95], [96].

2.5 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) é uma técnica de LLM com mecanismos de recuperação de informação. Assim, antes de gerar uma resposta, o sistema pesquisa automaticamente documentos relevantes em fontes de conhecimento externas, disponibilizando essa informação ao modelo como contexto adicional. Desta forma, o modelo deixa de depender apenas do conhecimento adquirido durante o pré-treino, podendo fundamentar as suas respostas em informação mais atual, específica e verificável. Esta abordagem é particularmente útil quando se pretende reduzir a ocorrência de alucinações e melhorar a atualidade, a transparência e a fiabilidade das respostas geradas [97], [98], [99]

2.5.1 Componentes Básicos de RAG

Um pipeline de Retrieval-Augmented Generation (RAG) é constituído por três componentes principais: indexação dos documentos, recuperação da informação e geração da resposta. A fase de indexação engloba todas as etapas necessárias para preparar os documentos para posterior consulta, incluindo o pré-processamento, a normalização, a segmentação do conteúdo, a criação de embeddings e o armazenamento numa base de dados vetorial. A componente de recuperação tem como objetivo identificar os segmentos mais relevantes em função da questão colocada pelo utilizador. Por fim, a componente de geração utiliza o contexto recuperado para produzir respostas fundamentadas em informação proveniente de fontes confiáveis (Figura 2) [98].

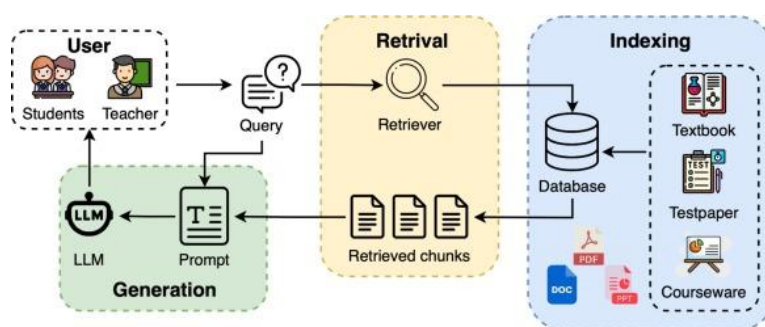


Figura 2: Esquema de um sistema de RAG. Fonte: [98]

Embora estas componentes desempenhem funções distintas, o seu funcionamento é fortemente interdependente, sendo determinante para o desempenho global de um sistema RAG.

2.5.2 Pré-processamento e document parsing de documentos científicos

Uma das primeiras etapas da fase de indexação consiste no pré-processamento dos documentos científicos. Durante esta etapa os documentos são extraídos, estruturados e preparados para posterior segmentação e indexação. A qualidade desta etapa influencia diretamente o desempenho do sistema, uma vez que erros na extração ou perda de informação propagam-se para todas as fases subsequentes, comprometendo a recuperação de contexto e, consequentemente, a qualidade das respostas geradas pelo modelo de linguagem [100].

Este processo é frequentemente designado na literatura por *document parsing*, correspondendo à conversão de documentos não estruturados, como ficheiros PDF, numa representação estruturada composta por texto, títulos, tabelas, figuras e outros elementos semânticos. No contexto de artigos científicos, esta tarefa é particularmente desafiante devido à elevada complexidade estrutural destes documentos e à necessidade de preservar corretamente a relação entre os seus diferentes elementos [101].

Ao contrário de outros tipos de documentos, os artigos científicos apresentam informação distribuída por diferentes componentes estruturais, não se limitando ao texto corrido. Para

além do texto principal, incluem frequentemente tabelas, figuras, gráficos, equações e respectivas legendas, que contêm informação científica relevante para a compreensão dos resultados apresentados. Conforme salientado por Ga *et al.* [102], uma parte substancial da informação presente nos artigos científicos perde-se quando estes elementos não são devidamente recuperados e integrados no processo de extração, uma vez que tabelas e figuras contêm informação científica relevantes que não se encontram descritos no texto principal. As figuras presentes nos artigos científicos melhoram a compreensão do conteúdo, funcionando muitas vezes como um complemento indispensável ao texto. Gráficos, fluxogramas, esquemas e imagens ilustrativas condensam informação que seria mais difícil ou mais demorado de perceber apenas na forma textual. Por esse motivo, a extração automática e a recuperação de imagens melhora a qualidade de um sistema baseado em RAG [102].

A preservação destes diferentes elementos documentais torna, por isso, a extração de informação de artigos científicos um processo tecnicamente exigente, devido à complexidade estrutural e semântica característica deste tipo de documentos. Neste contexto, os principais desafios podem ser agrupados em três categorias principais:

- Extração do texto principal do documento

Os artigos científicos apresentam uma grande variedade de elementos estruturais, tais como cabeçalhos, rodapés, títulos, subtítulos, legendas, figuras, tabelas, notas de rodapé e referências bibliográficas. Apesar de serem essenciais para a organização e compreensão do documento, quando estes elementos não são devidamente identificados e tratados, podem introduzir ruído no processo de extração [103], [104].

Para além da diversidade de elementos textuais que compõem os artigos científicos, importa destacar que também existe uma grande diversidade relacionada com o *layout* e organização visual dos artigos científicos. Os documentos variam quanto ao tipo e tamanho de letra, a utilização de diferentes estilos para os títulos e subtítulos, bem como a disposição espacial das tabelas, figuras e respectivas legendas. Para além disso, alguns artigos incluem notas na lateral da página, cabeçalho e rodapé, quando estes elementos não são corretamente identificados, texto relativo a estes elementos pode aparecer incorretamente entre os segmentos de texto extraído do documento [105].

Acresce ainda a organização do texto numa ou várias colunas, algo que é comum nos artigos biomédicos publicados. Muitas ferramentas extraem o texto da página de forma sequencial, de cima para baixo. Mas, quando o documento se encontra dividido em duas colunas, o texto extraído vai misturar conteúdo de ambas as colunas, originando frases incompletas e sem sentido. Nestes casos, a extração de texto exige a utilização de técnicas capazes de identificar a ordem lógica do texto [103], [106].

No que diz respeito à utilização de LLMs para a extração de texto de artigos científicos, embora estes modelos apresentem um elevado desempenho em tarefas de compreensão e geração de texto, persistem limitações quando são aplicados à extração literal de informação científica dos artigos. Por exemplo, existe uma tendência para resumir ou reformular o conteúdo dos documentos, mesmo quando o objetivo é fazer uma extração fiel do texto original. Este

comportamento decorre da natureza generativa destes modelos, que são otimizados para produzir texto coerente e contextualizado, mas não necessariamente para preservar integralmente a estrutura e formulação originais. Consequentemente, podem ocorrer alterações semânticas subtis, omissões ou simplificações que comprometem o rigor científico [107].

Zhang *et al.* [101] reforçam esta ideia, ao referirem que a utilização de modelos multimodais generalistas, como o GPT-4, Qwen-VL e a série LLaMA, não assegura, atualmente, a precisão estrutural que é exigida para os artigos científicos. Estes modelos retornam resultados com alucinações e têm dificuldades em lidar com documentos muito longos, uma vez que as limitações de contexto dos modelos podem dificultar a recuperação precisa do contexto [101].

- Extração de tabelas

Outro aspeto relevante no contexto da recuperação de informação prende-se com a extração de tabelas. Durante a extração de tabelas devem-se considerar duas etapas principais. A primeira etapa é a análise de *layout*, que serve essencialmente para detetar tabelas no documento. A segunda etapa é o reconhecimento de estrutura, ou seja, compreender a organização interna da tabela, nomeadamente, cabeçalhos, colunas, linhas, células; em documentos imagem, acrescenta-se *Optical Character Recognition* (OCR) para detetar o texto e convertê-lo para um formato digital legível [108].

Apesar dos avanços significativos das abordagens baseadas em ML, DL e LLMs, a extração automática de tabelas continua a apresentar diversos desafios, especialmente quando se trata de documentos com estruturas tabulares complexas. Entre os principais desafios destacam-se a complexidade estrutural, resultante da presença de cabeçalhos hierárquicos e células unidas, o que dificulta a identificação da sua estrutura e a correta interpretação do conteúdo. Outro desafio importante é a diversidade de formatos e estilos. As tabelas podem apresentar diferentes organizações visuais, incluindo a ausência de linhas delimitadoras, diferentes estilos de formatação e *layouts* bastante distintos, dificultando a generalização dos métodos de extração. Por fim, existem limitações computacionais, uma vez que os algoritmos devem conciliar elevada precisão com eficiência computacional, permitindo a sua utilização em tempo real e em ambientes com recursos computacionais limitados [104], [109].

- Extração de imagens

Os artigos científicos incluem frequentemente figuras, como gráficos, fluxogramas, esquemas e imagens ilustrativas, que apresentam resultados experimentais, descrevem metodologias ou resumem informação que ajuda na compreensão do estudo.

A extração automática destes elementos continua a ser um desafio técnico relevante, devido à complexidade estrutural dos artigos científicos e à necessidade de preservar corretamente o contexto associado a cada figura. Um dos principais desafios consiste em distinguir as figuras científicas de outros elementos gráficos presentes no documento, como logótipos de editoras, cabeçalhos, rodapés ou elementos decorativos, cuja inclusão pode introduzir ruído nos dados extraídos.

Outro desafio importante prende-se com a associação entre cada figura e a respetiva legenda, uma vez que esta contém o contexto necessário para a correta interpretação do conteúdo visual. Esta tarefa é dificultada pelo facto de nem sempre existir uma correspondência explícita entre a figura e a legenda. Além disso, podem ocorrer erros na identificação de legendas, em que referências a figuras presentes no corpo do texto são incorretamente classificadas como legendas [110].

Os desafios anteriormente descritos evidenciam várias dificuldades relacionadas com a extração de informação em artigos científicos. A complexidade estrutural destes documentos, que incluem múltiplas secções hierárquicas, tabelas e imagens, impõe desafios que ultrapassam a simples interpretação de texto contínuo. Neste contexto, embora os LLMs apresentem um desempenho elevado em tarefas de compreensão e geração de texto, a sua aplicação direta ao *document parsing* continua a apresentar diversas limitações.

No que diz respeito às tabelas, um estudo realizado por Deng *et al.* [111] demonstra que os LLMs e os Modelos de Linguagem de Grande Escala Multimodal (MLLMs) falham perante a leitura de tabelas com estruturas mais complexas, como por exemplo colunas ou células unidas. No caso dos MLLMs, verificam-se ainda limitações na leitura do texto presente nas tabelas, podendo ocorrer confusão entre caracteres ou palavras visualmente semelhantes. Estes modelos apresentam também dificuldades na identificação correta da estrutura tabular, nomeadamente na delimitação de linhas e colunas, conduzindo frequentemente a interpretações incorretas da organização da tabela. Adicionalmente, o processamento de tabelas extensas constitui outro desafio, uma vez que a elevada quantidade de informação pode levar os modelos a resumir ou simplificar o conteúdo da tabela, omitindo valores relevantes ou relações importantes entre os dados apresentados [107], [111].

Face às limitações identificadas, a utilização exclusiva de LLMs para a extração de informação de artigos científicos não garante a preservação rigorosa da estrutura documental, nem a recuperação correta de todos os elementos relevantes. Consequentemente, torna-se necessário recorrer a ferramentas especializadas de *document parsing*, capazes de estruturar corretamente o conteúdo dos documentos. De acordo com Adhikari *et al.* [112], os métodos de *document parsing* podem ser classificados em abordagens *rule-based* e *learning-based*. As abordagens baseadas em regras utilizam heurísticas previamente definidas para organizar e extrair informação dos documentos. Por outro lado, as abordagens baseadas em aprendizado recorrem a modelos de *deep-learning* treinados a partir de grandes volumes de dados para aprender padrões de *layout*, conteúdo textual e estrutura documental.

Segundo o artigo [112], estas bibliotecas que utilizam abordagens de *rule-based* apresentam um desempenho inferior quando são utilizadas em documentos científicos. Tal deve-se, sobretudo, à presença de expressões matemáticas complexas, como fórmulas e outros elementos visuais, que são frequentemente incorporados com o texto do artigo, resultando numa mistura de conteúdo textual e não textual. Estas bibliotecas apresentam muitas limitações em termos de adaptabilidade a diferentes estruturas e formatos. Ao contrário das

abordagens *learning-based*, que demonstram maior capacidade de generalização perante documentos com *layouts* complexos, como os artigos científicos [112].

No que diz respeito à extração de tabelas, as bibliotecas que utilizam métodos *rule-based* demonstram um bom desempenho quando são aplicados a tabelas simples, caracterizadas por bordas bem definidas e espaçamento uniforme entre colunas. No entanto, o desempenho degrada-se em documentos com *layouts* mais complexos, como artigos organizados em múltiplas colunas ou que incluem linhas gráficas em cabeçalhos, rodapés e outros elementos da página. Nestes casos, estas bibliotecas tendem a gerar um número elevado de falsos positivos, identificando incorretamente regiões do documento como tabelas. Em contraste, modelos baseados em aprendizado, como o TATR, apresentam desempenho superior, conseguindo adaptar-se a diferentes estruturas e *layouts* documentais [112].

Zhang *et al.* [101] descrevem a utilização de *pipelines* modulares para extrair o conteúdo de documentos. Nestes casos, o processamento do documento é dividido em várias etapas e cada uma responsável por uma tarefa específica. Tipicamente, estes *pipelines* iniciam-se com uma fase de análise de *layout*, cujo objetivo é identificar os elementos estruturais do documento, como blocos de texto, títulos, imagens e expressões matemáticas, bem como determinar as suas coordenadas espaciais e a ordem de leitura. Segue-se a extração de conteúdo propriamente dita, que pode incluir vários módulos especializados, nomeadamente:

1. Extração de texto, recorrendo a técnicas de OCR para converter o texto presente nas imagens em formato legível por máquina.
2. Extração de expressões matemáticas, focada na deteção de símbolos e estruturas matemáticas
3. Extração de dados e estrutura de tabelas, geralmente realizada em duas etapas, a deteção das regiões tabulares e reconhecimento da sua estrutura interna, incluindo cabeçalhos, linhas, colunas e células.
4. Reconhecimento de gráficos, que identificam diferentes tipos de representações gráficas e extrair os dados, convertendo a informação visual em formatos estruturados, como tabelas;

Para além disso, o artigo destaca ainda a existência de bibliotecas que combinam a utilização de modelos que combinam a utilizam de OCR com DL. Estas abordagens demonstram capacidade para preservar a estrutura semântica e visual dos documentos, incluindo expressões matemáticas e *layouts* em múltiplas colunas. Contudo, continuam a apresentar limitações relacionadas com erros de interpretação do conteúdo, elevados custos computacionais, problemas de escalabilidade e reduzida diversidade linguística [101].

2.5.3 Indexação de documentos

Concluída a etapa de pré-processamento, a fase de indexação prossegue com a segmentação dos documentos, a criação de *embeddings* e o armazenamento da informação numa base de dados vetorial.

Nesta fase, a informação proveniente de fontes externas, como bases de dados institucionais, repositórios académicos, documentos técnicos ou literatura científica e médica, é organizada e indexada de forma a poder ser posteriormente recuperada e utilizada pelo modelo de linguagem [98], [99].

Nem sempre existe este processo de indexação num sistema RAG. Existem abordagens em que o sistema recebe diretamente, como entrada, textos externos, por exemplo, resumos de documentos, que servem como fonte imediata de recuperação. Nestes casos, o sistema não depende de uma base de conhecimento previamente indexada, mas sim do conteúdo que o modelo recebe no momento da consulta [99].

Toda esta fase de indexação é realizada previamente, ou seja, antes de correr o modelo, é construída uma base de conhecimento externo, que será um complemento ao conhecimento do LLM. Para garantir que a recuperação ocorre de forma correta, os dados passam por um processo de pré-processamento, no qual são limpos, normalizados e segmentados em unidades menores chamadas *chunks*. De seguida, os *chunks* são convertidos em representações numéricas, usando modelos de *embeddings* e são armazenados em base de dados vetoriais. Essa estrutura permite a recuperação rápida e precisa das informações mais relevantes [98]. A fase de indexação inclui três componentes principais [98], [113]:

1. Segmentação (*Text Splitting*). Os documentos são divididos em unidades menores, denominados *chunks*, para facilitar o processamento da informação. Podem usar-se diferentes estratégias de divisão dos documentos, por exemplo: por caracteres ou palavras; por frases ou parágrafos; ou com sobreposição, em que cada *chunk* partilha algumas frases com o anterior, de forma a manter a continuidade semântica.
2. Vectorização. Cada *chunk* é convertido num vetor numérico usando modelos apropriados, permitindo medir similaridade semântica entre conteúdos.
3. Armazenamento e indexação vetorial. Os *embeddings* são armazenados numa base de dados vetorial, juntamente com metadados relevantes.

Assim, a indexação oferece a estrutura semântica necessária para consultas mais eficazes nos conjuntos de documentos.

2.5.4 Recuperação de textos

A Recuperação é a fase em que são identificadas as passagens de texto mais relevantes para responder ao utilizador. Para tal, diferentes pesquisadores propuseram diferentes métodos de representação textual, nomeadamente: *sparse retrieval*, *dense retrieval* e *contextual retrieval*. Estas representações mapeiam consulta e passagens para um espaço comum, no qual a relevância é estimada por medidas de similaridade (por exemplo, cosseno ou produto interno) [114].

Os Modelos de *Sparse Retrieval* correspondem a métodos baseados em correspondência lexical, ou seja, na presença de palavras idênticas entre a pergunta do utilizador e os documentos

recuperados [98]. O método mais básico é o *one-hot encoding*, onde cada palavra é representada com um vetor numérico (tão grande quanto o tamanho do vocabulário) e apenas um valor é “1” (a posição da palavra no vocabulário), todos os outros são “0”. Desta forma, este método assinala a presença ou ausência de palavras [114].

Nos Modelos de *Sparse Retrieval* destaca-se os métodos de *Bag-of-Words*, TF-IDF (*term frequency and inverse document frequency*) e BM25 (*Best Matching 25*). *Bag-of-Words* é uma melhoria do *one-hot encoding*, ao realizar a contagem das palavras nos documentos (frequência de termos). TF-IDF aplica uma ponderação às contagens, combinando a frequência do termo no documento com o inverso da frequência do termo no documento, desta forma, reduz o peso de palavras muito frequentes no documento. BM25 é um aperfeiçoamento da técnica de TF-IDF, que ajusta a frequência de termo, de forma a que repetições adicionais contem cada vez menos, e normaliza a pontuação pelo comprimento do documento, desta forma, tem em consideração a frequência de termos, a raridade com que aparecem no documento e o comprimento do documento [114].

Ao contrário dos Modelos de *Sparse Retrieval*, a *Dense Retrieval* baseia-se numa análise semântica, isto é, na capacidade de representar e comparar o significado de palavras subjacentes de palavras, frases ou documentos, independentemente da correspondência literal de termos. Assim, a *Dense Retrieval* mapeia as consultas do utilizador e as passagens de texto para *dense embeddings* num espaço vetorial comum. A relevância é calculada pela similaridade entre os vetores, tipicamente a similaridade do cosseno ou o produto interno, onde os valores mais altos indicam maior afinidade semântica entre a pergunta e os documentos. Deste modo, recuperam-se textos semanticamente próximos da consulta, mesmo na ausência de palavras idênticas.

Entre as primeiras abordagens tem-se os modelos word2vec e GloVe [98]. Com o surgimento dos *Transformers*, os modelos pré-treinados começaram a ser usados para pesquisa e *ranking* de texto. Entre estes, surgiram o BERT e derivações deste, como o SciBERT (treinado com texto científico geral) e o BioBERT (treinado com artigos biomédicos). O Sentence-BERT (SBERT) é também uma modificação do BERT, mas com uma arquitetura *siamese* ou *triplet*, capaz de encontrar pares de frases semanticamente similares em grandes conjuntos de dados, mantendo a precisão dos modelos BERT, mas reduzindo o custo computacional [115].

Nos Modelos de *Dense Retrieval*, a relevância entre a consulta e os trechos de texto é estimada segundo a similaridade do cosseno e às vezes o produto interno [98].

2.5.5 Geração da Resposta

Nesta fase, o modelo generativo produz a resposta ao utilizador a partir do *prompt*. Num sistema RAG, o *prompt* é contextualizado com textos recuperados da base de conhecimento externa. Para tal, utiliza-se um *template* de *prompt*, uma estrutura padronizada que define como combinar a consulta do utilizador com as passagens de texto recuperadas e como os apresentar ao modelo [98].

Em geral, o *template* do *prompt* inclui três elementos [98]:

- Consulta do utilizador (pergunta ou pesquisa).
- Bloco de contexto com os excertos de texto mais relevantes, retirados da base de conhecimento.
- Instruções e restrições que orientam o modelo a responder apenas com base nesse contexto, especificando formato de saída e outras diretrizes para minimizar alucinações.

Desta forma, o LLM recebe o *prompt*, combina o conhecimento aprendido no pré-treinamento (e, se houver, no *fine-tuning*), com a informação recuperada e prioriza essa informação para gerar uma resposta. Para preparar e combinar o contexto (ou seja, a informação recuperada), a abordagem mais simples é a concatenação direta, onde os trechos recuperados são anexados à consulta. Embora seja funcional, pode levar a que o contexto exceda o limite de *tokens* permitidos pelo modelo ou pode trazer problemas como o excesso de informação. Para diminuir estes problemas, recorre-se a técnicas como [113]:

- Filtragem por relevância: seleciona-se as passagens de texto mais relevantes usando filtros nos metadados.
- Compressão de texto: reduz-se o texto ao mais importante, removendo *tokens* irrelevantes ou resumindo os textos.
- *Re-Ranking*: reordena-se o texto recuperado com modelos mais precisos, para incluir primeiro os excertos mais relevantes.

2.5.6 Diferentes abordagens de RAG

O conceito de RAG não se restringe a uma única arquitetura, mas abrange um conjunto de variações que exploram diferentes formas de melhorar a interação entre a consulta do utilizador, o processo de recuperação e a geração da resposta. Estas abordagens distinguem-se, sobretudo, nas estratégias utilizadas para reescrever consultas, selecionar documentos, resumir informação ou integrar o contexto no *prompt*. Duas dessas arquiteturas são a *Naive RAG* e a *Graph RAG* [97].

A *Naive RAG* baseia-se num processo de *Retrieve-Read*, em que os documentos são armazenados numa base de dados e os segmentos de textos mais próximos semanticamente da consulta (os top-k *chunks*) são concatenados ao *prompt*, como mostra na Figura 3(a). Embora seja um método eficaz, apresenta algumas limitações como a recuperação de trechos de texto pouco relevantes. Isto leva a que, quando é inserida uma consulta mais complexa, o modelo RAG apresente falta de coerência e redundância nas respostas [97].

A *Graph RAG* organiza os documentos sob a forma de grafos, em que os nós correspondem a passagens de texto e as arestas representam relações semânticas entre eles, como mostra na Figura 3(b). Esta estrutura permite recuperar a passagem de texto mais relevante e os conteúdos vizinhos no grafo, tornando a resposta mais completa [97].

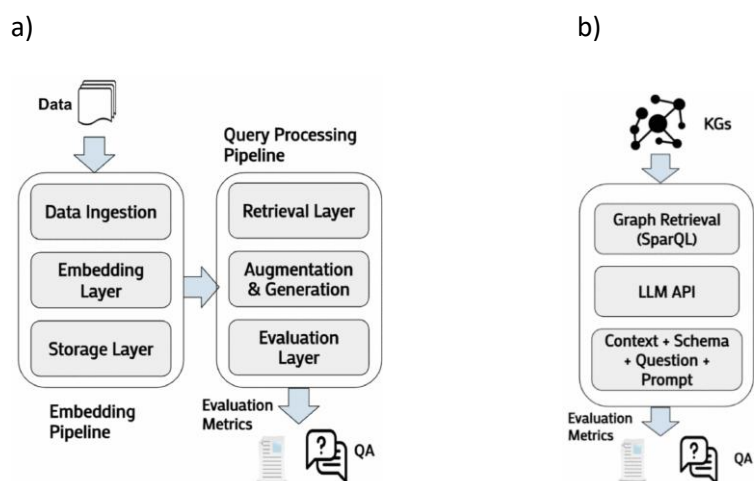


Figura 3: a) Arquitetura de *Naive RAG*. b) Arquitetura de *Graph RAG*. Fonte: [97]

As diferentes abordagens de RAG procuram responder às limitações dos sistemas mais simples, introduzindo mecanismos que melhoram a recuperação de informação e a utilização do contexto durante a geração da resposta. A escolha da arquitetura mais adequada depende das características do domínio de aplicação, da estrutura dos documentos e dos requisitos da tarefa.

2.6 Métricas de Avaliação para Modelos de Embeddings e Modelos de Linguagem

Nesta secção, são apresentadas as principais métricas utilizadas na avaliação de modelos de *embeddings* e de LLMs, de forma a compreender como estas métricas quantificam o desempenho dos modelos.

Para avaliar os modelos de *embeddings*, são utilizadas métricas como *Precision*, *Recall*, *F1-Score* e *nDCG*. No entanto, é necessária uma base de dados de referência (*ground truth*), que contenha os documentos considerados relevantes para um conjunto de consultas previamente definido. Estas métricas têm como objetivo quantificar o grau de correspondência entre os *chunks* recuperados pelo modelo e os *chunks* presentes na *ground truth*, permitindo assim avaliar a qualidade do sistema de recuperação de informação [116].

Os significados destas métricas de avaliação são descritos a seguir [117], [118]:

- *Precision* – Refere-se à proporção de documentos corretamente recuperados entre os k primeiros resultados devolvidos pelo sistema. Esta métrica avalia a qualidade dos resultados apresentados ao utilizador, penalizando a presença de informação irrelevante. A métrica *Precision@k* pode ser expressa como:

$$precision@k = \frac{\text{documentos relevantes recuperados @k}}{k} = \frac{TP@k}{TP@k + FP@k}$$

Onde os Verdadeiros Positivos (TP) correspondem aos documentos simultaneamente relevantes e recuperados, e os Falsos Positivos (FP) correspondem aos documentos recuperados, mas que não são relevantes.

Trata-se de uma métrica importante para avaliar a precisão em sistemas de recuperação de informação, especialmente quando a qualidade dos resultados é importante. A métrica de *Precision* é usada em vários trabalhos da área de Recuperação de informação [119], [120].

- *Recall* – Refere-se à proporção de *chunks* relevantes que foram corretamente recuperados pelo modelo, em relação ao total de *chunks* relevantes. No contexto de um modelo de recuperação de informação baseado numa pergunta do utilizador, o total de chunks relevantes corresponde ao conjunto de todos os documentos necessários para responder adequadamente à questão formulada, sendo este valor representado pela soma dos TP e FN. A métrica *Recall@k* pode ser expressa como:

$$recall@k = \frac{\text{chunks relevantes recuperados @k}}{\text{total de chunks relevantes}} = \frac{TP@k}{TP@k + FN@k}$$

É uma métrica importante para medir o grau de cobertura na recuperação de informação, sobretudo quando é importante identificar o maior número possível de documentos relevantes. Esta métrica é frequentemente usada como critério de avaliação experimental em diversos estudos de Recuperação de Informação [121], [122].

- *F1-Score* - Define-se como a média entre a *Precision* e a *Recall*. O F1-Score sintetiza o desempenho do modelo relativamente à identificação correta dos documentos e à cobertura de todos os documentos relevantes. Esta métrica é normalmente aplicada para avaliar sistemas de AI e ML [123], [124].

A métrica *F1-Score* pode ser expressa como:

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

- *Normalized Discounted Cumulative Gain* (nDCG) [125] – Mede a qualidade da ordenação dos resultados, tendo em conta a relevância dos itens e penalizando resultados relevantes que aparecem em posições mais baixas na lista. É obtido pela normalização do DCG pelo iDCG (ordenação ideal), permitindo comparar o desempenho do modelo com uma classificação perfeita.

Uma classificação perfeita resultaria num nDCG de 1, enquanto qualquer desvio da classificação ideal resultaria num nDCG inferior a 1.

No entanto, para avaliar os modelos de linguagem, estas métricas não são suficientes [116]. Avaliar os LLMs é uma tarefa mais complexa, por causa das respostas dos modelos serem diversificadas e dependentes do contexto.

Assim, muitas vezes, a avaliação dos LLM é realizada por especialistas humanos. Esta é a melhor forma de conseguir ter uma avaliação exata, pois os especialistas conseguem analisar a

relevância das respostas e apontar os pontos fracos do modelo. Contudo, esta forma de avaliar é vulnerável a enviesamento, por estar sujeita ao julgamento humano e à variabilidade entre anotadores. Para reduzir este efeito, seria necessário recorrer a múltiplos avaliadores (utilizadores e peritos), no entanto, a recolha destas avaliações costuma ser demorada e consumir muitos recursos, tornando-se difícil uma avaliação a larga escala [126].

Devido a isso, os métodos de avaliação dos LLMs devem equilibrar o julgamento humano, com a eficiência e rapidez dos métodos automáticos. Assim, surgem os *LLM-as-a-judge*, que procuram integrar o melhor destes dois métodos de avaliação [126].

2.6.1 LLM-as-a-Judge

Os *LLM-as-a-Judge* surgem para substituir as avaliações realizadas por humanos e as métricas estatísticas. Estes modelos geram avaliações interpretativas, que permitem ter uma avaliação mais abrangente sobre como funciona o modelo, e consegue-se perceber os pontos fortes e fracos do modelo. Além disso, os *LLM-as-a-Judge*, oferecem uma forma escalável e semelhante à avaliação realizada por especialistas humanos, mas sem o problema de elevados custos e demoras [126], [127].

Gu et al. [126] propõem um guia prático para a aplicação do *LLM-as-a-Judge*, organizado em quatro etapas principais.

- **Fase de Planeamento**
Delimitar o objetivo de avaliação, como a avaliação seria conduzida por humanos e que critérios utilizariam. Reunir exemplos confiáveis de avaliações realizadas por humanos, que possam servir de referência.
- **Design do prompt.**
Construir instruções explícitas para o LLM avaliador, incluindo critérios de avaliação, escala de avaliação e formato de saída, entre outros. Gu et al. [126] apresentam diferentes estratégias para estruturar estes *prompts*, consoante o objetivo da avaliação. Uma abordagem frequente consiste na utilização de escalas numéricas, nas quais o modelo atribui uma pontuação a diferentes critérios, como a utilidade, relevância, precisão ou nível de detalhe da resposta (por exemplo, numa escala de 1 a 10). Outra estratégia baseia-se em julgamentos binários, em que o modelo responde apenas com "sim" ou "não" a uma determinada questão, como verificar se a informação disponibilizada é suficiente para responder à pergunta. Também são utilizadas abordagens de comparação par a par, nas quais o modelo compara duas respostas e identifica qual apresenta melhor qualidade, sendo esta estratégia útil para estabelecer hierarquias ou ordenar diferentes alternativas. Por fim, podem ser utilizados *prompts* de múltipla escolha, em que o modelo seleciona a opção mais adequada entre um conjunto de respostas previamente definidas.

- Escolher o modelo.
Escolher o modelo mais adequado à tarefa, privilegiando a capacidade de raciocínio e a capacidade de seguir instruções complexas.
- Padronizar a saída.
Garantir que os resultados são claros e comparáveis, usando um formato de *output*. Desta forma, evita respostas vagas e facilita a análise automática. Para esse efeito, podem ser utilizados diferentes formatos de saída, como escalas numéricas (por exemplo, de 1 a 5), respostas binárias (por exemplo, "Sim/Não" ou "Correto/Incorreto") ou formatos estruturados, como JSON, quando a avaliação é integrada em processos automáticos.

2.6.2 Métricas específicas para RAG

Neste âmbito, a qualidade das respostas geradas em sistemas de RAG é normalmente avaliada através de três dimensões principais: relevância contextual, relevância da resposta e confiança na resposta. Estas métricas permitem analisar, respetivamente, se a informação recuperada é adequada ao contexto da pergunta, se a resposta gerada responde corretamente à questão colocada e se a resposta é sustentada por evidências fiáveis provenientes das fontes recuperadas [99].

- Relevância de Contexto

A relevância de contexto mede o grau em que os documentos recuperados estão relacionados com a consulta inicial do utilizador. Essa avaliação considera fatores como o alinhamento temático, a utilidade das informações apresentadas e a presença de redundância no conteúdo recuperado [99].

Uma abordagem utilizada é definir a razão entre número de frases do contexto consideradas relevantes para responder à pergunta e o número total de frases recuperadas. Outra abordagem, o cálculo do número de frases relevantes pode ser feito através da soma dos scores de similaridade semântica entre cada frase e a pergunta do utilizador [99].

- Confiança na Resposta

A confiança na resposta refere-se ao grau em que a resposta gerada está fundamentada no contexto recuperado. Esta métrica serve para verificar se as informações apresentadas não contradizem o conteúdo original. Dessa forma, contribui para diminuir o problema de alucinações, no qual o LLM introduz informações não suportadas pelos dados disponíveis [99].

Este processo pode ser feito, primeiro identificando as frases presentes na resposta gerada e, de seguida, calcular a proporção dessas frases que estão efetivamente contidas no contexto recuperado. Quanto maior a correspondência entre as frases da resposta e o conteúdo do contexto recuperado, maior o valor atribuído à confiança na resposta. O inverso desta métrica, é a taxa de alucinação, que quantifica a proporção de frases geradas que não se encontram no contexto recuperado [99].

- Relevância da Resposta

A relevância da resposta avalia o grau em que a resposta gerada é pertinente para a pergunta do utilizador. Esta métrica tem em conta aspetos como a compreensão da intenção da consulta, a pertinência do conteúdo apresentado e a completude da resposta [99].

Para realizar esta avaliação, primeiro, é gerada uma resposta para a pergunta original do utilizador. De seguida, um LLM é instruído a formular múltiplas perguntas com base nessa resposta. A similaridade semântica, medida através da similaridade do cosseno, entre essas perguntas geradas e a pergunta original é então calculada, sendo utilizado o valor médio dessas similaridades como métrica final. Esta avaliação baseia-se na premissa de que, se a resposta abordar corretamente a questão inicial, então as perguntas derivadas deverão ser semanticamente próximas da pergunta original [99].

3 Revisão de Literatura

Este capítulo apresenta uma revisão da literatura sobre os métodos utilizados para apoiar a atualização do conhecimento dos profissionais de saúde. Para isso, procede-se à análise de artigos científicos que existem, com o objetivo de reunir, avaliar e sintetizar os principais métodos atualmente disponíveis.

A evolução do conhecimento científico deve-se ao número crescente de estudos e tratamentos inovadores que se realizam ao longo dos anos. Esta evolução é muito importante para melhorar a qualidade do atendimento médico, mas também dificulta o trabalho dos profissionais de saúde. Estes precisam de identificar rapidamente as informações que são mais relevantes para a prática clínica. Neste contexto, têm surgido soluções digitais com o propósito de facilitar a filtragem, recuperação e atualização científica dos clínicos.

3.1 Metodologia

Esta revisão sistemática tem por objetivo identificar os desafios enfrentados pelos médicos na prática clínica, nomeadamente no que diz respeito à necessidade de atualização do conhecimento científico. Para tal, adotou-se uma revisão sistemática da literatura, guiada pelas diretrizes de Kitchenham [128].

3.1.1 Perguntas de Pesquisa

Para compreender de que forma a tecnologia pode melhorar o conhecimento dos médicos e, assim, melhorar o atendimento prestado às pacientes, define-se como questão principal:

“De que forma têm sido aplicadas tecnologias digitais e técnicas de Inteligência Artificial para apoiar a análise de literatura científica e a atualização do conhecimento médico na saúde?”

Para responder esta questão principal, torna-se importante responder a estas subquestões:

RQ1: “Quais sistemas ou soluções baseadas em tecnologias digitais têm sido aplicados na área da saúde, com o objetivo de apoiar a atualização do conhecimento médico e acelerar a análise de literatura científica?”

RQ2: “Quais tecnologias e técnicas de Inteligência Artificial têm sido aplicadas na análise de literatura científica em contextos clínicos?”

RQ3: “Quais são os principais desafios identificados na aplicação dessas tecnologias, particularmente no contexto da saúde?”

A crescente complexidade do conhecimento médico e o aumento exponencial das publicações científicas representam desafios relevantes para a prática clínica. Entre 2016 e 2022, o número anual de artigos científicos publicados globalmente cresceu cerca de 47%, passando de 1,9 para 2,8 milhões. Este crescimento resulta de fatores como a expansão de instituições de ensino e investigação, o aumento de financiamento, os avanços tecnológicos e a intensificação da colaboração internacional [129]. Embora este cenário reflita um progresso importante na geração de conhecimento, também levanta preocupações relativamente à capacidade dos profissionais de saúde em acompanhar e integrar de forma eficiente esta vasta quantidade de informação na sua prática clínica diária.

Tendo isto em conta, a primeira questão procura identificar quais os sistemas ou soluções baseadas em tecnologias digitais têm sido aplicados na área da saúde, com o intuito de perceber quais os métodos de tecnologias que existem, e de forma geral, como funcionam e qual o desempenho.

O rápido desenvolvimento da AI e a sua crescente adoção na área da saúde têm impulsionado o desenvolvimento de ferramentas para análise automática da literatura científica. Neste contexto, torna-se relevante compreender quais as técnicas e abordagens utilizadas por estas ferramentas, sendo apresentada nesta questão um resumo das técnicas identificadas.

Apesar das vantagens associadas à utilização de tecnologias digitais na área da saúde, a sua implementação continua a enfrentar diversos desafios. Questões relacionadas com a ética e a privacidade dos dados, bem como a resistência dos profissionais de saúde à adoção de novas tecnologias, constituem algumas das principais barreiras à aceitação e integração destas soluções na prática clínica [130]. Neste contexto, a terceira subquestão desta investigação centra-se na identificação dos principais obstáculos relacionados com a aplicação de tecnologias digitais no contexto clínico, com o objetivo de oferecer uma visão abrangente sobre os fatores que ainda dificultam a sua plena integração na prática médica.

Em conjunto, estas três subquestões permitem, fazer uma análise abrangente do estado atual da utilização de tecnologias digitais e de AI no apoio à prática clínica. Com esta abordagem pretende-se compreender não apenas o que tem sido feito até ao momento, mas também identificar lacunas e falhas nos sistemas atuais para perceber como desenvolver soluções mais eficazes para os profissionais de saúde.

3.1.2 Definições de Estratégia de Pesquisa

A metodologia de pesquisa desta revisão sistemática de literatura está estruturada em três fases, garantindo a transparência e a rastreabilidade de todas as decisões metodológicas. Na primeira fase, selecionam-se as fontes de informação que serão usadas no estudo; na segunda fase, definem-se os termos de pesquisa e elaboram-se as *queries*; na terceira e última fase, selecionam-se os estudos e retiram-se os dados essenciais. Esta organização sequencial assegura não só a reprodutibilidade do processo, mas também minimiza potenciais vieses e maximiza a identificação dos resultados mais relevantes.

- Definição das Fontes de Informação

A primeira fase é a identificação das fontes de informação que serão usadas para o presente estudo. A seleção destas fontes de informação deve garantir representatividade e relevância dos dados. Para esta revisão sistemática da literatura foram selecionadas três bases de dados disponíveis online e frequentemente usadas para procurar informações sobre pesquisas científicas e médicas.

Tabela 1: Bases de Dados Online

Base de dados	URL
Pubmed/MEDLINE	https://pubmed.ncbi.nlm.nih.gov/
Arxiv	https://arxiv.org/
Web of Science	https://www.webofscience.com/

- Definição dos termos de pesquisa

A segunda fase desta revisão sistemática de literatura consiste na formulação de um conjunto de termos de pesquisa que permitam responder à questão principal.

Tabela 2: Termos de Pesquisa

Termos de Pesquisa	
Medicina	("medicine" OR "clinical" OR "healthcare" OR "health care") AND
Análise de documentos	("research papers" OR "information retrieval" OR "document analysis" OR "scientific literature") AND
Tecnologia e IA	("digital technology" OR "semantic analysis" OR "semantic indexing" OR "artificial intelligence" OR "generative AI" OR "text mining" OR "machine learning" OR "deep learning" OR "natural language processing")

- Processo de seleção de estudos e extração de dados

Durante o processo de seleção dos estudos e de extração dos dados, procedeu-se à filtragem dos artigos não relevantes para este estudo, com o intuito de reter apenas aqueles que respondem às três sub-questões, de forma clara e fundamentada. Para tal, foram definidos critérios de inclusão e exclusão, apresentados nas Tabelas 3 e 4. Posteriormente, procedeu-se à selecção dos estudos e à extração sistemática dos dados com base no modelo PRISMA

(*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*). Este modelo organiza o processo em quatro fases distintas: identificação, triagem, avaliação da elegibilidade e, por último, inclusão dos estudos considerados relevantes para a análise final.

Os critérios de inclusão (Tabela 3) definem as condições essenciais para que os estudos sejam considerados nesta revisão literária. Devem ser incluídos estudos que apresentem evidência empírica proveniente de experiências controladas, preferencialmente ensaios clínicos randomizados, que demonstrem a eficácia de soluções tecnológicas. Adicionalmente, os estudos deverão abordar tecnologias aplicadas à análise de literatura científica e estar inseridos no domínio da saúde.

Tabela 3: Critérios de inclusão de artigos

Critérios de Inclusão:
IC1. Evidência empírica derivada de estudos com controlo de variáveis, preferencialmente ensaios clínicos randomizados, que demonstrem a eficácia ou impacto de soluções tecnológicas em cenários de saúde.
IC2. Inclusão de sistemas computacionais ou soluções baseadas em Inteligência Artificial, NLP, Mineração de Texto ou Modelos de linguagem, com aplicação direta à automação/análise de literatura científica.
IC3. Foco em domínios da saúde, com especial relevância para obstetrícia, saúde materna, contextos clínicos relacionados.

A Tabela 4 apresenta os critérios de exclusão aplicados aos artigos, detalhando as condições que levaram à desqualificação dos estudos para esta revisão. Para assegurar a atualidade e a acessibilidade dos dados, foram incluídos apenas estudos publicados entre janeiro de 2020 e junho de 2025, em inglês. Foram também excluídos artigos cuja versão integral não estivesse disponível, assim como aqueles que abordassem soluções tecnológicas fora do domínio da saúde ou que não se centrassem na automação, análise de literatura científica ou em técnicas de extração de texto.

Tabela 4: Critérios de exclusão de artigos

Critérios de Exclusão:
EC1. Artigos duplicados.
EC2. Estudos publicados em idiomas diferentes do inglês.
EC3. Trabalhos publicados antes de janeiro de 2020 ou após junho de 2025.
EC4. Artigos sem acesso completo ao texto.
EC5. Artigos de revisão, opiniões ou editoriais.
EC6. Artigos que não estão relacionados com contextos clínicos, especialmente fora das áreas sobretudo os que não tenham ligação com saúde e obstetrícia.

A Figura 4 ilustra o diagrama PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*), que representa visualmente o processo de seleção dos estudos incluídos nesta revisão literária. Este diagrama detalha cada uma das etapas principais do método PRISMA, onde em cada fase é descrita o número de artigos considerados e excluídos, proporcionando mais transparência neste processo de seleção.

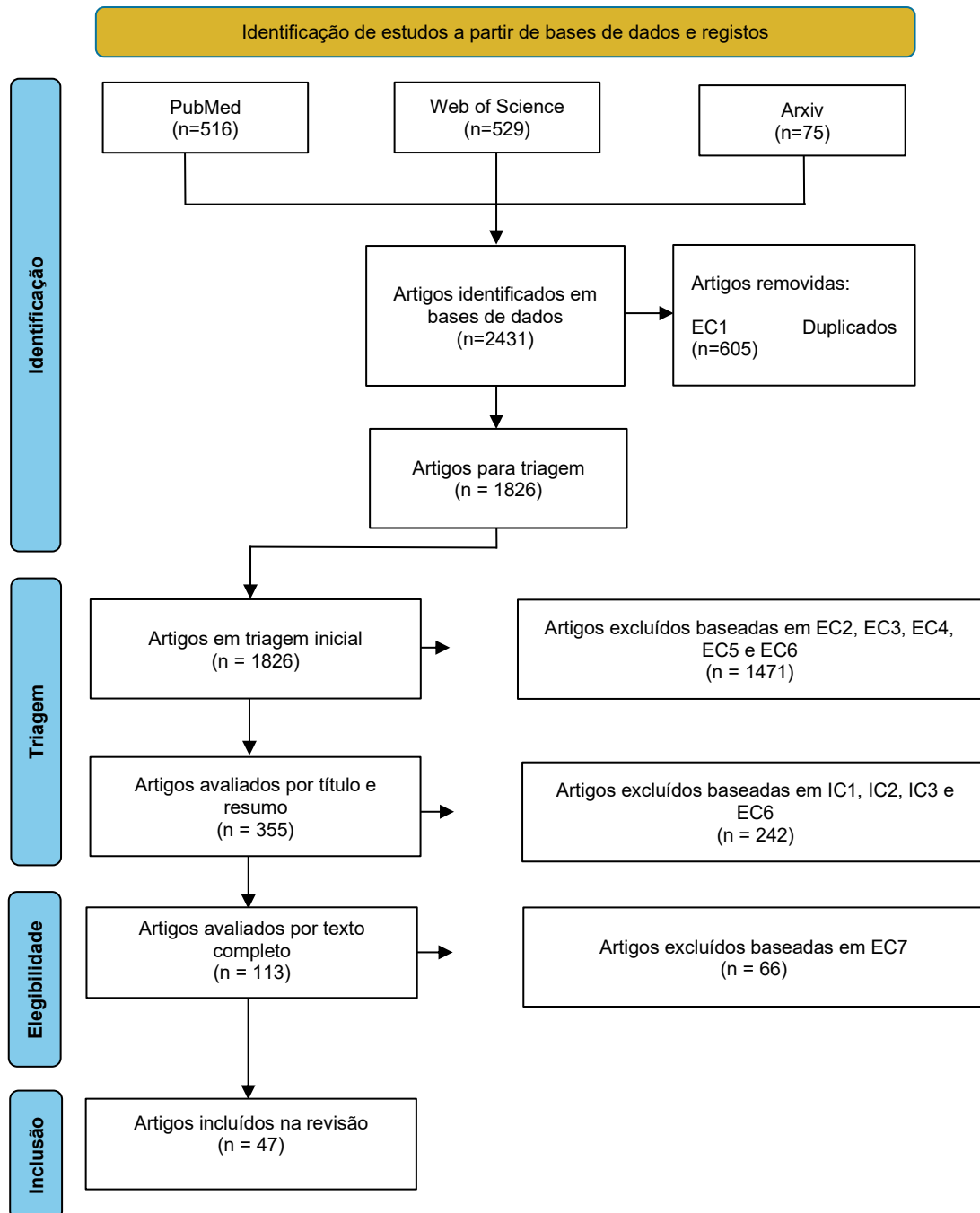


Figura 4: Fluxograma PRISMA do processo de identificação, seleção, elegibilidade e inclusão dos estudos.

3.2 Resultados

Nesta secção, procedeu-se a uma análise dos 47 artigos seleccionados anteriormente, para responder às subquestões desta revisão literária. Pretende-se analisar cada um dos artigos individualmente, de forma a identificar padrões e lacunas relevantes para este estudo. E por fim, obter um estudo que oriente investigações futuras neste domínio.

3.2.1 Questão 1: Quais sistemas ou soluções baseadas em tecnologias digitais têm sido aplicados na área da saúde, com o objetivo de apoiar a atualização do conhecimento médico e acelerar a análise de literatura científica?

A atualização científica dos profissionais de saúde é muito importante para a prática clínica. Todos os dias, novos artigos científicos são publicados, trazendo descobertas e recomendações que podem influenciar o diagnóstico, tratamento e tomada de decisão clínica. No entanto, devido ao elevado número de publicações, os profissionais de saúde têm dificuldades em acompanhar os avanços mais recentes nas suas áreas de atuação. Como resposta a este problema, têm surgido várias tecnologias para facilitar a pesquisa e análise da literatura científica.

A partir dos artigos analisados, foi possível identificar três categorias principais de tecnologias utilizadas no apoio à atualização do conhecimento médico, nomeadamente: pesquisa avançada e ordenação de artigos científicos; identificação de relações entre entidades biomédicas; e, agentes conversacionais.

- Pesquisa Avançada e Ordenação de Artigos Científicos

O primeiro grupo de sistemas de pesquisa avançada e ordenação de literatura científica, permitem ao utilizador submeter uma consulta, sob a forma de palavras-chave, pergunta estruturada ou texto livre, e obter uma lista de artigos científicos relevantes.

As primeiras plataformas desenvolvidas para responder a este problema focaram-se na implementação de mecanismos de filtragem e reordenação de artigos científicos. Como exemplos destacam-se a *Europe PubMed Central* (Europe PMC) [131] e a *Health Information Research Unit* (HiRU) [132]. Estas ferramentas permitem a pesquisa de artigos científicos com base em palavras-chave, recorrendo a filtros simples, como autor, ano de publicação ou tipo de estudo. Os resultados da pesquisa também podem ser organizados segundo diferentes critérios, tais como relevância ou data de publicação do artigo.

Com o progresso das técnicas de IA, os sistemas de pesquisa passaram a integrar mecanismos automatizados de ordenação e priorização de literatura científica. Desta forma, os artigos científicos passaram a ser apresentados segundo a relevância que têm para responder a uma pesquisa específica. Neste contexto, a *RetriLite* [133] é uma aplicação desenvolvida para recuperar artigos biomédicos relacionados com genes ou medicamentos. A partir da pesquisa submetida pelo utilizador, a aplicação gera automaticamente *queries* estruturadas para retornar os artigos mais relevantes da base de dados MEDLINE, com destaque nas palavras-

chaves identificadas em cada documento. Além disso, apresenta metadados estruturados, que incluem informações clínicas importantes, como o tipo de tumor, os medicamentos em estudo e os potenciais efeitos adversos. Os autores Zhang *et al.* [134] propõem um algoritmo para recuperar artigos biomédicos cujo conteúdo contenha informações semelhantes em relação a uma pesquisa sobre uma doença.

Durante a pandemia de COVID-19, foram desenvolvidos diversos sistemas especializados na recuperação de literatura científica relacionada com esta temática. Um exemplo é o CO-Search [135], que funciona como um motor de busca que retorna publicações biomédicas consideradas relevantes para a pesquisa.

Para além destas abordagens baseadas em listar os artigos científicos, algumas soluções tecnológicas oferecem representações mais acessíveis e interativas. Para isso, recorrem a diferentes representações gráficas como grafos e mapas de calor (*heatmap*). Por exemplo, o COVID-19 Explorer [136], para além da lista de artigos, mostra um mapa semântico de artigos científicos (Figura 5) que permite compreender melhor a proximidade entre diferentes publicações que têm a mesma temática. Outros sistemas, como o KnowVID-19 [137], organizam os artigos em tabelas filtráveis e representações gráficas que ajudam a explorar as relações entre temas e subtemas de investigação.

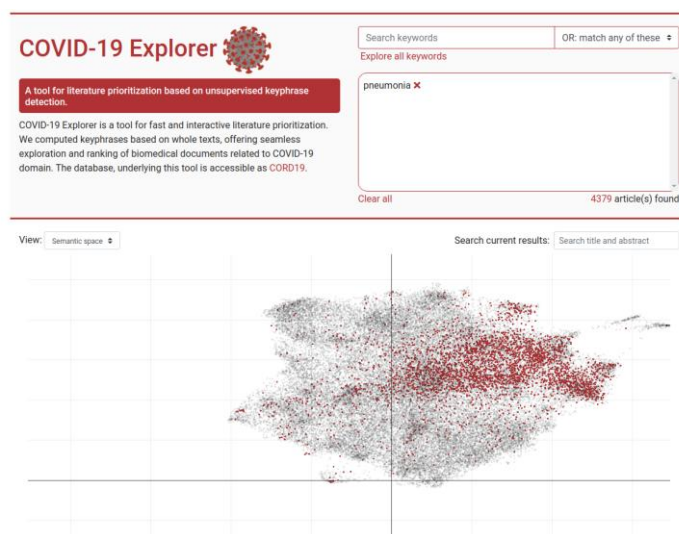


Figura 5: Mapa semântico de artigos científicos da aplicação *COVID-19 Explorer*. Adaptado de: [136].

De forma semelhante, o KnowCOVID-19 [138] apresenta listas de artigos relevantes, acompanhadas de tabelas estruturadas, nas quais os artigos são classificados de acordo com a pirâmide de níveis de evidência, um modelo hierárquico da medicina que organiza os estudos de acordo com a confiança científica que têm, desde relatos de casos até revisões sistemáticas. Adicionalmente, os artigos são categorizados por áreas clínicas, como diagnóstico e tratamento.

Zerva *et al.* [139] apresentam um motor de busca para a pesquisa de literatura científica relacionada com a COVID-19 e gráfico de bolhas, onde aparecem os termos mais relacionados com a pesquisa (Figura 6). De forma semelhante, o TopicTracker [140] integra diferentes tipos

de visualização para a exploração de literatura científica relacionada com a COVID-19, incluindo mapas de redes semânticas que representam os conceitos biomédicos e as relações semânticas entre eles.

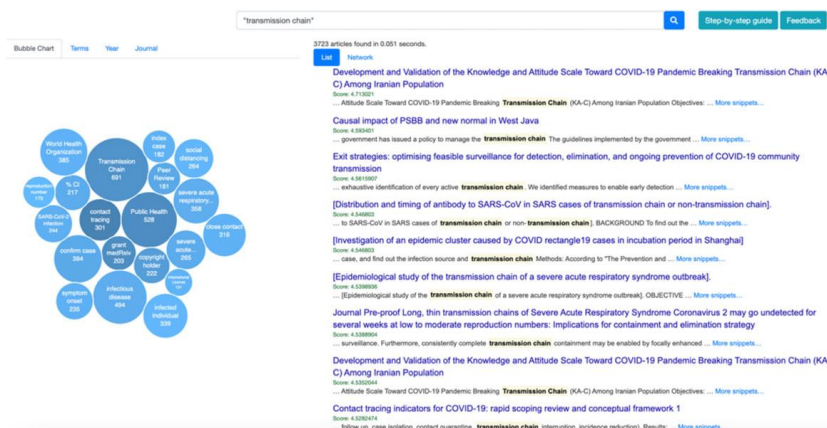


Figura 6: Gráfico de bolhas e lista dos artigos mais relevantes apresentados pela aplicação.

Adaptado de: [139].

Urru *et al.* [141] e Abd-alrazaq *et al.* [142] que organizam a literatura sobre a COVID-19 em diferentes tópicos, contribuindo para uma identificação mais rápida de estudos sobre terapias, vacinas e protocolos clínicos.

- Identificação de Relações entre Entidades Biomédicas

A segunda categoria engloba tecnologias focadas na identificação de relações entre entidades biomédicas. Trata-se de abordagens mais específicas, que permitem identificar associações entre conceitos como doenças, fármacos, genes ou biomarcadores. A partir da pesquisa de uma entidade biomédica específica, o sistema retorna outras entidades semanticamente relacionadas, frequentemente acompanhadas do tipo de relação identificada (por exemplo, “causa”, “trata”, “associado a”).

Um exemplo deste tipo de aplicações é a FooDis [143], desenvolvida para identificar relações entre alimentos e doenças, classificando se o alimento atua como causa ou tratamento, e apresentando as frases dos artigos biomédicos que apoiam essas relações. De forma semelhante, o MRAgent [81] apoia os profissionais de saúde na investigação de doenças. Esta ferramenta analisa autonomamente a literatura científica para identificar pares de exposição-desfecho, mais especificamente, relações entre fatores de risco e doenças reportadas em diversos estudos clínicos.

O ClinIQIR [144] tem como finalidade reduzir erros de diagnóstico através de uma abordagem baseada em recuperação de informação. O sistema analisa descrições clínicas de pacientes e estabelece correspondências semânticas com a literatura biomédica do PubMed, recuperando artigos que descrevem casos ou contextos clínicos semelhantes. Esta informação é assim, utilizada para apoiar no diagnóstico clínico.

Bauer *et al.* [145] procuram facilitar a análise de grandes volumes de literatura científica ao encontrar relações entre os 30 tipos de cancro mais frequentes e 270 fármacos utilizados nos

seus tratamentos. Os resultados são apresentados num grande mapa de calor (*heatmap*) com a relação entre estas duas entidades biomédicas, facilitando o acesso à informação e permitindo que os médicos se mantenham atualizados quanto a novos instrumentos de avaliação, recomendações clínicas e avanços na prática médica, para cada tipo de cancro.

No artigo Hsueh *et al.*[146] desenvolveram um gráfico de conhecimento interativo, onde os profissionais de saúde podem explorar as relações entre entidades biomédicas relacionadas com tratamentos/doenças existentes ou desconhecidos. A abordagem proposta organiza a informação sob a forma de grafos, nos quais os nós representam entidades biomédicas e as arestas correspondem aos diferentes tipos de relação identificados na literatura científica.

A figura 7 apresenta um exemplo dos grafos gerados, evidenciando a interligação entre múltiplas entidades. Por exemplo, a entidade “*Poor symptom control*” encontra-se associada à entidade “*increased asthma*” através da relação “*associated with*”. Desta forma, pode-se identificar facilmente padrões, dependências e possíveis implicações clínicas.

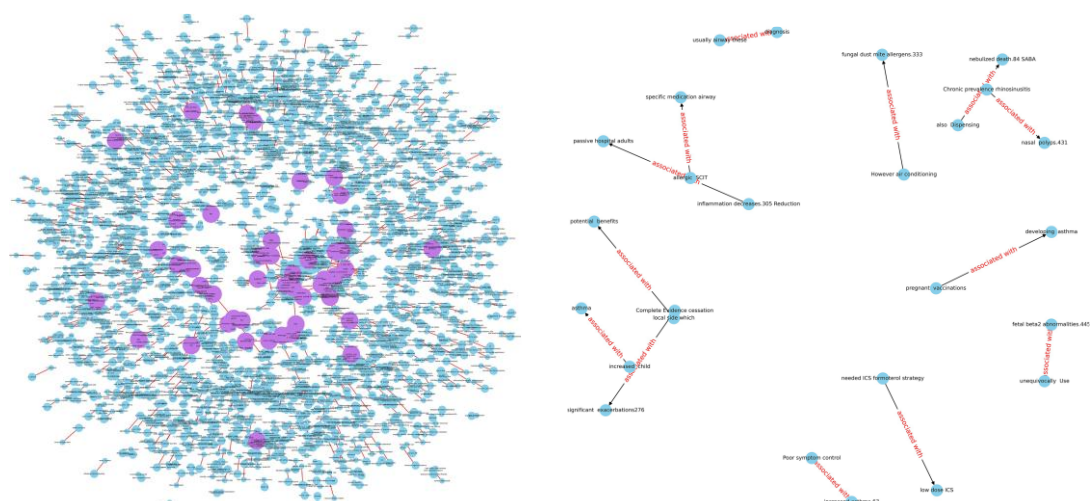


Figura 7: (a) Grafo de conhecimento gerado com restrição à entidade de entrada, apresentando ligação a apenas uma entidade biomédica; (b) subgrafo de conhecimento construído com base na relação “*associated with*”. Adaptado de [122].

Franklin *et al.* [147] abordam a necessidade de compreender a aplicabilidade dos estudos de coorte à prática clínica. Para esse efeito, propõem uma ontologia de coortes de estudo que permite padronizar a representação das características dos grupos de pacientes. O sistema automatiza a extração de informação a partir de tabelas presentes em publicações científicas, transformando esses dados em grafos de conhecimento utilizáveis em sistemas de suporte à decisão clínica. A abordagem é avaliada com base em 41 publicações científicas em formato PDF citadas nas diretrizes de 2019 da *American Diabetes Association*, permitindo aos profissionais de saúde avaliar rapidamente a relevância dos estudos para os seus próprios pacientes.

- Agentes Conversacionais

A terceira categoria corresponde a agentes conversacionais que permitem ao utilizador formular questões em linguagem natural e obter respostas personalizadas geradas por um LLM. Nestes casos, a resposta obtida é personalizada, gerada pelo agente conversacional e baseada em informação extraída de publicações científicas. Estes sistemas combinam, frequentemente, mecanismos de recuperação de informação e modelos generativos, permitindo fundamentar as respostas em evidência científica previamente indexada. Em muitos casos, incluem ainda a referência explícita às fontes utilizadas, promovendo maior transparência e rastreabilidade da informação apresentada.

Vários trabalhos têm explorado esta abordagem em domínios clínicos específicos. O AlzheimerRAG [148] é um agente conversacional baseado em RAG, destinado ao apoio à análise de casos clínicos, com foco em estudos de caso sobre a Doença de Alzheimer publicados na PubMed. Trata-se de uma aplicação multimodal, integrando dados textuais e visuais, o que possibilita a recuperação de informação biomédica relevante de acordo com a questão colocada pelo utilizador.

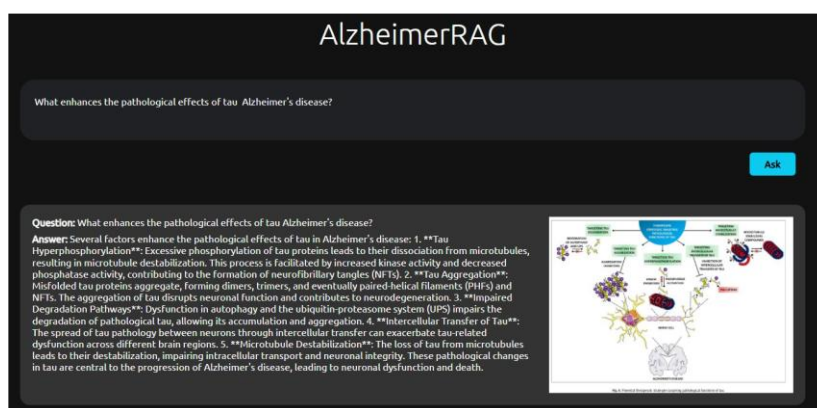


Figura 8: Agente Conversacional AlzheimerRAG. Adaptado de [148].

De forma semelhante, o EyeGPT [149] é um agente conversacional especializado na área da Oftalmologia. Neste sistema, foram utilizados conjuntos de dados públicos para incorporar terminologia específica do domínio oftalmológico. Adicionalmente, no contexto de RAG, o EyeGPT recorre a um conjunto de 14 livros de oftalmologia como fonte de conhecimento, permitindo fundamentar as respostas em literatura especializada.

O sistema HoPE [150] segue uma abordagem análoga, estando adaptado ao domínio da gravidez e saúde materna. Utiliza RAG para oferecer respostas contextualizadas com base em normas clínicas, incluindo informações sobre amamentação, nutrição, riscos, vacinas, entre outras.

Amirian *et al.* [151] propõem um agente conversacional para a área da reabilitação psicossocial, baseado em artigos científicos da PubMed, com o objetivo de oferecer respostas contextualizadas aos profissionais de saúde.

Yu *et al.* [152] avalia a possibilidade de utilizar o modelo LLaMA 2 para recomendar diagnósticos de COVID-19 em adultos. Nesta abordagem, os *Care Pathway Mayo* (CPMs), fundamentados em diretrizes clínicas, são utilizados como referência para a tomada de decisão. Através da análise do contexto clínico do paciente, incluindo sintomas e histórico de vacinação, o modelo aplica os percursos clínicos definidos nos CPMs, sugerindo as etapas clínicas apropriadas para cada situação.

Por fim, Gudivada *et al.* [153] propõe um modelo personalizado de apoio ao conhecimento médico, focado na área da diabetes, com o objetivo de auxiliar os profissionais de saúde na gestão de grandes volumes de informação. Para isso, o sistema é alimentado com dados específicos sobre diabetes, incluindo um livro dedicado à diabetes tipo 1 e dados públicos governamentais com informação sobre casos de diabetes. Esses dados são posteriormente integrados na plataforma *IBM Watson Discovery* e recuperados pelo agente conversacional.

3.2.2 Questão 2: “Quais tecnologias e técnicas de Inteligência Artificial têm sido aplicadas na análise de literatura científica em contextos clínicos?”

Na secção anterior foram apresentadas diversas soluções digitais que contribuem para a atualização contínua dos profissionais de saúde, através da análise automatizada de literatura científica. Contudo, para além da descrição funcional dessas aplicações, torna-se essencial compreender quais as tecnologias e técnicas de AI que sustentam o seu funcionamento, bem como de que forma a escolha dessas abordagens influencia a qualidade, relevância e precisão dos resultados obtidos.

Nesse sentido, este capítulo tem como objetivo aprofundar o conhecimento sobre as tecnologias e técnicas de AI utilizadas nas diferentes aplicações identificadas, analisando os métodos adotados.

De forma geral, a maioria das soluções analisadas integra técnicas de NLP, dado que a literatura científica biomédica é predominantemente textual e não estruturada. O NLP permite transformar grandes volumes de texto em representações estruturadas e analisáveis computacionalmente, automatizando tarefas que tradicionalmente exigiam leitura e interpretação manual.

As abordagens identificadas seguem, tendencialmente, um *pipeline* estruturado composto por quatro etapas principais: aquisição de documentos biomédicos; pré-processamento e normalização; indexação e recuperação de documentos e métodos avançados de análise textual.

Embora estas etapas sejam comuns à maioria dos sistemas, as técnicas específicas aplicadas variam significativamente entre aplicações, em função dos objetivos clínicos, do volume de dados considerado e dos requisitos de precisão. Assim, nas secções seguintes, apresenta-se uma análise detalhada de cada etapa, com o propósito de caracterizar as técnicas mais utilizadas e compreender o impacto das escolhas metodológicas nos resultados.

A aquisição de documentos biomédicos constitui a primeira etapa dos sistemas que utilizam NLP, sendo o PubMed/MEDLINE a fonte de dados mais utilizada. Em diversas aplicações, os artigos são recuperados por meio da biblioteca *Entrez Programming Utilities* (E-utilities) disponibilizada pelo *Nacional Center for Biotechnology Information* (NCBI). Esta biblioteca permite que as aplicações submetam pesquisas estruturadas, enriquecidas com filtros como ano de publicação, tipo de artigo ou idioma, e que essas aplicações recuperem resumos e metadados para processamento posterior [140], [143], [148], [154].

Para aumentar a quantidade de artigos recuperados e garantir que é utilizada a terminologia correta, algumas aplicações expandem as pesquisas dos utilizadores utilizando ontologias biomédicas e vocabulários médicos. Nos artigos revistos, utilizaram MeSH e UMLS na aplicação RetriLite [133], do UMLS Metathesaurus na aplicação CliniquIR [144] e do SNOMED-CT na aplicação CERC [155].

Uma estratégia alternativa consiste no *download* de grandes conjuntos de artigos para serem processados *offline*. Esta abordagem dá acesso completo a todo o artigo biomédico e permite que se utilizem análises computacionais mais detalhadas. Vários estudos também combinam múltiplas fontes de literatura, incluindo Google Scholar e Web of Science [146], [156].

Durante a pandemia da COVID-19, muitos trabalhos recorreram a conjuntos de dados específicos como o CORD-19 (COVID-19 Open Research Dataset), um repositório criado no início de 2020 e continuamente atualizado que agrega centenas de milhares de artigos biomédicos provenientes de diversas fontes, incluindo do bioRxiv e do medRxiv [135], [136], [138], [139], [157].

Além de artigos científicos, algumas aplicações optaram por utilizar *datasets* de diálogos médico-paciente [158] e, no caso específico do EyeGPT [149], conteúdos provenientes de livros didáticos de oftalmologia. Outros sistemas optaram por recorrer a *guidelines* clínicas como fonte de conhecimento, com o objetivo de assegurar o acesso a recomendações terapêuticas atualizadas e validadas para cada área específica. Um exemplo desta abordagem é a aplicação HoPE [150], que utiliza *guidelines* na área da obstetrícia e saúde materna para proceder à atualização do agente conversacional.

Após a recolha dos documentos, é realizada a etapa de pré-processamento e normalização do texto biomédico, com o objetivo de reduzir ruídos e preparar os textos para a análise automatizada. Essas etapas são importantes para os contextos biomédicos, nos quais os artigos contêm terminologia complexa, metadados estruturados e elementos não textuais.

Nas operações de pré-processamento é comum aplicar *stemming* ou lematização para reduzir palavras às suas formas canónicas [155], [141], [159] e [150] e remover *stop words* e caracteres irrelevantes (como números ou *tokens* muito curtos) [135]. Alguns sistemas excluem blocos de texto irrelevantes, como no sistema HoPE, onde excluem-se imagens e tabelas [150] e no sistema [159], que exclui metadados do PubMed (IDs de revista, dados de autores, termos não ingleses), recorrendo a bibliotecas em Python, como Gensim.

Após o pré-processamento, os documentos são convertidos em *embeddings*, isto é, representações vetoriais numéricas que codificam o conteúdo textual, e posteriormente indexados em bases de dados apropriadas. Esta etapa é muito utilizada em agentes conversacionais baseados em RAG, como o sistema apresentado em [151], o EyeGPT [149] e o AlzheimerRAG [148].

Nestes casos, a recuperação de informação baseia-se na proximidade entre os *embeddings* da pesquisa do utilizador e os *embeddings* previamente indexados. Este processo é geralmente realizado através do cálculo de medidas de similaridade entre os *embeddings* da consulta e os *embeddings* dos documentos armazenados em índices lexicais ou vetoriais. Os documentos ou parágrafos que obtêm as pontuações mais elevadas são então selecionados e utilizados pela aplicação para responder à consulta.

Neste contexto, as aplicações analisadas recorrem a uma das três abordagens de representação vetorial e recuperação: *sparse embeddings*, *dense embeddings* e abordagens híbridas que combinam ambas.

Sparse Retrieval é um exemplo clássico da área de *Information Retrieval*, baseado na correspondência lexical explícita entre a pesquisa do utilizador e os documentos da coleção. A recuperação é realizada pela comparação direta entre a pesquisa e os documentos. Entre as técnicas mais utilizadas destaca-se o TF-IDF e BM25 [133], [155].

Zeng *et al* [133] referem a utilização do método TF-IDF, uma técnica estatística que estima a relevância de um termo num documento com base na sua frequência local e raridade, atribuindo maior peso a palavras mais raras e informativas. Esta técnica é igualmente aplicada em CERC [155] e no trabalho apresentado por Bose *et al.* [159], onde é implementada a partir da biblioteca Apache Lucene.

A aplicação PIRO [160] e o trabalho apresentado por Zerva *et al.* [139] implementam a técnica BM25, um modelo probabilístico que introduz normalização em função do comprimento do documento, evitando que textos longos sejam favorecidos.

Dense Retrieval é uma abordagem de recuperação de informação baseado na representação que representa a pesquisa e os documentos através de *dense embeddings* aprendidos por modelos neuronais, permitindo medir a similaridade semântica no espaço vetorial. Um exemplo deste tipo de abordagem é o modelo Word2Vec, que aprende *dense embeddings* ao analisar os contextos em que as palavras ocorrem, de forma que palavras com significados semelhantes sejam mapeadas para vetores próximos no espaço vetorial. No trabalho realizado por Amirian *et al.* [151], este modelo é utilizado para recuperar os documentos mais similares com a pesquisa do utilizador. De forma semelhante, o trabalho de Bauer *et al.* [145] utilizou este modelo para explorar relações entre doenças e fármacos.

O modelo Doc2Vec estende a lógica do Word2Vec para gerar representações vetoriais não apenas de palavras, mas também de frases e documentos completos. O COVID-19 Explorer [136] utiliza este modelo para transformar cada artigo num *embedding*, ou seja, numa representação vetorial que resume o significado geral do documento. De seguida, a dimensionalidade dos

embeddings é reduzida utilizando o *Uniform Manifold Approximation and Projection*, e são posteriormente utilizados para criar mapas interativos onde os artigos semanticamente mais semelhantes aparecem mais próximos.

Os métodos mais recentes são os *embeddings* contextuais gerados por modelos baseados em transformes, como BERT, BioBERT e Sentence-BERT. Esses modelos atribuem representações distintas à mesma palavra dependendo do contexto em que ocorre, o que resulta numa maior precisão semântica. O HoPE [150] adota um pipeline de duas etapas usando *embeddings* contextuais: um *retriever* (*Sentence-BERT Paraphrase-Multilingual-MiniLM*) que indexa e recupera os documentos mais relevante.

Os métodos híbridos combinam *sparse embeddings* e *dense embeddings*, combinando as vantagens de ambos. Por um lado, recorrem a técnicas lexicais, como TF-IDF e BM25, que asseguram precisão na correspondência de termos específicos. Por outro lado, incorporam *contextual embeddings* gerados por modelos de linguagem pré-treinados, como BERT, BioBERT ou Sentence-BERT, desta forma captura relações semânticas e ultrapassa as limitações associadas à recuperação exclusivamente baseada em correspondência lexical.

O artigo [134] exemplifica uma abordagem híbrida que inicia a recuperação de informação com o BM25, aprimorado com termos MeSH, para selecionar os 1000 artigos mais relevantes. De seguida, aplica o algoritmo *k-means clustering* para agrupar frases semanticamente semelhantes dos resumos, selecionando a frase mais representativa em cada grupo e reduzindo a redundância da informação apresentada.

Seguindo uma estratégia semelhante, o CO-Search [135] combina, durante a indexação, três técnicas complementares: TF-IDF, BM25 e Sentece-BERT (SBERT). Na fase de recuperação, a consulta do utilizador é convertida em *embeddings* semânticos através do SBERT e é calculada a similaridade do cosseno entre a consulta e os documentos indexados. Em paralelo, são obtidas as pontuações produzidas pelo TF-IDF e pelo BM25, sendo as três classificações posteriormente combinadas através da técnica *Reciprocal Rank Fusion*, permitindo integrar evidencias lexicais e semânticas na ordenação final dos resultados.

Já o estudo [161], voltado para a medicina de precisão, combina os modelos probabilísticos BM25 e *Divergence From Randomness* com filtros demográficos, para recuperar e reordenar documentos de acordo com as características clínicas e demográficas do paciente. Estes modelos estimam a relevância dos documentos a partir da correspondência lexical entre os dados do paciente e o conteúdo dos documentos. Adicionalmente, o BERT cria *dense embeddings* que capturam a relação semântica entre a *query* e os documentos, permitindo eliminar resultados sem significado clínico relevante e reduzindo, consequentemente, a ocorrência de falsos positivos.

Por fim, o artigo [142] aplica a mesma lógica para identificar lacunas através do BERTopic. Este modelo gera *embeddings* semânticos com o Sentence-BERT e agrupa documentos semanticamente semelhantes, recorrendo posteriormente ao *class-based* TF-IDF (c-TF-IDF) para identificar os termos mais representativos de cada tópico. Desta forma, os *embeddings*

asseguram o agrupamento dos documentos de acordo com o seu significado, enquanto o c-TF-IDF facilita a interpretação dos tópicos identificados.

Alguns dos estudos analisados recorrem a métodos avançados de análise textual para extrair informação dos documentos biomédicos, através da identificação de entidades, relações e outros padrões semânticos. Entre os métodos mais utilizados destacam-se as técnicas de *Named Entity Recognition* (NER) e *Relation Extraction* (RE), bem como os métodos de *clustering* aplicados ao *topic modelling*.

As técnicas de NER e RE são utilizadas para identificar entidades biomédicas e as relações existentes entre elas, permitindo estruturar a informação clínica e revelar padrões e relações entre entidades biomédicas importantes para o conhecimento clínico. Dependendo do objetivo do estudo, alguns artigos revistos concentram-se apenas na identificação das entidades biomédicas que ocorrem no texto, enquanto outros incluem a normalização dessas entidades em ontologias biomédicas (por exemplo, MeSH e UMLS) e a extração de relações entre entidades. Para realizar esta tarefa, as abordagens utilizadas variam entre a utilização de modelos supervisionados, *transformers* e, mas recentemente, LLMs com *prompt engineering*.

O FooDis [143] é um *pipeline* de extração de informação dos resumos de artigos científicos do PubMed, tem como objetivo identificar automaticamente relações de causa ou tratamento entre entidades alimentares e doenças. O processo inicia-se com a extração de entidades relacionadas com doenças, utilizando a ferramenta SABER (*Sequence Annotator or Biomedical Entities and Relations*) e com a extração de entidades alimentares, utilizando métodos NER, como o BuTTER e o FoodNER. De seguida, selecionam as frases que contenham uma entidade alimentar e uma doença, e terminam com a determinação da relação que existe entre essas entidades, distinguindo relações de causa, tratamento ou ausência de relação, utilizando o modelo SAFFRON, um modelo de extração de relações semânticas.

O RetriLite [133] é uma ferramenta que aplica técnicas de NER e de análise de relações entre entidades para refinar os resultados da recuperação de artigos científicos. O NER é utilizado no início para identificar e normalizar entidades relevantes, tais como genes, medicamentos e conceitos relacionados com cancro. Por fim, utilizam a biblioteca spacy para analisar a coocorrência de entidades na mesma frase, uma forma de inferir associações entre as entidades e reduzir os falsos positivos.

Da mesma forma, Hsueh et al. [146] usa a biblioteca spaCy para extrair tripletos Sujeito–Relação–Objeto. Neste caso, a ideia é construir *knowledge graphs* médicos, capazes de extrair, organizar e refinar o conhecimento clínico (doenças, diagnósticos, tratamentos, medicamentos) a partir de artigos científicos.

O MedScaleNER [162] usa LLMs com *prompts* para extrair entidades ligadas a escalas médicas em artigos científicos em chinês. Dado um texto, o modelo identifica menções relevantes e rotula-as em três categorias: nomes de escalas (p. ex., “MMSE”, “HADS”), conceitos avaliados, e itens/descriptores de medida. O sistema adota técnicas de *few-shot learning* para mostrar ao

modelo alguns exemplos anotados que ilustram o resultado esperado, e usa *Chain-of-Thought prompting*, que incentiva ao processamento de passos intermédios de raciocínio.

Alguns artigos utilizam técnicas de *clustering* para organizar os dados de forma que elementos semelhantes estejam dentro de um mesmo grupo. Esta abordagem é utilizada no NLP, uma vez que permite estruturar grandes volumes de documentos textuais, que normalmente são dados não estruturados. Um caso em que é utilizado é para topic modelling, em que o objetivo é descobrir automaticamente os temas principais que aparecem num conjunto de textos.

Um algoritmo muito utilizado é o *Latent Dirichlet Allocation* (LDA), um modelo probabilístico que permite representar cada documento como uma mistura de tópicos, sendo cada tópico definido por uma distribuição de palavras [154]. Uma extensão deste modelo é o *Structural Topic Modeling* (STM), que é conceptualmente semelhante ao LDA, mas incorpora metadados dos documentos no processo de modelação. Para além disso o STM permite correlações entre diferentes tópicos, o contrário do LDA, que assume independência entre tópicos [141].

Outro exemplo é o *K-means*, que combinam *embeddings* gerados por modelos de linguagem pré-treinados com algoritmos de *clustering*. Um exemplo é o BERTopic, que utiliza *embeddings* obtidos a partir de modelos baseados em BERT para representar semanticamente os documentos, aplicando posteriormente algoritmos de *clustering* para agrupar textos semanticamente semelhantes. Nesta abordagem, cada cluster corresponde a um tópico, que é posteriormente interpretado através da extração dos termos mais representativos.

Urru et al [141] aplicaram o algoritmo STM aos resumos de artigos científicos sobre SARS-CoV-2 e COVID-19. Neste estudo analisam a frequência e coocorrência de palavras ao longo dos artigos para estimar a probabilidade de determinados conjuntos de palavras estarem semanticamente relacionados. Os autores testaram diferentes números de tópicos, variando entre 8 e 10, e selecionaram a solução mais adequada com base em critérios de qualidade específicos, nomeadamente a coerência semântica, que avalia se as palavras associadas a cada tópico são semanticamente consistentes com o respetivo tema, e a exclusividade, que verifica se as palavras de um tópico são distintas. Como resultado, identificaram um conjunto de 10 tópicos, aos quais foram atribuídos rótulos de acordo com as palavras mais frequentes e exclusivas, como “Impacto da COVID-19 na economia”, “Biologia molecular” e “Telemedicina”.

No estudo [154] sobre distúrbios da fala infantil, os autores realizam uma comparação entre técnicas de *topic modelling*, como LDA e BERTopic. O LDA mostrou rapidez e interpretabilidade, identificando tópicos clínicos como TDAH e estresse parental, embora o BERTopic tenha produzido tópicos mais diversificados, apesar do maior custo computacional.

Alguns dos estudos analisados descrevem agentes conversacionais concebidos para fornecer respostas fundamentadas em evidência clínica. Para esse efeito, as aplicações identificadas nesta revisão da literatura recorrem a arquiteturas de RAG ou modelos LLM ajustados (*fine-tuned*), integrando conhecimento proveniente da literatura científica. Os trabalhos analisados referem ainda Retalgumas técnicas aplicadas aos LLMs, que ajudam a melhorar a resposta obtida e a reduzir a probabilidade de ocorrer alucinações.

O AlzheimerRAG [148] é um agente conversacional baseado em RAG, que integra um modelo multimodal para gerar respostas fundamentadas em artigos científicos. Este *chatbot* utiliza um LLM ajustado ao domínio clínico, pois foi aplicada a técnica de *fine-tuning* com literatura biomédica e conjuntos de perguntas e respostas clínicas. É utilizado um modelo multimodal, que integra texto e imagens de artigos biomédicos. Os *embeddings* textuais e visuais são combinados através de um mecanismo de *cross-modal attention fusion*, que produz representações multimodais. Que serão armazenadas e utilizadas na fase de recuperação do modelo.

O HoPE é também um agente conversacional, que utiliza técnicas de NLP para identificar a intenção do utilizador e recuperar a informação da base de dados deste sistema. Este *chatbot* tem uma etapa pormenorizada no que diz respeito à interpretação da pergunta feita pelo utilizador (representado na Figura 9). Primeiro passa por uma etapa de reconhecimento da intenção do utilizador, onde é classificada a intenção da consulta e é atribuída uma pontuação de confiança. Quando a pontuação é baixa, aplica-se o módulo de gestão de confiança, que integra várias técnicas de NLP, extração e normalização de entidades com recurso a ontologias, e a pergunta é reescrita. A recuperação de informação é realizada através de modelos Sentece-BERT, combinando uma arquitetura *bi-encoder* para similaridade semântica com um *cross-encoder* para *re-ranking*, permitindo selecionar respostas mais precisas e contextualmente adequadas.

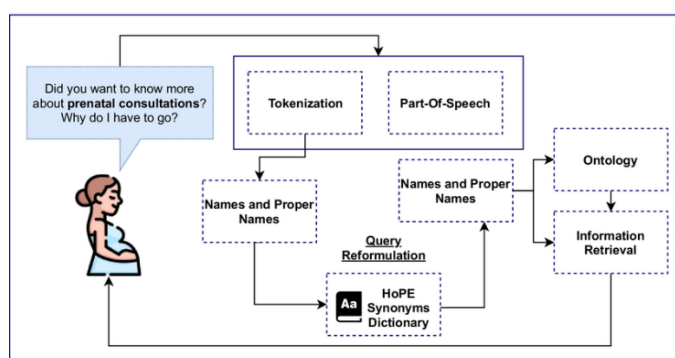


Figura 9: Arquitetura do sistema HoPE para processamento da pergunta do utilizador. Adaptado de [150].

Para além disso, o sistema HoPE por ser um modelo utilizado para língua portuguesa, compara a performance de três modelos multilingue, o BERTimbau, Paraphrase-Multilingual-MiniLM e BERT-Multilingual.

3.2.3 Questão 3: “Quais são os principais desafios identificados na aplicação dessas tecnologias, particularmente no contexto da saúde?”

A adoção de sistemas de atualização do conhecimento médico exige uma análise crítica das suas limitações. Dado que estas tecnologias operam em contextos de elevado risco, frequentemente associados à tomada de decisões clínicas e ao tratamento de dados sensíveis,

diversos estudos têm procurado identificar as suas limitações, compreender onde falham e apontar os aspetos em que necessitam de ser melhoradas.

No âmbito dos LLMs a literatura identifica diversas limitações que condicionam a sua utilização em ambiente clínico. Uma das mais relevantes está relacionada com os dados utilizados no treino do modelo, uma vez que o seu desempenho depende diretamente da qualidade, representatividade e diversidade desses dados. Neste contexto, vários estudos destacam que, quando os modelos são treinados com dados provenientes de populações específicas, podem apresentar dificuldades em generalizar para outros contextos clínicos. Este enviesamento pode resultar na produção de recomendações menos precisas ou inadequadas para doentes pertencentes a populações minoritárias [152], [163].

Outro aspeto a considerar na utilização de LLMs para apoio ao diagnóstico é a variação do estado clínico: a condição do doente pode alterar-se de minuto a minuto. Para que um LLM seja útil na prática clínica, deveria acompanhar o paciente em tempo real e assimilar uma série de dados multimodais, como exames imagiológicos, sinais vitais e análises laboratoriais [144], [158].

Os LLM podem gerar alucinações, ou seja, informações imprecisas, irrelevantes ou inventadas, sobretudo quando são expostos a dados desconhecidos ou interpretam a informação de forma incorreta [148], [149], [150], [152]. Observou-se, por exemplo, no MedpromptExtract [164] que o sistema fazia suposições que não eram sempre verdadeiras em contexto real, como por exemplo, assumir que se deveria utilizar anestesia geral para qualquer cirurgia.

Adicionalmente, estes modelos são frequentemente percebidos como “caixas-negras”, na medida em que o seu processo de tomada de decisão não é transparente e interpretável, o que dificulta a análise do seu comportamento e a identificação de erros [165].

Do ponto de vista ético e de privacidade, os sistemas devem cumprir regulamentos que protejam a privacidade do paciente. Isso implica práticas como, prevenir o acesso não autorizado ou o uso indevido de dados de saúde sensíveis [152], [166], [167].

Existe também um desafio relacionado com as avaliações das tecnologias que usam LLMs. Os autores Xu et al. referem que, geralmente, as avaliações são conduzidas em cenários hipotéticos, o que não reflete a complexidade da prática médica. Para além disso, referem que os LLMs tendem a ter bom desempenho em tarefas estáticas, como responder a perguntas para as quais foram treinados. Contudo, revelam dificuldades quando são adaptados ao mundo real, onde a resposta exige um raciocínio de múltiplas etapas. Na prática clínica, por exemplo, o diagnóstico de um doente não se baseia apenas na análise dos sintomas apresentados. O médico tem de considerar diferentes hipóteses de diagnóstico, planear a realização de exames complementares e interpretar os seus resultados para, só então, chegar a um diagnóstico final [158], [166]. Também é referido o facto de haver falta de avaliações para os LLMs, sendo que as avaliações realizadas por especialistas humanos são trabalhosas e não escaláveis, e as métricas automáticas não são tão eficientes quanto o julgamento humano [166].

Alguns artigos referem o risco de os profissionais de saúde se tornarem excessivamente dependentes de sistemas automatizados, levando a uma diminuição das suas próprias competências ou a uma excessiva confiança em diagnósticos ou informações geradas por AI [149], [163].

Parente [168] refere nos elevados requisitos computacionais necessários para executar um LLM. Os modelos de maior dimensão requerem um hardware especializado, nomeadamente GPUs de elevada capacidade de processamento e memória, sendo frequentemente disponibilizados apenas através de serviços na *cloud*. Esta abordagem implica custos adicionais associados à utilização destes modelos, que podem tornar-se significativos, quando existe um elevado número de interações ou conversas prolongadas no *chatbot*. Embora alguns modelos de menor dimensão possam ser executados localmente, esta opção pode implicar uma redução do desempenho quando comparada com os modelos de maior dimensão ou com soluções disponibilizadas na *cloud*.

3.3 Discussão de Resultados

Nesta secção procede-se à análise crítica e comparativa dos principais resultados identificados nos estudos incluídos na presente revisão da literatura.

3.3.1 Questão 1: Quais sistemas ou soluções baseadas em tecnologias digitais têm sido aplicados na área da saúde, com o objetivo de apoiar a atualização do conhecimento médico e acelerar a análise de literatura científica?

Com base na análise dos estudos incluídos na presente revisão sistemática, verifica-se um crescente desenvolvimento de sistemas e soluções digitais aplicados na área da saúde com o objetivo de apoiar a atualização do conhecimento médico e acelerar a análise da literatura científica. De forma geral, as tecnologias identificadas podem ser agrupadas em três grandes categorias: (i) plataformas baseadas em filtros e pesquisa estruturada em bases de dados, (ii) sistemas de recuperação avançada e análise semântica suportados por AI e (iii) agentes conversacionais baseados em modelos de linguagem de grande escala.

Os primeiros sistemas de pesquisa, como a Europe PubMed [131] e a *Health Information Research Unit* [132], apresentam maior confiança, porque a seleção é controlada pelo utilizador. Apresentam filtros (como por ano, autor, tipo de estudo) que facilitam pesquisas direcionadas. Contudo, estas ferramentas são muito dependentes da formulação de uma *query* correta e da experiência do utilizador. Pesquisas muito gerais, retornam um excesso de resultados, enquanto pesquisas demasiado específicas podem não retornar resultados relevantes. Além disso exigem a leitura e seleção manual dos artigos, tornando o processo demorado e trabalhoso.

Aplicações mais recentes, como RetriLite [133], CO-Search [135] e CliniqIR [144], representam uma evolução significativa nos sistemas de recuperação de informação biomédica. Estas

aplicações recorrem a técnicas de AI com o objetivo de identificar automaticamente os artigos mais relevantes para um determinado contexto clínico. Estas aplicações organizam e priorizam os resultados de forma automática, reduzindo o tempo e o esforço do utilizador. No entanto, o seu desempenho depende da qualidade e da quantidade de documentos disponíveis, podendo ser limitado quando utilizam apenas títulos e resumos ou quando aplicados a áreas com menor volume de literatura, como doenças raras ou temas pouco estudados.

Os agentes conversacionais baseados em modelos de linguagem de grande dimensão, como o EyeGPT [149], HoPE [150] e AlzheimerRAG [148], representam uma evolução nos sistemas de apoio à decisão clínica, ao permitirem interação em linguagem natural e geração de respostas contextualizadas. Estas soluções tornam a pesquisa mais acessível, pois não exigem a formulação de *queries* de pesquisa, e resumem informação proveniente de múltiplos artigos. Em alguns casos, como no AlzheimerRAG, incluem ainda integração multimodal, combinando texto e imagem para apoiar a análise clínica.

Apesar destas vantagens, estes sistemas apresentam desafios importantes. Existe o risco de geração de informação incorreta, especialmente quando as fontes não são suficientemente robustas ou completas. O seu desempenho depende fortemente da qualidade do corpus utilizado e pode ser afetado por ambiguidades linguísticas ou termos pouco frequentes, limitação identificada no HoPE [150]. Além disso, a aplicação em contexto clínico exige validação rigorosa, e o recurso a estes modelos implica, frequentemente, maior exigência computacional.

Por fim, os estudos que utilizam grafos referem limitações relacionadas com a escalabilidade e a facilidade de utilização. Em particular, é mencionada a dificuldade em interagir com grafos que apresentam muitas ligações entre os elementos, o que pode tornar a visualização confusa e difícil de interpretar [139], [146].

3.3.2 Questão 2: “Quais tecnologias e técnicas de Inteligência Artificial têm sido aplicadas na análise de literatura científica em contextos clínicos?”

Tal como referido anteriormente, nos últimos 5 anos, a maioria das técnicas aplicadas em ferramentas de análise de literatura científica, é baseada em IA. A análise destas tecnologias aplicadas à recuperação e processamento de informação biomédica evidencia avanços no processo de pesquisa e análise de artigos, mas também apresenta várias limitações que condicionam o desempenho das ferramentas em diferentes contextos.

No que diz respeito à recolha e organização inicial da informação, a maioria das ferramentas utiliza bases de dados com artigos científicos, principalmente a PubMed. Algumas ferramentas optam por usar ontologias [155], como forma de remover redundâncias (por exemplo, “AMI”, “*Acute Myocardial Infarction*”, “*heart attack*”, passaram a ter o mesmo conceito) e inconsistência de dados, desta forma, conseguem normalizar os termos clínicos e siglas durante o pré-processamento dos artigos científicos. Em relação à pesquisa de artigos utilizando apenas o título e resumo que aparece no site PubMed, os estudos referem que esta abordagem afetou

negativamente a recuperação, uma vez que determinados detalhes relevantes, como nomes específicos de medicamentos, surgem apenas no texto completo [133], [144], [145].

A utilização de normas clínicas é uma abordagem adotada por alguns dos sistemas médicos, para apoiar a tomada de decisão. Estas diretrizes garantem a prestação de cuidados baseados nas melhores evidências científicas disponíveis e no consenso de especialistas [152]. No entanto, estudos indicam que as normas clínicas podem tornar-se rapidamente desatualizadas, uma vez que a evolução contínua do conhecimento médico faz com que cerca de metade destas diretrizes perca a sua validade no prazo de cinco a seis anos [158].

As técnicas de pré-processamento (como tokenização, *stemming*, remoção de *stop words*) constituem a base dos sistemas de análise textual. A sua principal vantagem está no facto de normalizarem o texto e reduzirem o ruído, melhorando o desempenho dos modelos subsequentes.

Os métodos lexicais, como TF-IDF e BM25 destacaram-se por serem mais simples, mais eficientes computacionalmente pela simplicidade e mais transparentes, sendo por isso, ainda usados nos sistemas de recuperação de informação. No entanto, também se mostraram muito limitados, pois dependem de palavras-chaves exatas e têm dificuldades em lidar com sinónimos ou variações terminológicas ou mesmo relações implícitas entre conceitos [135]

De forma a ultrapassar estas limitações, surgem as abordagens baseadas em *dense embeddings*, como Word2Vec e doc2vec, que permitem ir além da simples coocorrência lexical, descobrindo ligações indiretas e hipóteses de associações que os métodos TF-IDF ou BM25 dificilmente captariam. Ainda assim, estas representações permanecem fixas, não conseguindo adaptar-se ao contexto em que os termos surgem [135]. Modelos mais recentes, baseados em arquiteturas Transformer como BERT, BioBERT e SBERT conseguiram atribuir significados distintos às palavras consoante o contexto. Esta adaptabilidade traduz-se num melhor desempenho em pesquisas clínicas complexas e um maior suporte em tarefas avançadas, como a modelação de tópicos. Em contrapartida, estes métodos implicam custos computacionais elevados [134], [135], [150], [157].

A combinação destes métodos de *embeddings*, foram testados no artigo [135]. Este artigo teve como resultado que confirma que a hibridização lexical-semântica constitui a melhor estratégia para recuperar conteúdo biomédico. Modelos como BM25/TF-IDF obtêm bons resultados quando a *query* contém terminologia específica, enquanto o SBERT amplia a capacidade de captar relações de significado mais abrangentes. A combinação destes modelos revelou-se superior, ao compensarem mutuamente as suas limitações [134], [135], [161]. No entanto, o sistema apresentou fragilidades em consultas mais abstratas e amplas, como “quais as diretrizes para triagem de pacientes infetados com coronavírus?” ou “o que causa morte por COVID-19?”. Nestes casos, a ausência de termos-chave claros dificulta tanto a correspondência lexical como a recuperação puramente semântica [135].

A modelagem de tópicos constitui outra linha relevante de investigação, permitindo organizar grandes volumes de literatura e identificar tendências emergentes. Os métodos clássicos, como

o *Latent Dirichlet Allocation* (LDA), oferecem rapidez, mas tendem a produzir tópicos pouco distintos. Por outro lado, modelos mais recentes como o BERTopic aumenta o nível de detalhe da análise e capta melhor as relações semânticas, resultando numa maior diversidade de tópicos, embora exija mais recursos computacionais [154].

Apesar dos benefícios demonstrados pelas abordagens híbridas e pela integração de múltiplas técnicas, importa salientar que a combinação de métodos de recuperação diferentes, pode não ser igualmente adequado para todos os contextos de pesquisa. Por exemplo, a utilização de uma recuperação híbrida (combinando BM25 e BioBERT) combinada com *clustering* (técnica de *k-means*), teve bons resultados em pesquisas exploratórias, no entanto, introduziu ruído quando é aplicado a pesquisas mais específicas, levando a que a aplicação tenha um desempenho inferior em pesquisas mais específicas, como aquelas relacionadas com genes, doenças raras ou características clínicas pouco descritivas [134].

A fase mais recente desta evolução tecnológica é marcada pela adoção de LLMs e arquiteturas baseadas em RAG. Estes sistemas permitem síntese de informação, raciocínio contextual e interação em linguagem natural, aproximando a análise automatizada da prática clínica. A integração de RAG demonstrou reduzir significativamente a ocorrência de alucinações [166]. Aplicações como o AlzheimerRAG reportam melhorias na recuperação e síntese de informação específica do domínio, com baixas taxas de alucinação (6%) [148]. A aplicação HoPE, também baseada em RAG, demonstrou características positivas, como boa capacidade preditiva numa velocidade razoável e capacidade de lidar com uma grande variedade de estruturas de frases na pesquisa [150].

No que diz respeito à engenharia de *prompts*, a literatura indica que a combinação de técnicas como *One-Shot* (fornecimento de um exemplo de saída) e *Few-Knowledge* (inclusão de informação adicional para orientar a análise) conduz a respostas mais completas e precisas, superando a aplicação isolada de cada técnica [156], demonstrando a importância de fornecer *prompts* detalhados aos modelos de linguagem.

3.3.3 Questão 3: “Quais são os principais desafios identificados na aplicação dessas tecnologias, particularmente no contexto da saúde?”

Em relação à questão 3, os resultados evidenciam limitações relevantes no uso de LLMs em contexto clínico e apontam, simultaneamente, soluções para diminuir essas mesmas limitações.

Um dos desafios apresentados eram sobre a dificuldade de os modelos responderem detalhadamente a questões mais complexas. Os modelos tendem a omitir passos intermédios essenciais à decisão diagnóstica. Uma resposta para este problema é o CoT, que decompõe o problema em etapas sequenciais e explicita o raciocínio, melhorando o desempenho do modelo em questões complexas [169].

Existe uma grande preocupação com a geração de alucinações por parte dos LLMs, que colocam em causa a confiabilidade do modelo. A recomendação dada para mitigar este problema, é

fundamentar as respostas com fontes verificadas. A literatura sugere pipelines de verificação factual e usar RAG, para que a resposta gerada seja baseada numa base de conhecimento confiável [148], [149], [150].

Outro desafio está relacionado com a privacidade dos dados dos pacientes. Para resolver este problema, os artigos [152], [160], ressaltam a necessidade da adesão a regulamentos de proteção de dados de saúde, prevenindo acesso não autorizado e uso indevido de informação sensível. Outra solução foi dada no artigo [164], onde o MedPromptExtract integra a ferramenta EIGEN para anonimização de resumos de alta, identificando classes como nome, ID do paciente e localização para mascaramento sistemático. O objetivo é produzir dados prontos a usar, preservando a anonimidade e reduzindo significativamente os custos de anonimização face a anotações manuais extensivas.

3.4 Conclusão

A revisão da literatura evidencia uma evolução significativa nos sistemas desenvolvidos para apoiar os profissionais de saúde na atualização do conhecimento médico. Esta necessidade tornou-se particularmente evidente face ao crescimento exponencial das publicações científicas, que gera sobrecarga informacional e dificulta a atualização contínua dos clínicos. A rapidez com que novos dados são produzidos torna impraticável a análise exclusivamente manual da evidência disponível, reforçando a importância de soluções tecnológicas capazes de automatizar e otimizar este processo.

Grande parte dos estudos analisados centra-se no contexto da COVID-19, período marcado por um aumento súbito e massivo de publicações científicas. Este cenário funcionou como catalisador para o desenvolvimento e teste de sistemas de recuperação, organização e síntese de literatura biomédica. Contudo, identificam-se também aplicações em áreas clínicas específicas, como a oftalmologia (EyeGPT) e a doença de Alzheimer (AlzheimerRAG), demonstrando que estas abordagens são transferíveis para diferentes domínios clínicos, desde que adaptadas às respetivas fontes de informação e contextos.

Nos últimos anos, observa-se uma transição clara para abordagens baseadas em NLP e LLM. Mais recentemente, destacam-se arquiteturas que integram RAG, capazes de recuperar passagens relevantes da literatura, resumir informação e responder em linguagem natural. Estes sistemas citam as fontes em que se baseiam as respostas, o que contribui para reduzir o risco de alucinações e aumentar a transparência do modelo. Além disso, permitem acelerar o acesso à informação clinicamente relevante sem exigir pesquisa manual extensiva ou exploração de grafos complexos.

Apesar dos avanços, persistem desafios importantes. Observam-se limitações no raciocínio em múltiplas etapas e na resposta a questões clinicamente complexas, especialmente em cenários que exigem integração de múltiplas variáveis contextuais. Mantêm-se também preocupações relacionadas com vieses nos dados de treino, que podem influenciar o desempenho do modelo

em determinadas populações ou patologias. Acrescem riscos associados à privacidade e segurança dos dados, incluindo a possibilidade de exposição indevida de informação sensível.

Para mitigar estes problemas, a literatura sugere várias estratégias complementares: utilização de arquiteturas RAG com citação explícita das fontes; aplicação de técnicas de *Chain-of-Thought* para explicitar o raciocínio do modelo; integração sistemática de ontologias biomédicas; e implementação rigorosa de processos de anonimização e proteção de dados.

Adicionalmente, torna-se evidente que a incorporação de normas clínicas constitui um fator crítico para aumentar a segurança, fiabilidade e aplicabilidade clínica dos sistemas baseados em LLMs. O HoPE exemplifica esta abordagem ao basear as suas respostas em diretrizes oficiais e protocolos de saúde. De forma semelhante, outros estudos demonstram que a integração de *guidelines* estruturadas e percursos clínicos permite aos modelos aumentar a probabilidade de dar uma resposta que seja adaptável ao contexto clínico.

4 Solução Proposta

A revisão literária revelou algumas falhas presentes nas atuais soluções tecnológicas utilizadas pelos profissionais de saúde para se manterem atualizados. Desta forma, este capítulo apresenta uma solução de um agente conversacional que utiliza RAG, como forma de criar uma tecnologia de fácil acesso e interação, mas que ao mesmo tempo, entregue informação confiável para os profissionais de saúde.

Desta forma, o capítulo começa com uma explicação da arquitetura do sistema, seguida de uma descrição do funcionamento da fase de interação com o utilizador.

4.1 Arquitetura do Sistema

Como referido anteriormente, a atualização científica dos profissionais de saúde é fundamental para garantir um atendimento ao paciente de elevada qualidade. A publicação diária de novos estudos científicos contribui para a evolução da evidência científica, tornando indispensável a atualização constante dos profissionais de saúde. Por sua vez, a atualização das normas de orientação clínica, fundamentadas nessa evidência, permite uniformizar a prática clínica, apoiar a tomada de decisão e contribuir para a prestação de cuidados mais seguros e eficazes.

O sistema desenvolvido recorre à tecnologia de IA, em particular à AI generativa, utilizando LLMs para responder em linguagem natural. Esta abordagem é utilizada através de um sistema de interação simples e intuitivo, capaz de responder em tempo real. No entanto, apesar das vantagens associadas à utilização de LLMs, estes modelos podem gerar respostas incorretas ou não fundamentadas, designadas por *alucinações*, bem como produzir informação desatualizada ou sem suporte em fontes científicas fiáveis. Para mitigar estas limitações, a arquitetura do sistema integra a abordagem RAG, que combina a capacidade de geração de texto dos LLMs com um mecanismo de recuperação de informação. Antes de criar uma resposta, o sistema

recupera os documentos mais relevantes de uma base de conhecimento, oferecendo um contexto adicional baseado em evidência científica.

A arquitetura geral da solução encontra-se representada na Figura 10.

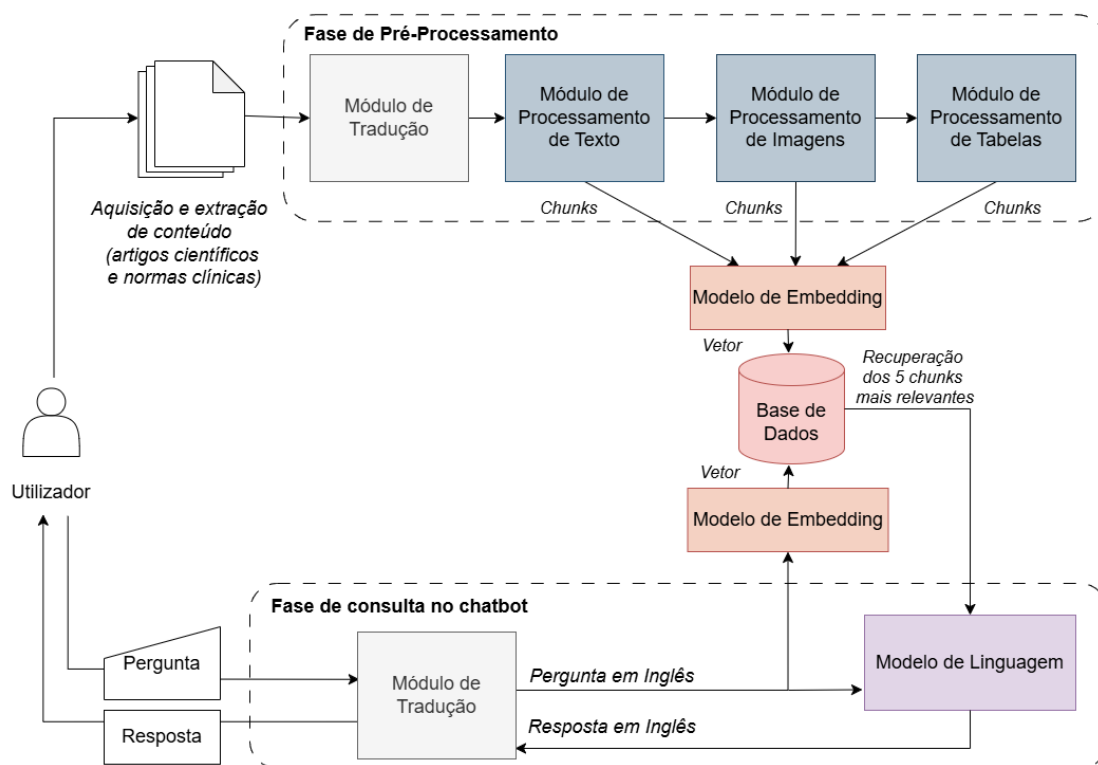


Figura 10: Arquitetura geral do sistema proposto

A fase de ingestão é iniciada através do carregamento dos documentos pelo utilizador e da extração automática do respetivo conteúdo. Nesta etapa, é realizada a análise da estrutura dos documentos, permitindo identificar os diferentes elementos que os constituem, nomeadamente texto, imagens e tabelas. Para esta extração é utilizada a ferramenta PaddleOCR, que reconhece alguns dos elementos presentes no documento e preserva a ordem de leitura.

Após a extração do conteúdo, os elementos textuais são encaminhados para um módulo de tradução, que começa por identificar automaticamente o idioma do documento e, caso necessário, traduz o conteúdo para inglês. Esta etapa permite uniformizar o idioma da informação antes do processamento específico do texto, das imagens e das tabelas.

De seguida, o processo de ingestão inclui módulos responsáveis pelo processamento e estruturação de cada tipo de conteúdo. O módulo de texto identifica e organiza o conteúdo textual, preservando a hierarquia das secções e subsecções do documento. O módulo de imagens identifica as figuras existentes, associa-lhes as respetivas legendas e prepara o conteúdo para posterior indexação e recuperação. De forma análoga, o módulo de tabelas identifica as tabelas presentes nos documentos, preservando a sua estrutura e informação para

posterior utilização durante a geração de respostas. Em conjunto, estes módulos servem para estruturar automaticamente o conteúdo textual e visual dos documentos, convertendo informação originalmente não estruturada em representações organizadas que podem ser integradas na base de dados do sistema.

Após o processamento de cada tipo de conteúdo, a informação é segmentada em *chunks* e convertida em *embeddings*, que representam semanticamente os diferentes elementos do documento. Estes *embeddings* são posteriormente armazenados numa base de dados vetorial, organizada em três coleções distintas, correspondentes ao texto, às imagens e às tabelas.

Na fase de consulta do *chatbot*, o utilizador submete uma pergunta em linguagem natural através da interface do *chatbot*. À semelhança do processo de ingestão, a pergunta é encaminhada para o módulo de tradução, de forma a uniformizar o idioma da consulta com o dos conteúdos armazenados na base de dados vetorial. Esta questão é então processada e convertida num vetor de *embedding*, permitindo representar semanticamente a consulta.

De seguida, o sistema realiza uma pesquisa na base de dados vetorial, recorrendo à similaridade entre *embeddings*, de forma a identificar os *chunks* mais relevantes armazenados nas diferentes coleções (texto, imagens e tabelas). Os resultados recuperados são agregados e organizados como contexto para o modelo de linguagem.

Posteriormente, este contexto é enviado para um LLM, que gera a resposta final em linguagem natural, tendo em consideração a informação recuperada. Adicionalmente, o sistema disponibiliza ao utilizador o contexto recuperado que serviu de base à resposta gerada, permitindo consultar os *chunks* e os documentos de origem utilizados pelo modelo.

4.2 Módulo de processamento do texto

O módulo de processamento do texto tem como objetivo transformar o conteúdo textual dos documentos numa representação estruturada, preservando as relações hierárquicas existentes entre títulos, secções, subsecções e parágrafos. Esta etapa permite manter a organização lógica do documento original, assegurando que a informação extraída permanece associada ao respetivo contexto estrutural.

Durante o processo de segmentação, os parágrafos são convertidos em unidades independentes de processamento, podendo perder a associação direta com as secções e subsecções onde se encontravam inseridos no documento original. Esta separação pode resultar na perda de informação contextual, uma vez que a interpretação de determinados segmentos pode depender da relação estabelecida com os títulos e níveis hierárquicos superiores a que pertencem.

Conforme ilustrado na Figura 11, o módulo encontra-se organizado em três etapas sequenciais responsáveis pela reconstrução da estrutura lógica do documento. A primeira etapa consiste na identificação do título principal do documento, correspondente ao nível hierárquico superior

da estrutura textual. Este elemento é posteriormente utilizado como informação de contexto pelos módulos responsáveis pelo processamento de tabelas e imagens, permitindo associar cada elemento ao documento de origem. Desta forma, todas as tabelas e imagens extraídas mantêm a referência ao título principal do documento a que pertencem.

A segunda etapa é responsável pela identificação dos títulos e subtítulos presentes no documento, bem como pela determinação das respetivas relações hierárquicas. Este processo permite reconstruir a organização das diferentes secções e subsecções, estabelecendo a ligação entre cada elemento textual e o respetivo contexto estrutural.

Por fim, a terceira etapa consiste na associação da informação hierárquica previamente identificada aos parágrafos extraídos. O resultado deste processo é uma representação textual estruturada, na qual cada segmento mantém a ligação com a organização original do documento, facilitando o processamento e a utilização da informação nas etapas subsequentes do sistema.

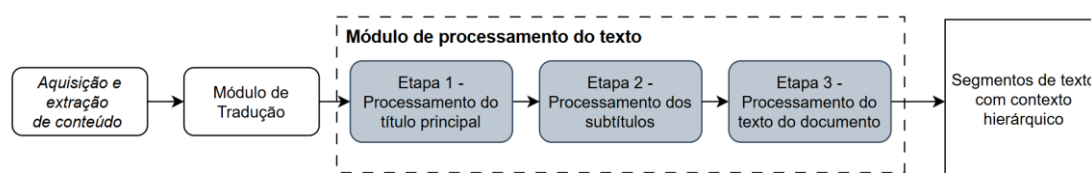


Figura 11: Etapas do módulo de processamento do texto

A Figura 11 apresenta uma visão geral do módulo de processamento do texto, ilustrando as diferentes etapas necessárias para transformar o texto extraído numa representação estruturada.

4.3 Módulo de processamento das imagens

O módulo de processamento das imagens tem como objetivo identificar as figuras presentes nos documentos e gerar uma representação estruturada do seu conteúdo, integrando a descrição da imagem, a respetiva legenda e o título principal do documento.

A inclusão de imagens no agente conversacional funciona como um complemento à informação textual, uma vez que muitos artigos científicos recorrem a figuras, diagramas e fluxogramas para representar visualmente conceitos, procedimentos clínicos ou processos de decisão. Desta forma, disponibilizar estes elementos durante a interação com o *chatbot* facilita a interpretação das respostas geradas pelo modelo, oferecendo ao utilizador informação visual que complementa a explicação textual.

Conforme ilustrado na Figura 12, o módulo inicia o processamento das imagens através da identificação das figuras presentes no documento, da respetiva associação às legendas e da ligação ao título principal do documento. Durante o processo de extração dos componentes documentais, a informação é preservada de acordo com a ordem de leitura original do

documento. Esta característica permite estabelecer uma associação entre cada figura e a sua legenda, uma vez que, na maioria dos casos, a legenda surge imediatamente após a respetiva figura na sequência de leitura extraída. Desta forma, as legendas são identificadas automaticamente.

Contudo, existem casos em que este método não funciona. Para estes casos, é aplicado um mecanismo complementar baseado na identificação de padrões linguísticos frequentemente utilizados na nomenclatura de figuras científicas, como “Figura”, “Fig.” ou equivalentes em inglês (“Figure”).

O passo seguinte consiste na geração de uma descrição textual das figuras extraídas, com o objetivo de converter a informação visual numa representação textual integrada no sistema. Para esta etapa, são aplicadas estratégias distintas de acordo com o tipo de conteúdo visual identificado. No caso de figuras que representam fluxogramas, é utilizado um modelo multimodal capaz de analisar diretamente a componente visual e gerar uma descrição textual do seu conteúdo. Por outro lado, quando a figura corresponde a uma imagem ilustrativa, a descrição é gerada usando os parágrafos que fazem referência a essa mesma imagem.

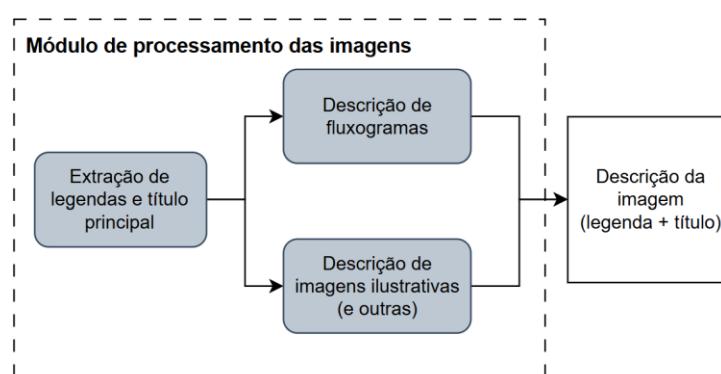


Figura 12: Módulo de processamento das imagens

Esta abordagem permite adaptar o processo de descrição às características de cada tipo de imagem, combinando a interpretação visual direta com a informação contextual disponibilizada pelo documento. Os detalhes da implementação destas estratégias são apresentados no capítulo seguinte.

4.4 Módulo de Processamento de Tabelas

O módulo de processamento de tabelas tem como objetivo identificar as tabelas presentes nos documentos e produzir uma representação estruturada do seu conteúdo, preservando a relação entre os dados tabulares, a respetiva legenda e o título principal do documento.

As tabelas representam uma componente relevante dos artigos científicos, sendo frequentemente utilizadas para apresentar resultados experimentais, recomendações clínicas, comparações entre métodos e valores de referência que podem não estar explicitamente

descritos no conteúdo textual do documento. Contudo, durante o processo de extração e indexação, a estrutura original das tabelas pode ser perdida, dificultando a interpretação correta da informação pelos modelos de linguagem. Para ultrapassar esta limitação, este módulo converte o conteúdo tabular numa representação textual estruturada, preservando as relações existentes entre cabeçalhos, linhas e colunas, bem como o contexto associado à tabela.

Conforme ilustrado na Figura 13, o processamento das tabelas inicia-se com a identificação da respetiva legenda e com a associação ao título principal do documento. Este processo segue uma abordagem semelhante à utilizada no módulo de processamento de imagens. No entanto, no caso das tabelas, é tido em consideração que a legenda aparece antes da tabela.

Sempre que presentes, as notas de rodapé associadas à tabela são igualmente identificadas e integradas na representação final, uma vez que podem conter informação complementar que apoie na interpretação dos dados apresentados.

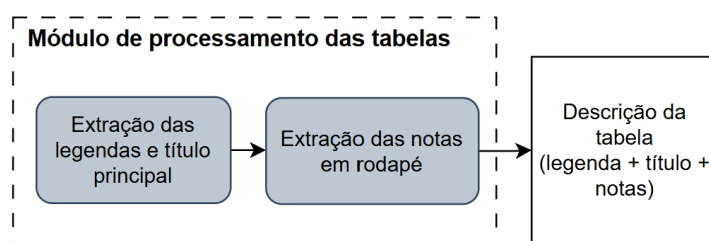


Figura 13: Módulo de Extração de Tabelas

O resultado permite que o conteúdo das tabelas seja posteriormente segmentado, indexado e recuperado pelo sistema, mantendo o contexto necessário para a geração de respostas pelo agente conversacional. Os detalhes da implementação desta abordagem são apresentados no capítulo seguinte.

5 Implementação

Este capítulo apresenta a implementação da solução desenvolvida, descrevendo os principais componentes que constituem o sistema. Inicialmente são apresentadas as tecnologias e os modelos de AI utilizados. De seguida, é detalhada a implementação dos módulos responsáveis pelo processamento de texto, tabelas e imagens, bem como o mecanismo de recuperação da informação. Por fim, é apresentada a interface da aplicação, ilustrando as principais funcionalidades disponibilizadas ao utilizador.

5.1 Ferramentas e tecnologias utilizadas

Nesta secção da tese são descritos os principais componentes tecnológicos utilizados no projeto, nomeadamente bibliotecas e *frameworks* que sustentam o agente conversacional desenvolvido nesta tese.

O desenvolvimento do sistema foi realizado no ambiente *Microsoft Visual Studio*. Recorreu-se à linguagem de programação *Python*, por ter várias bibliotecas orientadas à AI, NLP e ciência de dados.

O LangChain é um *framework open-source* que simplifica o desenvolvimento de aplicações baseadas em LLMs. As suas funcionalidades passam pela gestão de contexto, estruturação de *prompts* e coordenação do fluxo de dados entre diferentes componentes do sistema, o que reduz significativamente a complexidade da construção de aplicações de NLP [170], [171].

Para a indexação e recuperação semântica da informação, utilizou-se a base de dados vetorial Chroma, acedida através da classe `langchain_community.chat_models.ChatOllama` disponibilizada pelo LangChain. Esta classe permite armazenar, indexar e consultar representações vetoriais dos documentos, suportando a recuperação de informação com base na similaridade semântica [172].

Os *embeddings* geraram-se através da classe `HuggingFaceEmbeddings`, sendo posteriormente armazenados no Chroma para utilização durante o processo de recuperação. A Hugging Face é uma plataforma e comunidade colaborativa dedicada ao desenvolvimento e partilha de ferramentas, modelos e recursos para inteligência artificial e aprendizagem automática [173].

A interface web do sistema foi desenvolvida com recurso ao Streamlit, uma biblioteca que permite a criação de aplicações web interativas diretamente em Python. Esta ferramenta facilita a construção de interfaces funcionais e intuitivas sem necessidade de desenvolvimento extensivo em tecnologias de *front-end* [174].

5.2 Modelos de Embeddings

Para o sistema desenvolvido ao longo desta tese, o modelo de *embeddings* foi responsável pela representação vetorial dos documentos e das questões dos utilizadores, possibilitando a comparação semântica entre ambos e a recuperação dos documentos mais relevantes para a resposta. É por esse motivo, uma parte importante da etapa de recuperação de informação.

Embora exista uma grande diversidade de modelos de *embeddings*, foram seleccionados apenas aqueles considerados mais relevantes para este trabalho. Esta seleção procurou abranger modelos treinados em diferentes tipos de dados, incluindo modelos de domínio geral e modelos especializados em conteúdos biomédicos. Foram também utilizados modelos de diferentes dimensões, de forma a avaliar o compromisso entre a qualidade da representação semântica e a eficiência computacional.

Nessa perspetiva, dois dos modelos são generalistas e pertencem à biblioteca *Sentence Transformers*: o `all-MiniLM-L6-v2` e o `all-mpnet-base-v2`. O modelo `all-mpnet-base-v2` é reconhecido pelo elevado desempenho em tarefas de similaridade semântica, produzindo representações vetoriais de elevada qualidade. Por outro lado, o `all-MiniLM-L6-v2` apresenta um custo computacional significativamente inferior, sendo aproximadamente cinco vezes mais rápido, mantendo simultaneamente um desempenho competitivo em tarefas de recuperação de informação e pesquisa semântica [175], [176]. A inclusão destes dois modelos permite, assim, analisar o compromisso entre qualidade das representações e eficiência computacional.

Foram igualmente seleccionados dois modelos especializados no domínio biomédico. O PubMedBERT, um modelo baseado na arquitetura BERT, foi pré-treinado exclusivamente com literatura científica proveniente da base de dados PubMed, permitindo uma melhor adaptação à terminologia biomédica e clínica [177]. O MedEmbed é um modelo de *embeddings* desenvolvido especificamente para tarefas de recuperação de informação médica, recorrendo a uma estratégia de aprendizagem contrastiva baseada em *triplets* (consulta, documento relevante e documento negativo), treinados a partir de documentos do PubMed Central [178].

A Tabela 5 apresenta um resumo das principais características dos modelos considerados.

Tabela 5: Modelos de *Embeddings*

Modelo	Parâmetros	Tamanho Máximo do Input	Especialidade
all-MiniLM-L6-v2 ^a	22.7M	512	Conteúdo Geral
all-mpnet-base-v2 ^b	0.1B	512	Conteúdo Geral
NeuML/pubmedbert ^c	0.1B	512	Conteúdo Biomédico
abhinand/MedEmbed-small-v0.1 ^d	33.4M	512	Conteúdo Biomédico

^a <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>;

^b <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>;

^c <https://huggingface.co/NeuML/pubmedbert-base-embeddings>;

^d <https://huggingface.co/abhinand/MedEmbed-small-v0.1>;

5.3 Modelos de Linguagem de Grande Escala

No desenvolvimento do sistema, os LLM são utilizados em várias etapas da arquitetura proposta, sendo utilizados em diferentes tarefas, como a classificação da estrutura dos documentos, a geração de descrições de figuras e a resposta às questões dos utilizadores.

5.3.1 Infraestrutura Computacional

Neste trabalho foram utilizadas diferentes infraestruturas para a execução dos modelos de linguagem. O modelo LLaMA 3.1 8B foi executado localmente, num computador equipado com 32 GB de memória RAM e uma unidade de processamento gráfico (GPU) com 8 GB de memória dedicada.

Os modelos LLaMA 3.3 70B, Qwen3-32B e GPT-OSS-120B foram executados através da plataforma Groq [179], enquanto o modelo Kimi K2 foi executado recorrendo ao serviço Cloudflare Workers AI [180]. Estas plataformas disponibilizam infraestrutura otimizada para inferência de modelos de linguagem, permitindo eliminar as limitações de hardware associadas à execução local.

5.3.2 Modelos utilizados

Com o objetivo de avaliar o impacto da arquitetura e da dimensão dos modelos no desempenho do sistema, foram selecionados cinco modelos de linguagem pertencentes a diferentes famílias e com características distintas.

O Kimi K2.6 (*moonshotai/kimi-k2-instruct*) é um modelo do tipo *Mixture-of-Experts*, com aproximadamente um bilião de parâmetros totais e cerca de 32 mil milhões de parâmetros ativados durante a inferência. Este modelo foi desenvolvido para tarefas avançadas de NLP,

destacando-se em geração de texto, compreensão contextual e sistemas de pergunta-resposta [181].

O Qwen3-32B, desenvolvido pela Alibaba Cloud, pertence à mais recente geração da família Qwen e oferece um elevado desempenho em tarefas de compreensão e geração de linguagem natural, mantendo simultaneamente uma elevada eficiência computacional [182].

O GPT-OSS-120B, desenvolvido pela OpenAI, representa um dos modelos de maior dimensão considerados nesta investigação. Para além do seu desempenho geral em tarefas de raciocínio e geração de texto, estudos recentes demonstram o seu potencial em aplicações clínicas, nomeadamente na análise de registos médicos e identificação de riscos associados à medicação. Apesar dos resultados promissores, este estudo revela algumas limitações na interpretação do contexto clínico e na formulação de recomendações terapêuticas [183].

Também se avaliaram dois modelos da família LLaMA (*Large Language Model Meta AI*), desenvolvidos pela Meta: LLaMA 3.3 70B e LLaMA 3.1 8B. Estes modelos representam duas escalas distintas da mesma arquitetura, o que permite analisar a influência da dimensão do modelo no desempenho do sistema. Enquanto o modelo LLaMA 3.3 70B privilegia capacidade de raciocínio, compreensão contextual e qualidade da geração de texto, o LLaMA 3.1 8B apresenta requisitos computacionais substancialmente inferiores, sendo uma alternativa mais eficiente para casos em que os recursos computacionais são limitados [184].

Tabela 6: Configurações dos LLMs

Modelo	Organização	Parâmetros	Janela de contexto
Kimi K2.6	Moonshot AI	1 T (32 B ativos)	128 k
Qwen3-32B	Alibaba Cloud	32 B	32 k
GPT-OSS-120B	Open AI	120 B	131 k
LLaMA 3.3 70B	Meta	70 B	128 k
LLaMA 3.1 8B	Meta	8 B	128 k

5.3.3 Modelos utilizados no Pré-Processamento

A escolha do modelo de linguagem influencia diretamente a qualidade das respostas geradas, bem como os requisitos computacionais associados à sua execução. Neste trabalho, foram considerados dois modelos da família LLaMA para a fase de pré-processamento, nomeadamente o LLaMA 3.1 8B e LLaMA 3.3 70B, de forma a equilibrar o desempenho e eficiência.

A seleção do modelo a utilizar em cada tarefa foi realizada através de uma estratégia de escalonamento baseada na complexidade da tarefa e no desempenho observado durante a avaliação dos modelos. Inicialmente, foi privilegiada a utilização do modelo LLaMA 3.1 8B, devido ao menor custo computacional associado, menor tempo de inferência e reduzidos requisitos de memória. No entanto, para tarefas em que este modelo apresentou limitações ao

nível da compreensão do conteúdo ou da qualidade das respostas produzidas, foi selecionado o modelo LLaMA 3.3 70B, que apresenta maior capacidade de representação e melhores resultados em tarefas que exigem maior capacidade de raciocínio.

Desta forma, o modelo LLaMA 3.3 70B é utilizado apenas nas tarefas em que a capacidade adicional de compreensão e raciocínio é necessária, enquanto o LLaMA 3.1 8B é aplicado nas restantes etapas, onde apresenta desempenho adequado. Esta abordagem permite otimizar os recursos computacionais do sistema, mantendo simultaneamente a qualidade necessária nas tarefas de processamento mais exigentes.

5.4 Pipeline de extração de texto

Após a seleção das ferramentas de pré-processamento, foi desenvolvido um pipeline para extrair e estruturar automaticamente o conteúdo textual dos artigos científicos. O objetivo deste processo consiste em transformar documentos PDF numa representação estruturada, preservando a hierarquia do documento e a associação entre títulos, subtítulos e respetivos parágrafos.

A preservação da estrutura original do documento é um requisito fundamental para garantir que a informação extraída mantém o seu contexto semântico. Ao contrário de uma extração linear do texto do documento, esta abordagem é utilizada para identificar a organização lógica do documento, e estabelecer relações entre secções, subsecções e respetivo conteúdo.

A Figura 14 apresenta o fluxograma do pipeline implementado. O processo encontra-se dividido em três etapas principais:

- Etapa 1 - identificação e validação do título principal do documento;
- Etapa 2 - deteção e classificação hierárquica dos subtítulos;
- Etapa 3 - extração e associação dos parágrafos às respetivas secções;

Cada uma destas etapas desempenha um papel específico na tarefa de processamento do texto. A primeira é responsável por extrair o título, ou caso não seja identificado nenhum título, para criar um que represente o conteúdo do documento. A segunda extrai e identifica a organização hierárquica das diferentes secções e subsecções do documento. E por último, a terceira etapa, é responsável por associar os títulos e subtítulos aos parágrafos do documento. Para além disso, também é responsável por eliminar algumas secções que são irrelevantes para o sistema, nomeadamente, informação relacionada com os autores. O resultado é uma representação textual organizada e semanticamente coerente, que facilita a recuperação da informação na etapa seguinte.

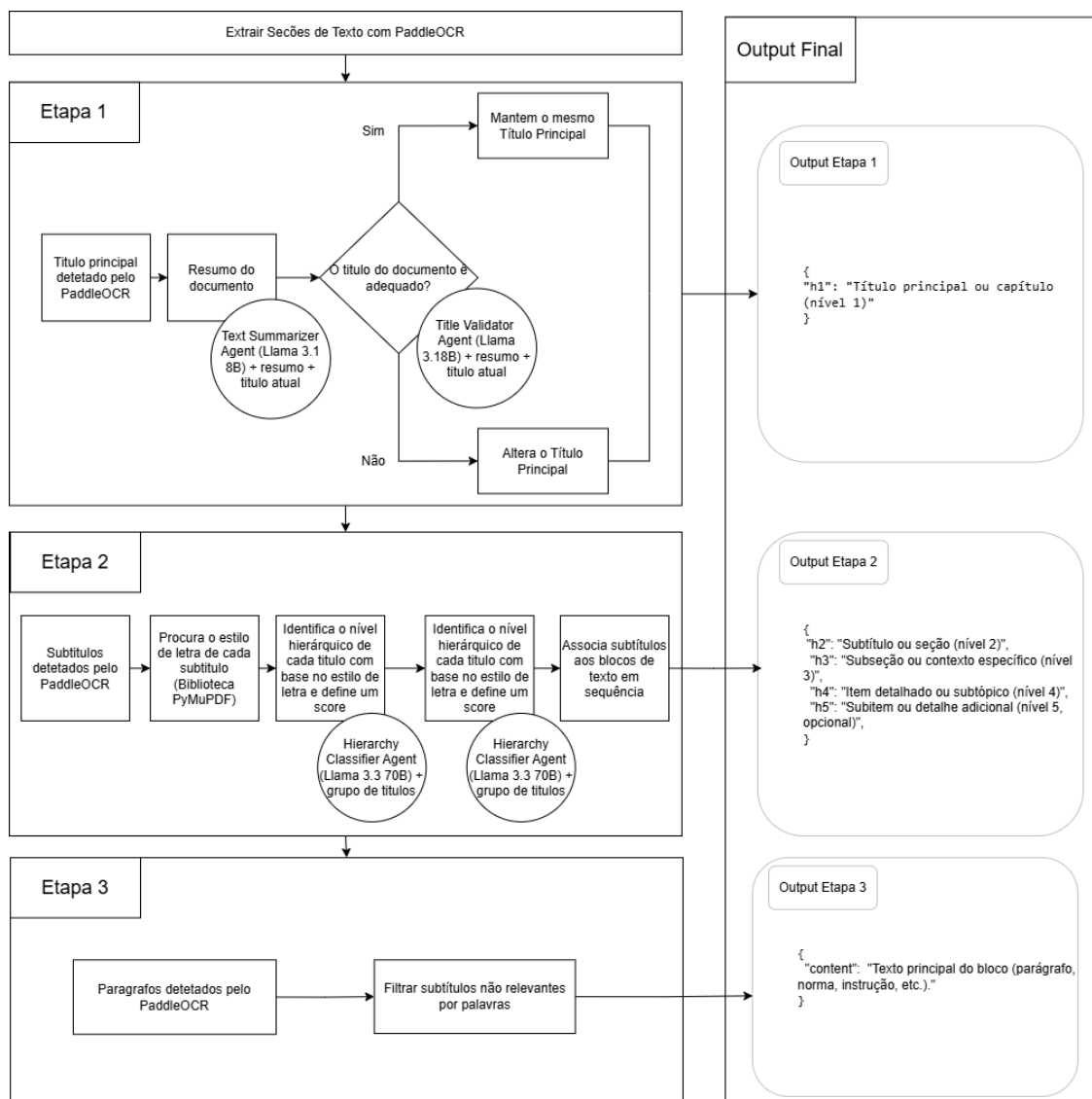


Figura 14: Diagrama de processamento das secções de texto

5.4.1 Etapa 1: Extrair título principal do Documento

A primeira etapa do pipeline consiste na identificação e validação principal do título principal do documento, o qual é posteriormente utilizado como representação global do conteúdo e armazenado como o nível hierárquico h1.

- Extração inicial do título através de OCR

Numa fase inicial, recorreu-se à biblioteca PaddleOCR para a deteção do título principal em documentos em formato PDF. Esta abordagem demonstrou bons resultados em documentos com estrutura consistente e formatação clara, permitindo identificar corretamente o título em múltiplos casos.

No entanto, durante a avaliação do sistema num conjunto mais alargado de documentos, foram observadas limitações relevantes. Em particular, verificaram-se situações em que:

- O título extraído não correspondia ao tema principal do documento;
- Existiam múltiplos candidatos plausíveis a título principal;
- O resultado do OCR apresentava ruído estrutural, especialmente em documento com formatação heterogénea

Estas limitações motivaram a introdução de uma etapa adicional de validação baseada em modelos de linguagem.

- Validação semântica do título com recurso a LLM

Para mitigar as limitações do OCR, introduziu-se uma etapa de validação semântica baseada em LLM. O objetivo desta etapa consiste em verificar se o título identificado pelo PaddleOCR é semanticamente coerente com o conteúdo global do documento e, caso não seja, propor uma alternativa mais representativa.

Como forma contextualizar o modelo acerca do conteúdo do documento, é gerado um resumo do documento, que servirá de contexto para o modelo avaliar o título.

- Geração do resumo do documento

No pré-processamento, o resumo tem duas finalidades principais: avaliar se o título identificado pelo PaddleOCR é adequado ao conteúdo do documento e, caso não seja, fornecer informação suficiente para que o modelo possa sugerir um novo título mais representativo.

Devido às limitações da janela de contexto dos modelos de linguagem, é adotada uma estratégia de segmentação do documento. Assim, o texto é dividido em blocos com um máximo de 350 palavras, permitindo que cada segmento seja processado individualmente pelo modelo LLaMA 3.1 8B, executado localmente.

Para cada bloco é gerado um resumo parcial, composto por um título curto e uma descrição objetiva do conteúdo. O *prompt* utilizado nesta etapa encontra-se no Código 1.

```
Read the following text and return a summary in JSON format:
{
  "title": "<short and informative title>",
  "abstract": "<concise and objective summary in English>"
}
#Rules:
- Do NOT include any information that is not explicitly present
in the provided text.
- Do NOT invent, infer, or guess missing details.
#Text: {parte}
```

Código 1: Estrutura do *prompt* para a criação dos resumos parciais do documento

Após o processamento de todos os blocos, os resumos parciais são agregados e novamente processados por um LLM, resultando num resumo global do documento. Este também é gerado utilizando o modelo LLaMA 3.1 8B. O *prompt* utilizado encontra-se no Código 2.

Combine the following summaries into a single coherent summary and return it in the same JSON format:

```
{{
  "title": "<general title>",
  "abstract": "<integrated summary in English>"
}}
```

Text: {resumo}

Código 2: Estrutura do *prompt* para criar o resumo global do documento

Esta estratégia permitiu reduzir a complexidade do documento mantendo uma representação semântica global eficiente para as tarefas de validação. Também permitiu utilizar um modelo com mais baixo custo computacional, sem alterar a qualidade dos resultados obtidos.

- Validação do título com base em LLM

Após a geração do resumo global, procede-se à validação do título identificado pelo PaddleOCR. Para este efeito, é utilizado o modelo LLaMA 3.1 8B, que recebe como entrada o título extraído pelo PaddleOCR e o resumo global do documento.

O modelo é instruído a avaliar a adequação do título relativamente ao conteúdo do documento, considerando possíveis problemas como:

- Título demasiado genérico
- Título incompleto ou corrompido
- Discrepância semântica entre o título e o conteúdo do documento

Sempre que o título é considerado inadequado, o modelo gera uma proposta alternativa mais representativa do conteúdo global.

A saída do modelo é estruturada em formato JSON, contendo os seguintes campos:

- evaluation: classificação do título (adequado ou inadequado);
- justification: justificação da decisão;
- suggestion: novo título proposto (quando aplicável).

Aplicou-se a técnica de *few-shot prompting*, através da inclusão de exemplos de entrada e saída, com o objetivo de estabilizar o formato da resposta e melhorar a consistência do modelo.

O *prompt* utilizado nesta etapa encontra-se no Código 3.

```
You will receive a document title and a summary of an obstetrics-related document.
```

```
Your task is to strictly evaluate the quality and specificity of the title.
```

```
Apply the following decision rules strictly and in order:
```

```
1) If the title appears corrupted, fragmented, contains random characters, OCR artifacts, or does not form a meaningful grammatical phrase → return: "Title corrupted or invalid."
```

2) If the title does NOT explicitly mention the specific medical condition, intervention, or clinical topic described in the content
→ return: "Title too generic."

3) Institutional or administrative titles (e.g., Ordinance, Directorate, Norm, Recommendation, Guideline) WITHOUT a specific clinical topic → ALWAYS return: "Title too generic."

4) Only classify as "Adequate" if the specific clinical topic is explicitly written in the title.

Strict constraints:

- Do NOT infer adequacy from the content alone.
- The specific clinical topic must be explicitly written in the title.
- Be strict.
- Do not add explanations outside the JSON.
- Justification must be concise (1-2 sentences).
- The suggestion must be realistic, specific, and medically accurate.
- The suggestion must explicitly include the main clinical topic described in the content.

Output format rules:

- Return JSON only.
- Always include all three fields: "evaluation", "justification", and "suggestion".
- If evaluation = "Adequate", suggestion must be null.
- If evaluation ≠ "Adequate", suggestion is mandatory.
- The suggestion field must NEVER be empty.
- Do not include any extra text outside the JSON.

Example:

Title: ORDINANCE OF THE GENERAL DIRECTORATE FOR HEALTH

Content: This document describes screening and management of gestational diabetes.

Output:

```
{{
  "evaluation": "Inadequate.",
  "justification": "The title is institutional and does not mention
gestational diabetes.",
  "suggestion": "Gestational Diabetes Screening and Management
Guideline"
}}
```

Now evaluate the following:
Title: {titulo}
Document content: {conteudo}

Código 3: Estrutura do *prompt* para a validação do título principal do documento

De forma a ilustrar o processo de validação do título, a Tabela 7 apresenta exemplos práticos, em que o modelo é utilizado para avaliar se o título detetado pelo PaddleOCR era adequado, tendo em consideração o conteúdo do documento.

Tabela 7: Exemplos de aplicação do *prompt* de validação de títulos.

Nome da Norma Clínica	Título Principal detetado pelo PaddleOCR	Resposta do Modelo LLaMA
Guideline: External Cephalic Version	SPOMMF GUIDELINES	{'evaluation': 'The title is too general', 'justification': 'The title 'SPOMMF GUIDELINES' does not accurately describe the document's content, which seems to focus on external cephalic version and its implications.', 'suggestion': ' External Cephalic Version: GUIDELINES '}
Guideline for Episiotomy	SPOMMF Clinical Guidelines for Episiotomy	{'evaluation': 'Adequate', 'justification': 'The title clearly describes the content of the document, which is a clinical guideline from the Portuguese Society of Maternal-Fetal Medicine on episiotomy.', 'suggestion': None }
Guideline for Screening for Short Cervix	SPOMMF Clinical Guideline	{'evaluation': 'The title is too generic', 'justification': 'The title 'SPOMMF CLINICAL GUIDELINES' does not accurately describe the content, which seems to focus on a specific procedure for screening for a short cervix.', 'suggestion': ' Short Cervix Screening: SPOMMF Clinical Guidelines '}
Guidelines for Thyroid Disease in Pregnancy	SPOMMF Clinical Guidelines for Thyroid Disease in Pregnancy	{'evaluation': 'Adequate', 'justification': 'The title clearly describes the topic and objective of the document, which is to provide clinical guidelines for thyroid disease in pregnancy.', 'suggestion': None }

Os resultados indicam que o modelo é capaz de distinguir entre títulos adequados e inadequados com base na sua correspondência semântica com o conteúdo do documento. Ou seja, o modelo consegue ter bons resultados a identificação de títulos genéricos ou pouco informativos.

Para além disso, nos casos em que o título é considerado inadequado, o modelo propõe uma alternativa que tende a ser semanticamente mais específica e alinhada com o conteúdo do documento.

Dado o desempenho observado com o modelo LLaMA 3.1 8B nesta tarefa, não se considerou necessário proceder à avaliação de modelos de maior dimensão, uma vez que os resultados obtidos foram satisfatórios no contexto experimental definido.

5.4.2 Etapa 2: Extrair subtítulos do Documento

Esta etapa tem como objetivo identificar os subtítulos presentes nos documentos e atribuir-lhes uma estrutura hierárquica coerente (entre os níveis h1 a h5), de acordo com a estrutura original do documento. Para isso, este processo baseia-se na combinação de técnicas de OCR, extração tipográfica e LLMs.

Inicialmente, os subtítulos são detetados através da biblioteca PaddleOCR. Esta ferramenta permite identificar regiões de texto potencialmente correspondentes a títulos e subtítulos. No entanto, durante a fase experimental, verificou-se que o PaddleOCR pode gerar falsos positivos, nomeadamente ao classificar legendas de tabelas e figuras como subtítulos do documento.

Para mitigar este problema aplicou-se um conjunto de regras heurísticas para excluir subtítulos com padrões incorretos, como por exemplo, iniciados pelas palavras “Tabela” ou “Figura”, e que frequentemente estão associados a legendas e não a subtítulos do documento.

- Abordagem inicial baseada apenas no conteúdo textual

Numa fase inicial do desenvolvimento desta etapa, explorou-se uma abordagem baseada exclusivamente na informação textual dos subtítulos identificados pelo PaddleOCR. O objetivo consistia em avaliar se um LLM seria capaz de inferir corretamente a hierarquia estrutural do documento recorrendo apenas ao conteúdo dos subtítulos e ao seu conhecimento prévio sobre a organização típica de documentos científicos.

Para este efeito, desenvolveu-se um *prompt* que recebia como entrada a lista de subtítulos extraídos do documento e solicitava ao modelo a atribuição de um nível hierárquico a cada um deles. O *prompt* incluía um conjunto de regras heurísticas destinadas a orientar a classificação, privilegiando, sempre que possível, a utilização da numeração presente nos títulos (por exemplo, 1, 1.1). Na ausência de numeração, eram consideradas outras convenções estruturais frequentemente utilizadas em artigos biomédicos, como a atribuição do nível principal a secções intituladas “Introduction” ou “Conclusion”. O Código 4 apresenta o *prompt* utilizado para esta etapa.

```
You will receive a list of titles from a document about obstetrics:
{list_titles}
I want a JSON structure with the hierarchical levels of the titles.

## Rules for level detection (heuristics, apply in this order):
1) **Structured numbering**: if the chapters are numbered, follow
the numbering
2) **Alphabetical lists** (only if there is no numbering):
- "a)", "b)", "c)" → treats them as subtitles of a higher-level
title.
Returns a JSON with "sections": [
{{ "title": "Title1", "level":1 }},
{{ "title": "Title2", "level":2 }},]
```

Código 4: Estrutura do *prompt* aplicado ao modelo de linguagem, para classificar os subtítulos por hierarquias

Embora esta abordagem tenha produzido resultados satisfatórios em documentos com uma estrutura explícita e baseada em numeração, verificaram-se limitações significativas em documentos cuja organização dependia predominantemente do conhecimento do LLM. Nestes casos, o modelo não dispunha de informação suficiente para distinguir corretamente os diferentes níveis hierárquicos.

A Tabela 8: apresenta um exemplo representativo destas limitações.

Tabela 8: Exemplo de classificação hierárquica obtida pela abordagem baseada exclusivamente na informação textual

Subtítulo Identificado	Nível esperado	Nível atribuído pelo LLM
<i>Hyperthyroidism:</i>	1	1
<i>Complications:</i>	2	3
<i>Clinical Management:</i>	2	4
<i>Management of the fetus/newborn</i>	2	4

Neste exemplo, observou-se que o modelo classificou corretamente o título principal *Hyperthyroidism*, mas atribuiu níveis hierárquicos excessivamente profundos às restantes secções. Em particular, os subtítulos *Clinical Management* e *Management of the fetus/newborn* foram interpretados como dependentes de *Complications*, quando, na realidade, constituem secções paralelas diretamente subordinadas ao título principal.

Estes resultados demonstram que a informação textual, por si só, nem sempre fornece contexto suficiente para inferir corretamente a estrutura hierárquica de um documento. Em muitos documentos científicos, a distinção entre níveis hierárquicos depende fortemente de características visuais, como o tamanho da fonte, o tipo de letra ou outros elementos de formatação, que não estavam disponíveis para o modelo nesta abordagem inicial.

Face a estas limitações, optou-se por complementar a informação textual com características tipográficas extraídas diretamente do documento PDF, enviando ao modelo evidências estruturais adicionais para suportar a classificação hierárquica. A metodologia adotada é descrita na secção seguinte.

- Integração da informação tipográfica

Devido aos resultados obtidos anteriormente, nesta etapa, procedeu-se à extração de informação estrutural adicional através da PDF PyMuPDF (fitz), uma biblioteca que permite obter características tipográficas diretamente do ficheiro PDF, nomeadamente:

- Tipo de fonte
- Tamanho da fonte
- Conteúdo textual

A biblioteca PyMuPDF extrai esta informação ao nível de cada palavra presente no documento, disponibilizando, para cada elemento textual, o respetivo tipo e tamanho de fonte. De forma a ser possível utilizar esta informação na tarefa de classificação hierárquica, implementou-se um processo de agrupamento dos elementos textuais adjacentes que apresentavam o mesmo tipo e tamanho de fonte, originando segmentos de texto com características tipográficas iguais.

Um exemplo simplificado desta estrutura encontra-se apresentado no Código 5.

```
[
  {
    "font": "SourceSansPro-Regular",
    "size": 20,
    "text": "4.2 Implementation considerations"
  },
  {
    "font": "SourceSansPro-Light",
    "size": 10,
    "text": "Dissemination of an executive summary containing
the recommendations..."
  }
]
```

Código 5: Exemplo de segmentos textuais com informação tipográfica extraída com PyMuPDF

Como a informação extraída pelo PyMuPDF correspondia a todo o conteúdo textual do documento, incluindo texto, títulos, legendas e outros elementos, foi necessário identificar quais destes segmentos correspondiam aos subtítulos previamente detetados pelo PaddleOCR. Para isso, comparou-se os subtítulos detetados pelo PaddleOCR com os segmentos extraídos pelo PyMuPDF, aplicando uma métrica de similaridade textual entre os dois conjuntos. Sempre que a similaridade entre dois textos era superior a um limiar definido de 0,9, considerou-se que existe correspondência entre ambos. Quando esta correspondência era estabelecida, o subtítulo era associado os respetivos atributos tipográficos (tipo e tamanho de fonte).

Após esta tarefa, classificaram-se os subtítulos quanto ao nível hierárquico utilizando um LLM. O LLM recebeu como entrada a lista de subtítulos enriquecidos com informação tipográfica e produziu como saída, a atribuição de um nível hierárquico entre h1 e h5.

Para orientar o modelo na classificação desta tarefa, foi desenvolvido um *prompt* contendo um conjunto de instruções que definem os critérios utilizados para a classificação dos elementos do documento. Estes critérios baseiam-se predominantemente em características estruturais e tipográficas, permitindo o modelo identificar a hierarquia e a organização do conteúdo. Entre os principais critérios considerados destacam-se:

- Presença de numeração hierárquica
- Tamanho relativo da fonte
- Tipo de fonte
- Padrões tipográficos consistentes no documento

Utilizou-se a relação semântica entre os subtítulos apenas como um critério auxiliar, para os casos em que os sinais estruturais eram insuficientes para uma decisão coerente. Adicionalmente, aplicou-se a técnica de *few-shot*, que incluiu um exemplo do tipo de resposta que o modelo deve retornar.

O código utilizado para esta etapa encontra-se no Código 6.

```
You will receive a list of titles with font type and size.
List of titles: {lista_titulos}
Your task is to assign each title a hierarchical level (1-5)
using ONLY visual cues (numbering, font type, font size).
Do NOT rely primarily on semantic interpretation.

-----
#PRIORITY ORDER (strict)
1. Prefix / Numbering (highest priority)
  - Detect explicit hierarchical prefixes (e.g., 1, 1.1, I, I.A,
  A.1).
  - If ALL titles contain consistent, ordered prefixes,
    infer levels directly from numbering.

2. Font characteristics
  - Larger font size → higher level.
  - Same font type AND same size as previous title → same level.
  - If all titles share the same font and size,
    font provides NO hierarchy signal.

3. Semantic relationship (fallback only)
  - Use semantic inference ONLY if numbering and font
    provide no structural distinction.
  - Do NOT override numbering or font-based structure.
-----
#HIERARCHY CONSISTENCY RULES
  - Levels cannot increase by more than +1 at a time.
  - A level 3 cannot appear without a preceding level 2.
  - If hierarchy requires assumptions beyond visual cues it is not
    structurally strong.
  - If two or more titles share the same numbering prefix (e.g.,
    "2.1. Definition" and "2.1. Definição"), they represent the SAME
    section.
    -- Return ONLY ONE title for that numbering prefix.
    -- When duplicates exist in different languages, keep the
    ENGLISH version only.
    -- Never return two titles with the same numbering prefix.
  - Return the exact same titles

NOTE: All titles must be in English.
-----
#OUTPUT FORMAT: Return JSON only

## Example output:
{{
  "sections": [
    {{ "title": "1. INTRODUCTION", "level": 1}},
    {{ "title": "1.1 Methodology", "level": 2}},
    {{ "title": "A) SAMPLE", "level": 2}},
```

```

    [{"title": "- Participants", "level": 3}]
  }}

```

Código 6: Estrutura do *prompt* para a definição dos níveis hierárquicos dos subtítulos

Para esta tarefa, utilizou-se inicialmente o modelo LLaMA 3.1 8 B. O modelo apresentou bons resultados, sempre que os títulos estavam corretamente numerados. No entanto, quando essa estrutura não era apresentada, o modelo apresentava várias falhas.

Nomeadamente, verificou-se que o modelo atribuiu níveis hierárquicos incorretos a alguns títulos. Um exemplo ocorreu no documento “*WHO recommendations on care for women with diabetes during pregnancy*”, em que foram observados dois tipos de problemas:

- Atribuição incorreta de níveis hierárquicos em algumas secções. Por exemplo, para o subtítulo “*Annex 2. Priority outcomes used in decision making*”, o modelo classificou nível 2, quando deveria corresponder ao nível 1.
- Falha na identificação de subtítulos. Ou seja, alguns subtítulos desta secção foram ignorados pelo modelo, como por exemplo “*Priority outcomes included for all PICOs*”, “*Critical maternal outcomes*” e “*Critical newborn outcomes*”.

Devido a estas limitações, optou-se por testar um modelo de maior dimensão, o LLaMA 3.3 70B, que possui uma maior capacidade de interpretação contextual. Alguns dos resultados da classificação hierárquica dos títulos, para o documento utilizado como exemplo, encontram-se na Tabela 9: Comparação dos níveis hierárquicos atribuídos pelos modelos LLaMA 3.1 8B e LLaMA 3.3 70B na classificação de subtítulosTabela 9.

Tabela 9: Comparação dos níveis hierárquicos atribuídos pelos modelos LLaMA 3.1 8B e LLaMA 3.3 70B na classificação de subtítulos

Exemplos de títulos	Modelo LLaMA 3.1 8B	Modelo LLaMA 3.3 70B	Nível hierárquico esperado
4 Dissemination and implementation	Nível 1	Nível 1	Nível 1
4.1 Dissemination and evaluation	Nível 2	Nível 2	Nível 2
5.3 Pharmacotherapy	Nível 2	Nível 2	Nível 2
5.3.1 Type 1 diabetes mellitus	Nível 3	Nível 3	Nível 3
Annex 2. Priority outcomes used in decision making	Nível 2	Nível 1	Nível 1
Priority outcomes included for all PICOs	--	Nível 2	Nível 2
Critical maternal outcomes	--	Nível 3	Nível 3
Critical newborn outcomes	--	Nível 3	Nível 3

Os resultados indicam que o modelo LLaMA 3.3 70 B apresenta maior consistência na recuperação da hierarquia dos títulos do documento, conseguindo identificar corretamente a maioria dos títulos e respetivos níveis hierárquicos, incluindo aqueles que não apresentam numeração explícita. Em particular, observou-se uma melhoria na capacidade de inferir relações hierárquicas em casos onde o modelo LLaMA 3.1 8B demonstrava várias limitações.

5.4.3 Etapa 3: Limpeza e preparação final das secções de texto

A terceira etapa do pipeline corresponde à fase de limpeza e preparação final das secções textuais extraídas dos artigos biomédicos. O principal objetivo desta etapa consiste na remoção de conteúdo estruturalmente irrelevante para tarefas subsequentes de indexação, representação semântica e recuperação de informação, reduzindo o ruído presente no corpus e melhorando a qualidade do processamento.

Desta forma, pretende-se excluir secções que não contribuem para o conteúdo clínico ou científico do documento, uma vez que estas podem introduzir informação redundante ou não informativa no processo de construção de representações textuais.

Segundo Cuconasu et al. [185], o desempenho de sistemas RAG depende não só da recuperação de conteúdo relevante, mas também da gestão de ruído no contexto, sendo que a presença de informação não relevante pode afetar o desempenho dos modelos. Neste sentido, a remoção de conteúdo não informativo, como dados relativos aos autores e secções como o índice e os agradecimentos, contribui para a qualidade do processamento subsequente.

- Filtragem de secções não informativas

Com base na análise exploratória do conjunto de documentos processados e nas convenções estruturais típicas de artigos biomédicos, identificaram-se secções que apresentam, de forma consistente, baixo valor informativo para os objetivos do sistema. A seguir, listam-se alguns exemplos de títulos destas secções:

- “References”;
- “Acknowledgements”;
- “Affiliations”;
- “Authors”;
- “Index”;

Estas categorias correspondem, em geral, a informações que não estão diretamente relacionadas com o conteúdo científico ou clínico do artigo. Assim, nesta etapa, procedeu-se à remoção desses elementos, uma vez que não eram relevantes para o funcionamento do sistema.

- Limitações na filtragem de informação acerca de autores

Durante esta etapa, explorou-se uma abordagem adicional para a identificação e remoção de blocos de texto contendo nomes de autores, contactos institucionais e endereços de email. Esta informação, frequentemente presente no início dos artigos, revelou-se problemática, uma vez que pode ser incluída inadvertidamente na base de dados de segmentos textuais e influenciar etapas subsequentes de recuperação de informação.

No entanto, os resultados desta abordagem não foram satisfatórios. A utilização de um modelo de linguagem para a identificação deste tipo de informação revelou algumas limitações. Verificou-se que o modelo apresentava dificuldades na distinção entre nomes próprios e termos utilizados em contexto biomédico, nomeadamente designações de procedimentos, técnicas ou instrumentos médicos, o que conduziu à remoção indevida de conteúdo relevante. Em simultâneo, observou-se também a não identificação consistente de alguns nomes de autores, sobretudo em casos de maior ambiguidade ou elevada frequência lexical, como “Oliveira” ou “Coelho”.

Face a estas limitações e ao risco de perda de informação cientificamente relevante, esta abordagem não foi integrada na versão final do sistema.

5.5 Pipeline de Extração de figuras dos documentos

Após a extração do texto, foi desenvolvido um pipeline dedicado à extração e processamento de figuras presentes nos documentos. O principal objetivo deste processo é descrever automaticamente o conteúdo de cada figura, convertendo a informação visual num texto estruturado. Desta forma, as figuras podem ser indexadas e posteriormente recuperadas pelo modelo RAG durante o processo de pesquisa.

O *pipeline* inicia-se com a análise dos resultados produzidos pelo PaddleOCR, uma ferramenta que localiza e extrai automaticamente todas as figuras dos documentos. De seguida, cada figura é submetida a uma etapa de classificação, na qual é atribuída uma de duas categorias: fluxograma ou imagem ilustrativa.

Em função da classificação atribuída, cada figura segue um fluxo de processamento específico. Apesar de serem aplicadas abordagens distintas, o resultado é o mesmo: a geração de uma descrição textual detalhada do conteúdo da figura e de um conjunto estruturado de metadados, que permitem a sua indexação e posterior recuperação pelo modelo RAG.

A Figura 15 apresenta o fluxograma correspondente ao *pipeline* de extração e processamento de figuras desenvolvido nesta etapa.

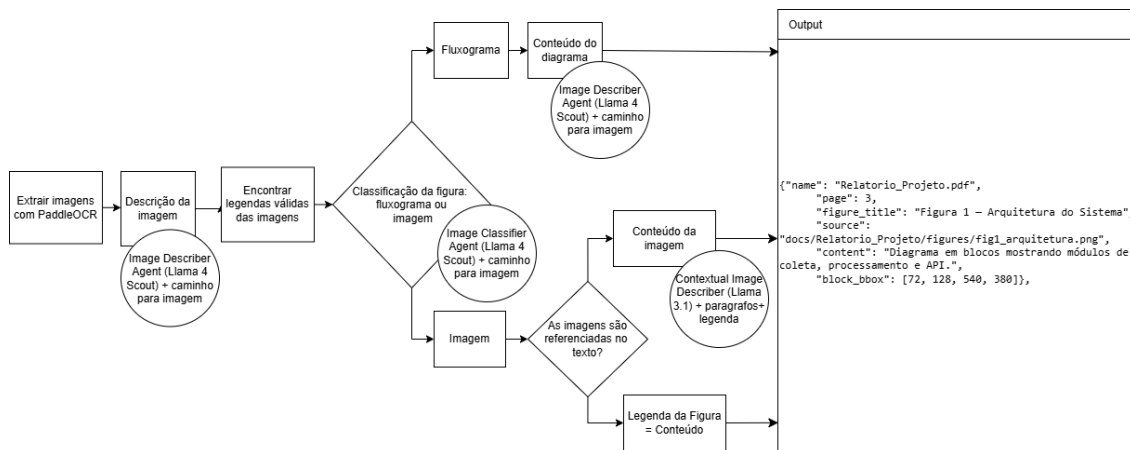


Figura 15: Fluxograma para descrever a extração de imagens de documentos

Conforme ilustrado na Figura 15, o pipeline organiza o processamento das figuras em diferentes etapas, desde a sua extração e classificação até à geração das descrições textuais e dos metadados associados. Esta abordagem adapta o processamento ao tipo de figura identificado, garantindo uma representação textual adequada para ser posteriormente indexado e recuperado por um modelo RAG. As secções seguintes descrevem detalhadamente cada uma das etapas que compõem este pipeline, apresentando os *prompts* desenvolvidos, bem como a fundamentação e o raciocínio adotados na definição da estratégia de processamento para cada tipo de figura.

5.5.1 Classificação do tipo de figura

As figuras presentes nos documentos científicos apresentam diferentes características. Enquanto fluxogramas e diagramas contêm frequentemente informação textual e relações espaciais entre elementos, como caixas, setas e sequências de decisão, imagens ilustrativas, fotografias clínicas ou ecografias dependem frequentemente do contexto textual fornecido pelo documento. Desta forma, a estratégia utilizada para gerar a descrição textual da figura deve ser adaptada ao tipo de conteúdo visual.

Para este trabalho, as figuras foram agrupadas em duas categorias:

- Fluxogramas: figuras que representam processos, algoritmos ou sequências de decisão através de elementos gráficos estruturados, como caixas, setas, ligações e etapas sequenciais. A interpretação destas figuras depende tanto do conteúdo textual presente na imagem como da relação espacial entre os diferentes elementos.
- Imagens Ilustrativas: figuras cujo objetivo principal é representar visualmente um conceito, procedimento, estrutura anatómica ou exemplo clínico. Incluem, por exemplo, ilustrações médicas, imagens clínicas, fotografias e ecografias. A sua

interpretação depende frequentemente do contexto textual presente na legenda e nos parágrafos que referenciam a respetiva imagem.

As figuras que não se encaixam em nenhuma destas categorias, ficam classificadas como “imagens ilustrativas”.

Numa fase inicial, foram avaliados três modelos multimodais para realizar diretamente a classificação das figuras, nomeadamente o LLaVA 7B, executado localmente, e os modelos LLaMA 4 Maverick 17B 128E e LLaMA 4 Scout 17B 16E, disponibilizados através da plataforma GROQ. Os resultados obtidos encontram-se na Secção 6.3, onde os modelos foram utilizados para identificar o tipo de figura (fluxograma ou imagem ilustrativa) e descrever a imagem.

Como foi observado, e apresentado nessa Secção, todos os modelos apresentaram falhas na classificação das figuras, mas acertaram a realizar a descrição da figura. Por esse motivo, experimentou-se uma segunda abordagem, onde o modelo multimodal era utilizado apenas para descrever as figuras, e posteriormente, utilizava-se um LLM para classificar a figura como imagem ilustrativa ou fluxograma, com base nessa mesma descrição. Uma vez que os resultados obtidos com esta abordagem revelaram menos falhas, optou-se por esta abordagem.

O modelo utilizado foi o LLaMA 4 Scout 17B 16E. O *prompt* utilizado na primeira etapa é simples e solicita apenas uma descrição objetiva do que aparece na figura, com um limite máximo de 50 palavras. A estrutura do *prompt* aplicado encontra-se no Código 7.

```
You are a visual content analyzer.  
Observe the image and describe briefly what you see visually,  
focusing only on the elements present (text, boxes, diagrams,  
or visuals).  
Respond with one short sentence with the limit of 50 words  
maximum.
```

Código 7: Estrutura do *prompt* aplicado ao modelo multimodal para descrever as figuras

O resultado enviou-se para um segundo LLM, com o objetivo de classificar a figura com base na descrição da figura. O *prompt* encontra-se no Código 8.

```
Based on the text, tell me if it corresponds to a flowchart or  
figure.  
The text is {texto}  
The response should be in JSON format, with the following format:  
{"type": "<flowchart/figure>"}
```

Código 8: Estrutura do *prompt* aplicado ao modelo multimodal para descrever as figuras.

5.5.2 Estratégias para descrição de figuras

Como referido anteriormente, para que o agente conversacional recupere as figuras mais relevantes em resposta às questões do utilizador, as imagens precisam ser descritas sob a forma

de texto. Esta representação estabelece, durante a etapa de recuperação de informação, uma correspondência semântica entre a *query* do utilizador e o conteúdo das figuras.

No entanto, nos documentos científicos, as figuras são frequentemente acompanhadas por legendas detalhadas e por referências ao longo do texto que explicam o seu conteúdo e contextualizam a informação apresentada. Por serem elaboradas pelos próprios autores, estas descrições tendem a ser mais completas, rigorosas e tecnicamente precisas do que aquelas geradas automaticamente por modelos multimodais de propósito geral. Desta forma, foram exploradas duas estratégias para representar o conteúdo das figuras: (i) uma abordagem baseada na informação disponível no próprio documento científico, recorrendo à legenda e aos parágrafos que fazem referência à figura; e (ii) uma abordagem baseada na utilização de um modelo de linguagem multimodal para gerar automaticamente uma descrição a partir da imagem.

Com base na avaliação experimental apresentada no Capítulo 6, foi adotada uma estratégia distinta para cada tipo de figura. Enquanto os modelos multimodais apresentaram melhores resultados na interpretação de fluxogramas, a utilização da informação textual presente no documento revelou-se mais eficaz para a descrição de imagens ilustrativas. As estratégias adotadas encontram-se descritas a seguir:

- Estratégia 1: Abordagem baseada no texto presente no documento científico para descrever figuras

Nesta abordagem, recorreu-se a um LLM para gerar automaticamente uma descrição da figura com base na informação textual presente no documento científico. Para tal, utilizaram-se duas fontes de informação complementares: a legenda associada à figura e os parágrafos do documento que fazem referência a essa figura. A identificação dos parágrafos relevantes foi feita através da deteção das referências às figuras no texto, tais como "Figura 3" ou "Fig. 3". Após a recolha da legenda e dos respetivos excertos textuais, esta informação foi enviada para o LLM, que gerou uma descrição do conteúdo da figura com base exclusivamente nesses elementos.

De forma a garantir que a descrição produzida refletisse fielmente a informação disponibilizada pelos autores, foi desenvolvido um *prompt* específico para orientar o modelo durante a geração da resposta. O modelo foi instruído a não inferir, completar ou introduzir informação visual que não estivesse explicitamente descrita no documento. O *prompt* utilizado nesta etapa encontra-se no Código 9.

```
You will receive several excerpts from an obstetric guideline:
{paragrafos}. And the caption of a figure: {legenda}
Based on these excerpts and the figure caption, describe in words
what the figure is likely showing.
Focus only on the information present in the text and in the
caption, without imagining or creating visual details beyond what
is implied in the text.
```

Código 9: Estrutura do *prompt* modelo multimodal LLaMA 4 Scout 17B 16E para descrever imagens

- Estratégia 2: Abordagem com Modelo Multimodal para descrever figuras

Nesta abordagem, recorreu-se a modelos multimodais capazes de processar simultaneamente informação visual e textual. Estes modelos abrangem uma componente visual, responsável pela extração de características da imagem, com uma componente linguística de grande escala, capaz de traduzir essas informações em descrições de linguagem natural.

Para implementar esta abordagem foi utilizado o modelo multimodal LLaMA 4 Scout 17B 16E, que permite interpretar imagens e gerar descrições textuais contextualizadas. A sua seleção baseou-se nos resultados obtidos durante a avaliação experimental, onde demonstrou maior capacidade para interpretar corretamente o conteúdo visual das figuras em comparação com os restantes modelos analisados.

Para esta etapa foi desenvolvido um *prompt* que teve como objetivo orientar o modelo para produzir descrições técnicas e relevantes para o domínio biomédico, garantindo simultaneamente consistência nas respostas.

O *prompt* orienta o modelo a resolver a tarefa de forma sequencial, começando por identificar os principais elementos do fluxograma, seguindo as ligações entre os diferentes nós e, por fim, construindo os percursos existentes. No final, é especificado um formato de resposta rigoroso, de forma a garantir que o resultado apresentado contém apenas os percursos identificados. O *prompt* utilizado nesta etapa encontra-se no Código 10.

```
You are extracting paths from a flowchart.
Follow the internal procedure below, but DO NOT show the reasoning
or steps.
INTERNAL PROCEDURE (do not output):
  1) Identify all MAIN NODES:
    - Treat each large outer box as a main node.
    - If theres an internal text inside a large box belong to that
    main node and are not separate nodes.
  2) Identify the START NODE:
    - Only when a main node has no outgoing arrows.
  3) Identify all END NODES:
    - Main nodes with no outgoing arrows.

  4) Identify all valid paths:
    - Build paths from the START NODE to each END NODE.
    - When constructing paths, consider only MAIN NODES.
    - Follow only explicit arrows/connectors between MAIN NODES.
    - Do not treat internal boxes as independent steps unless
    explicitly connected by arrows outside the main box.
  5) Represent internal content:
    - After constructing the path using MAIN NODES, include each
    node's internal content in parentheses after the node title.
  6) Include arrow labels:
    - If an arrow between two MAIN NODES has visible text, include
    it formatted as: -[label]→
    - If no label is visible, use: →

OUTPUT RULES (STRICT):
  - Output ONLY the final paths.
  - DO NOT output explanations.
```

- DO NOT output steps.
- DO NOT output reasoning.
- DO NOT output node lists.
- If anything other than the paths is written, the answer is incorrect.

OUTPUT FORMAT (ENGLISH ONLY):

Path 1:

Main Node A (internal content) → [label] → Main Node B (internal content) → Main Node C (internal content)

Código 10: Estrutura do *prompt* aplicado ao modelo multimodal para descrever as figuras

Este *prompt* que teve por objetivo transformar os fluxogramas clínicos em sequências estruturadas de decisão. Foi aplicado ao modelo multimodal LLaMA 4 Scout 17B 16E e testado em diferentes formatos de fluxogramas. Os resultados encontram-se no Anexo A.

Quando a figura era classificada como fluxograma, o seu processamento era realizado exclusivamente a partir do conteúdo visual, recorrendo a um modelo multimodal responsável por gerar uma descrição detalhada com base na análise da imagem.

De forma geral, o modelo identificou corretamente a maioria dos nós e das ligações entre etapas, permitindo converter os fluxogramas em representações textuais sequenciais. No entanto, observaram-se algumas limitações na interpretação de certos diagramas, particularmente quando os elementos gráficos eram mais complexos. Nomeadamente, em casos onde apareciam imagens de fundo (como em anexo Figura A1) e imagens com diferentes cores de caixas (como as figuras em anexo A2 e A3). Estes elementos, apesar de terem apenas uma função visual ou de organização no diagrama, podem ser interpretados incorretamente pelo modelo como indicadores estruturais do fluxo.

5.6 Pipeline de Extração de Tabelas

Nesta seção descreve-se a metodologia adotada para a extração e processamento das tabelas presentes nos documentos. O processo iniciou-se com a análise dos resultados obtidos pela PaddleOCR. Para cada tabela identificada, procedeu-se à deteção da respetiva legenda. Caso esta biblioteca não reconhecesse automaticamente a legenda, recorria-se a um LLM para inferir, a partir do conteúdo textual da página e da própria estrutura da tabela, qual a legenda mais provável.

Posteriormente, foi aplicada uma segunda etapa para identificar notas de rodapé associadas à tabela, tais como abreviaturas, definições ou esclarecimentos adicionais relevantes para interpretar a tabela. O resultado deste processo guardou-se em formato JSON, contendo a tabela estruturada e um conjunto de metadados.

O fluxograma da Figura 16 apresenta os processos adotados nesta etapa.

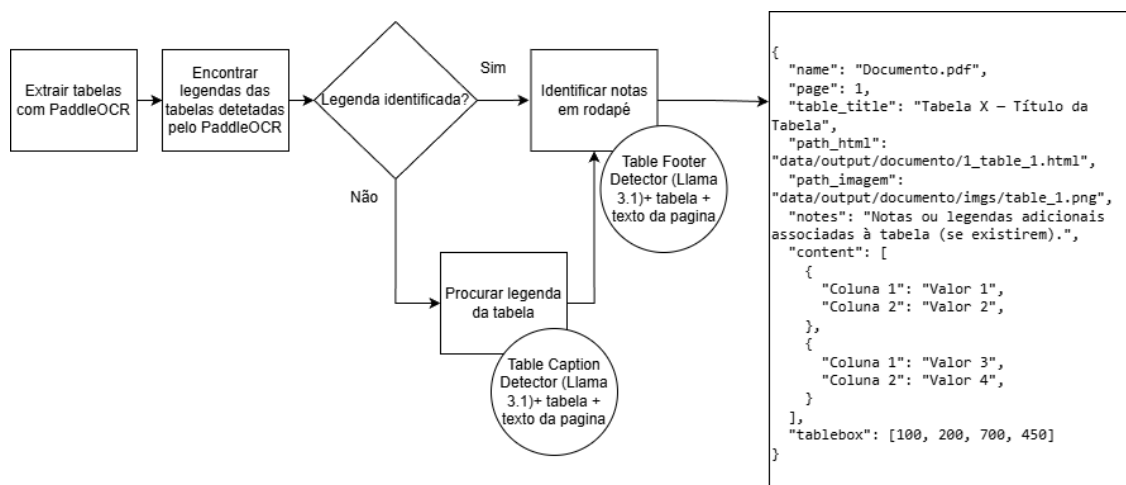


Figura 16: Fluxograma relativo à extração de tabelas

Nas normas e documentos clínicos, é frequente existirem notas relativas às tabelas contendo informações complementares, como a explicação de abreviaturas, a descrição de símbolos, condições experimentais ou exceções. Apesar de serem geralmente curtas, estas notas contêm um contexto adicional essencial para a correta interpretação dos dados apresentados nas tabelas, tornando a sua extração relevante.

Este processo foi realizado utilizando um LLM, responsável por analisar o conteúdo textual da tabela em conjunto com o texto completo da página onde esta se encontra. Este processo realizou-se enviando para um LLM um *prompt* que continha a tabela, e todo o texto da página onde ficava a tabela. O modelo foi instruído a identificar se existiam notas em rodapé associadas e, em caso afirmativo, a devolvê-la em formato JSON.

O *prompt* contém algumas orientações específicas, símbolos e convenções que costumam aparecer, como por exemplo ¹, ², *, ou †. Por fim, foi imposto um formato de resposta rigoroso, exigindo que o modelo devolvesse apenas um objeto JSON válido, sem texto adicional. O *prompt* utilizado encontrar-se no Código 11.

```

You are given a table extracted from a PDF and the full-page text
where the table appears.
Table: {tabela}; Page text: {pagina_conteudo_str};
Your task is to check whether the page contains any additional
notes, explanations of abbreviations or footnotes related to the
table above.
These may include:
- Explanation of numbers or explanatory symbols (such as 1, 2, *,
†) that appear in the table
Return ONLY a JSON object with the following structure, don't add
more text: {"notes_found": true or false, "notes": "If any notes
or explanations are found, include them here as a single text block.
If not, return an empty string.}}
  
```

Código 11: Estrutura do *prompt* para o LLM LLaMa 3.3 70B encontrar notas em rodapé relativas à tabela

O modelo produziu respostas no formato especificado, quer nos casos em que não foram identificadas notas em rodapé, quer quando estas estavam presentes. Seguem-se dois exemplos de respostas produzidas pelo modelo:

```
{'notes_found': False, 'notes': ''}  
{'notes_found': True, 'notes': 'IMC, índice de massa corporal; TEV, tromboembolismo venoso.\n1 Homozigotia para o fator V de Leiden ou para a mutação da Protrombina G20210A, dupla heterozigotia\npara fator V de Leiden e mutação da Protrombina G20210A, défice de antitrombina, síndrome de\nanticorpos antifosfolípidos.\n2 Apendicectomia, esterilização no pós-parto.\n3 Por exemplo, cancro, insuficiência cardíaca, lupus eritematoso sistémico ativo, poliartrópia\ninflamatória, doença inflamatória intestinal, síndrome nefrótica, diabetes mellitus tipo I com\nnefropatia, anemia falciforme, uso de drogas intravenosas.\n4 Heterozigotia para o fator V Leiden ou para a mutação da Protrombina G20210A, défice de\nproteína C, défice de proteína S, persistência de anticorpo antifosfolípido.\n5 Por exemplo, pielonefrite e pneumonia.'}
```

Código 12: Resultados obtidos com a utilização do LLM para a recuperação das notas de rodapé das tabelas.

Embora esta abordagem tenha produzido resultados satisfatórios, a sua eficácia depende da correta extração do texto e do bom funcionamento da biblioteca PaddleOCR. Erros introduzidos durante a conversão do documento ou a presença de múltiplas tabelas na mesma página podem dificultar a correta associação entre as notas em rodapé e a tabela correspondente.

5.7 Interface do Utilizador

A interface desenvolvida permite ao utilizador realizar todas operações necessárias para a utilização do sistema, incluindo o carregamento de novos documentos e a interação com o *chatbot*.

A Figura 17 apresenta a página inicial, onde o utilizador pode aceder às principais funcionalidades do sistema.



Figura 17: Interface principal do sistema de *Chatbot* baseado em RAG

Após selecionar a opção de adicionar documentos, é apresentada a interface ilustrada na Figura 18, que permite importar novos documentos para o processamento, e onde o utilizador tem de informar se o documento se trata de um artigo científico ou de uma norma clínica. Nesta interface o utilizador pode adicionar um ficheiro PDF e, ao carregar no botão “Guardar Documento”, o sistema inicia o processo de pre-processamento de documentos e de indexação.

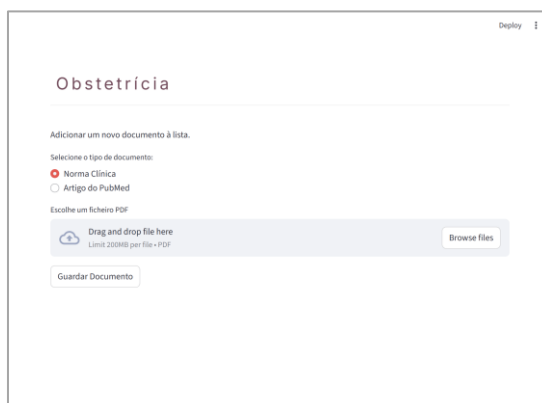


Figura 18: Interface de carregamento de documentos para ingestão na base de conhecimento

Depois de concluída a indexação, o utilizador pode interagir com o chatbot através da interface apresentada na Figura 20. As perguntas submetidas são processadas pelo sistema RAG, sendo apresentadas as respostas juntamente com os documentos utilizados como contexto, conforme o que aparece na Figura 19.

A fase de interação com o utilizador inicia-se após a seleção da funcionalidade *Chatbot* na interface principal da aplicação, Figura 17. Esta ação conduz à interface de conversação apresentada na Figura 19, desenvolvida para permitir a interação entre o utilizador e o sistema RAG.

A interface é composta por duas áreas principais. A área central corresponde ao espaço de conversação, onde são apresentadas as mensagens trocadas entre o utilizador e o *chatbot* e onde se encontra a caixa de texto destinada à introdução de novas questões. A barra lateral disponibiliza o histórico das conversas anteriormente realizadas, permitindo ao utilizador retomar sessões anteriores. Cada conversa é identificada pelo início da primeira questão submetida, bem como pela respetiva data e hora de criação.

Adicionalmente, a interface inclui um botão Novo Chat, que permite iniciar uma nova sessão de conversação, e um botão Voltar, utilizado para regressar à interface principal da aplicação.

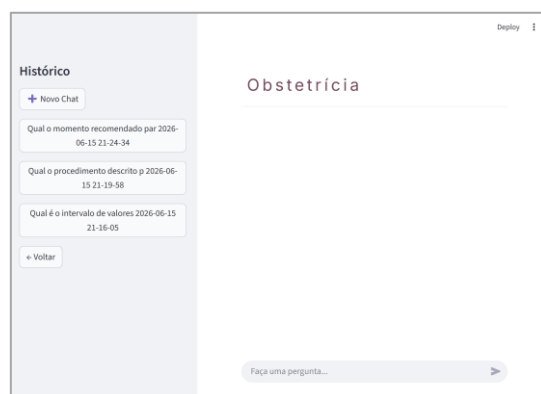


Figura 19: Interface do *chatbot* utilizada para interação com o sistema

Após a submissão de uma questão, o sistema processa o pedido e apresenta a resposta na interface do *chatbot*, conforme ilustrado na Figura 20.

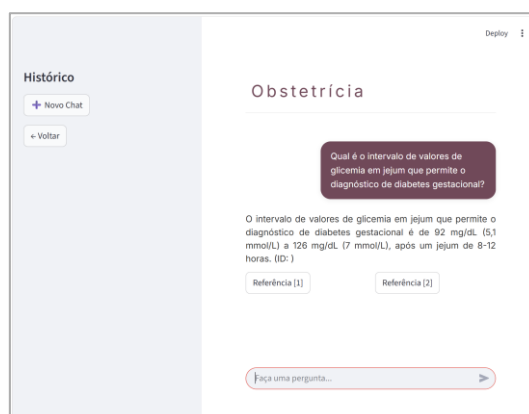


Figura 20: Interface do sistema de *chatbot* demonstrando uma interação com o utilizador, incluindo a submissão de uma pergunta e a respetiva resposta gerada pelo sistema.

Adicionalmente, o sistema disponibiliza ao utilizador as referências correspondentes ao contexto recuperado que suportou a geração da resposta. Cada referência pode ser seleccionada para consultar o segmento de texto, imagem ou tabela utilizado durante o processo de recuperação de informação, conforme ilustrado na Figura 21.

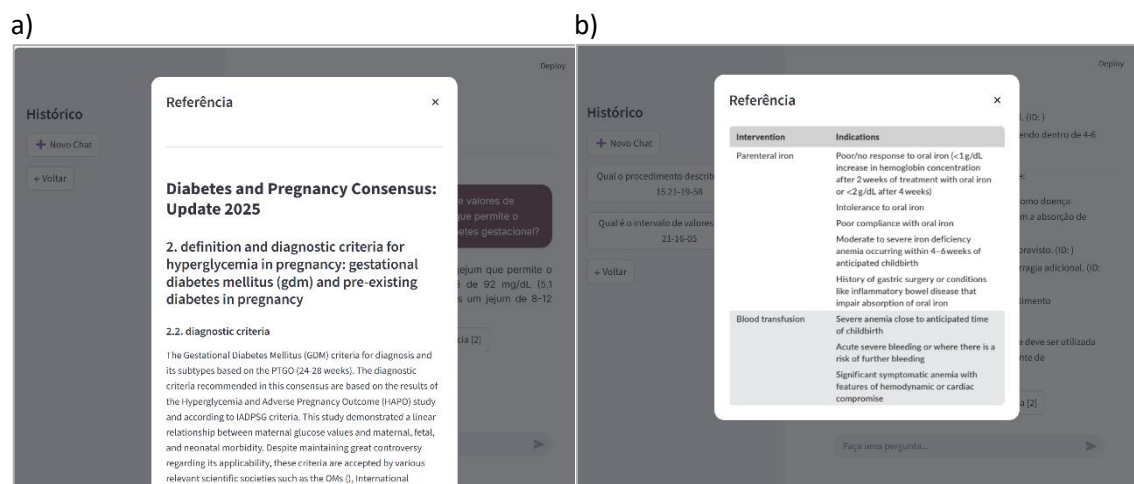


Figura 21: a) Exemplo de referência do tipo texto; b) Exemplo de referência do tipo tabela;

As funcionalidades apresentadas nesta secção demonstram a forma como o utilizador interage com o sistema, desde o carregamento dos documentos até à consulta das respostas geradas pelo *chatbot* e das respetivas referências. No capítulo seguinte, são apresentados os procedimentos de avaliação efetuados, bem como a análise dos resultados obtidos para os diferentes componentes do sistema.

6 Experimentação

Neste capítulo são apresentados os experimentos realizados para avaliar os diferentes componentes desenvolvidos no sistema proposto. A avaliação experimental foi dividida em duas vertentes principais. A primeira vertente incide sobre o processo de pré-processamento dos documentos científicos, onde são avaliadas as ferramentas utilizadas para extração e estruturação da informação documental. A segunda vertente corresponde à avaliação do sistema RAG, sendo analisado o impacto de diferentes modelos de *embeddings* na recuperação de informação e de diferentes modelos de linguagem na geração das respostas finais.

Para a avaliação do sistema RAG, foi criado um conjunto de perguntas baseado nos documentos científicos utilizados, permitindo analisar o desempenho das diferentes estratégias de recuperação e geração. Numa primeira etapa, são comparados diferentes modelos e estratégias de *embeddings*, com o objetivo de avaliar a capacidade de recuperar os elementos documentais mais relevantes. Posteriormente, são avaliados diferentes modelos de linguagem, responsáveis por gerar as respostas com base no contexto recuperado.

6.1 Avaliação de ferramentas para o pré-processamento de documentos

Nesta seção, foram avaliadas diferentes bibliotecas para usar na etapa de pré-processamento do sistema. O objetivo é avaliar o desempenho e encontrar a melhor ferramenta para extrair a informação proveniente de artigos científicos indexados no Pubmed, bem como em normas clínicas, que podem estar redigidas em português ou inglês.

A seguir apresentam-se algumas bibliotecas e modelos testados para detetar e extrair elementos como títulos/secções, tabelas e figuras.

- PyMuPDF

O PyMuPDF é uma biblioteca em Python, que dá ligação à MuPDF em linguagem C. É uma biblioteca *open-source* criada para extrair texto, tabelas e imagens disponibiliza o método `find_tables`, que automatiza a deteção e a extração de tabelas diretamente a partir do PDF. As grelhas identificadas podem ser extraídas (utilizando `table.extract`) e convertidas para DataFrames do Pandas. O PyMuPDF pode apoiar a extração automática de títulos e subtítulos. Quando o PDF disponibiliza estrutura de navegação (*bookmarks*), é possível reconstruir a hierarquia de seções através de `get_toc`. O resultado é uma lista com título, nível hierárquico e página. Em relação à extração de imagens, o modelo reconhece imagens de documentos pdf usando a função `extract_image`. O resultado é os bytes da imagem e metadados da imagem [186].

- PDFplumber

O PdfPlumber é uma biblioteca baseada no pdfminer, usada para analisar e extrair conteúdo de PDFs. A biblioteca é usada para detetar e extrair tabelas, com a função `extract_tables`. É capaz de extrair tabelas com linhas visíveis ou implícitas pelo alinhamento e tabelas com linhas ou colunas unidas. Também pode ser usada para identificar imagens, utilizando a `page.images`, que devolve as coordenadas das imagens que aparecem no documento em pdf. Por fim, pode ser usada para extrair texto, utilizando a `extract_words`, a biblioteca extrai texto de uma pagina em específico, ou para extrair o texto de todo o documento, usa-se a função `extract_text` [187].

- TATR

O TATR (Table Transformer), desenvolvido pela Microsoft, é um modelo de DL orientado para a deteção e estruturação de tabelas em documentos. A sua arquitetura baseia-se no DETR (DEtection TRansformer), um modelo de deteção de objetos baseado em Transformers, no entanto, o TATR foi treinado com a base de dados PubTables-1M, para ser capaz de identificar e estruturar tabelas de artigos científicos.

O TATR é composto pelos modelos “microsoft/table-transformer-detection”, responsável pela identificação de tabelas, e “microsoft/table-transformer-structure-recognition” responsável pelo reconhecimento da estrutura interna das tabelas, incluindo linhas e colunas [188].

- PaddleOCR

PaddleOCR (versão 3.1.0) é uma ferramenta *end-to-end*, desenvolvida pelo *PaddlePaddle*. Esta ferramenta combina modelos de DL com capacidade multimodal, permitindo a análise com documentos complexos em formato de imagem ou pdf. O PaddleOCR integra o PP-StructureV3, um pipeline de análise de layout para extrair informação estruturada de imagens e PDFs. Combina OCR, visão computacional e ML para identificar/segmentar texto, títulos, parágrafos, imagens, tabelas e outros elementos dos documentos[189]. É por isso, constituída por 7 módulos, cada um com uma tarefa específica [189]:

- a) *Layout Detection Module*: Segmenta a página em regiões semânticas, atribuindo a cada uma caixa delimitadora (coordenadas), um rótulo de classe e uma confiança. Deteta 23 classes, entre os quais tem *document title*, *paragraph title*, *texto*, *image* e *table*;

- b) *General OCR Subline*: Executa OCR nas regiões de texto identificadas, convertendo conteúdo de imagens ou PDFs em texto editável. Permite escolher o modelo por idioma. Para documentos em português é recomendado o `latin_PP-OCRv5_mobile_rec`, um modelo otimizado para alfabetos latinos.
- c) *Document Image Preprocessing Subline* (Opcional): classifica a orientação do documento e retifica a imagem (corrige rotação, inclinação e perspectiva dos documentos)
- d) *Table Recognition Subline* (opcional): reconhece estruturas de tabelas e extrai o conteúdo para formatos estruturados.
- e) *Seal Text Recognition Subline* (Opcional): reconhece o texto de selos/carimbos, útil para a análise de documentos oficiais.
- f) *Formula Recognition Subline* (Opcional): detecta e extrai automaticamente o conteúdo e a estrutura de fórmulas de documentos ou imagens.
- g) *Chart Parsing Module* (Opcional): interpreta gráficos visuais (como gráficos de barras, gráficos de linhas, etc) e converte para tabelas de dados estruturados.

Em conjunto, estes módulos permitem ao PaddleOCR realizar uma análise estruturada de documentos, extraindo não apenas o texto, mas também a organização lógica do conteúdo, incluindo títulos, parágrafos, tabelas, imagens e outros elementos gráficos.

- Nougat

O Nougat (*Neural Optical Understanding for Academic Documents*) é um modelo *Transformer* visual *encoder-decoder*, baseado na arquitetura Donut, que adapta técnicas recentes de *Visual Document Understanding* para documentos científicos. Este modelo utiliza um *encoder* visual baseado em *Swin Transformer* e um *decoder* autorregressivo semelhante ao mBART. O modelo é treinado de forma *end-to-end* com pares de imagens de páginas e texto em markup, gerados automaticamente a partir de artigos científicos (como arXiv e PMC). Durante o treino, o modelo aprende diretamente a reconhecer texto, layout e expressões matemáticas a partir da imagem, sem utilizar motores OCR tradicionais [190].

Ao contrário das abordagens tradicionais baseadas em OCR, o Nougat gera diretamente uma representação estruturada do documento em formato Markdown, preservando elementos como títulos, listas, tabelas, referências bibliográficas e expressões matemáticas [190].

A extração de informação de documentos PDF pode ser realizada recorrendo a diferentes abordagens, cada uma com características e capacidades distintas. A Tabela 10 apresenta uma síntese das principais bibliotecas utilizadas na extração e análise de conteúdo de documentos PDF, amplamente referidas na literatura. Para cada biblioteca são identificadas a abordagem adotada (baseada em regras, OCR ou Deep Learning), bem como as principais funcionalidades disponibilizadas, incluindo a extração de texto, tabelas, imagens e a análise da estrutura dos documentos.

Tabela 10: Ferramentas de *parsing* e extração de informação em documentos científicos

Biblioteca	Tipo de Abordagem	Funcionalidades Principais	Referências
PyMuPDF (fitz)	<i>Rule-based</i>	Extração de texto, imagens (extract_image), tabelas (find_tables), reconstrução de índice (get_toc)	[186]
PDFPlumber	<i>Rule-based</i>	Extração de texto (extract_text, extract_words), extração de tabelas (extract_tables), detecção de imagens (page.images).	[187]
PaddleOCR	<i>OCR + Deep Learning</i>	Análise de layout, extração de texto com OCR multilíngue, tabelas, fórmulas, imagens e gráficos.	[189]
Nougat	<i>Deep Learning</i>	Extração de texto, equações matemáticas e tabelas	[190]
TATR	<i>Deep Learning</i>	Deteção de tabelas e reconhecimento estrutural (linhas, colunas, cabeçalhos)	[188]

6.1.1 Resultados

Nesta secção apresentam-se os resultados e comparações relativos ao desempenho das ferramentas mencionadas anteriormente, na análise de documentos biomédicos, nomeadamente normas clínicas e artigos científicos em inglês e português.

Para esse efeito, analisaram-se as ferramentas quanto à sua capacidade de extrair tabelas com diferentes formatos, bem como imagens e títulos de secções de texto.

- pdfPlumber e PyMuPDF

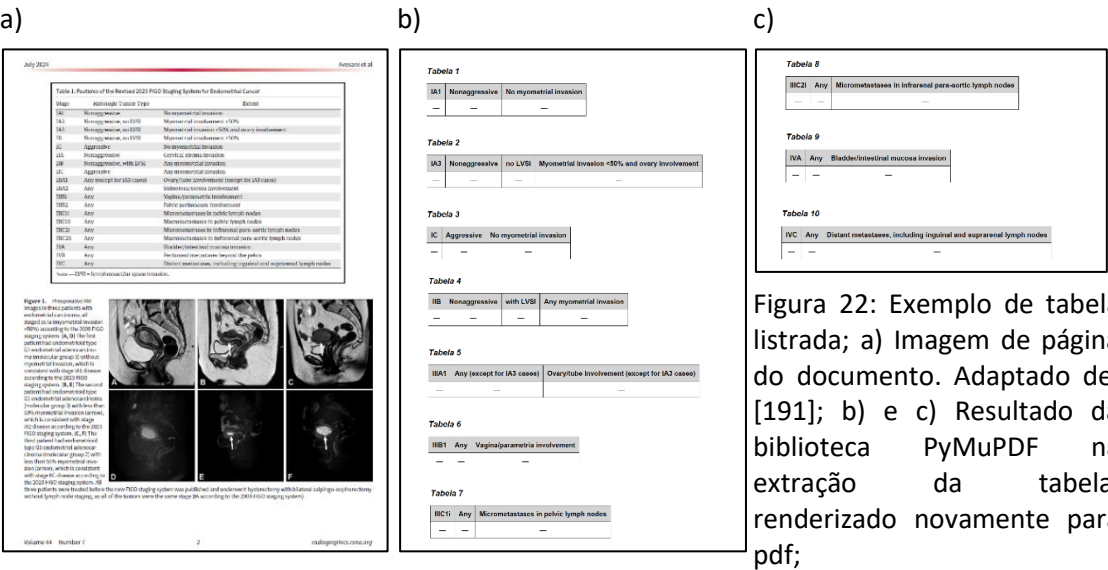
As ferramentas pdfPlumber e PyMuPDF enquadram-se nas ferramentas *rule-based*. Estas bibliotecas demonstraram um desempenho semelhante na identificação de tabelas com limites claramente definidos. Além disso, funcionaram corretamente para a extração de múltiplas tabelas presentes numa mesma página.

No entanto, em cenários que envolvem estruturas tabulares mais complexas, como na identificação de tabelas listradas, as bibliotecas apresentaram algumas limitações. Conforme ilustrado na Figura 22, a PyMuPDF interpretou cada linha colorida da tabela como uma tabela independente, segmentando a tabela em várias. Observou-se um comportamento idêntico na pdfPlumber. Este resultado sugere que estas bibliotecas apenas funcionam corretamente quando as tabelas apresentam um fundo de cor uniforme.

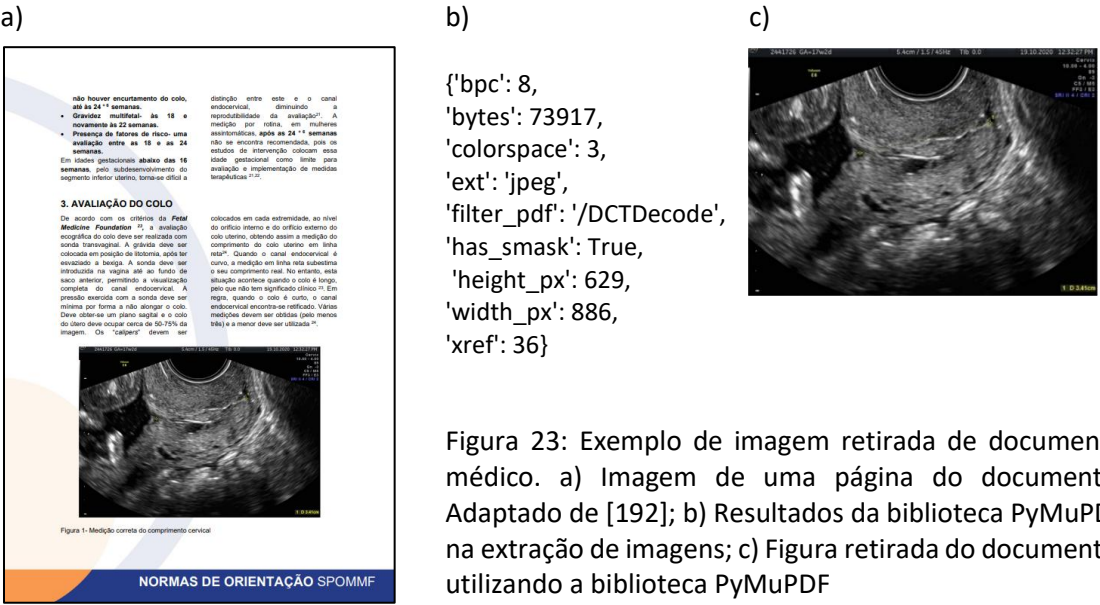
Estas bibliotecas não funcionaram quando foram aplicadas para extrair tabelas sem linhas verticais. Além disso, foram observados falsos positivos, uma vez que blocos de texto coloridos

delimitados por linhas são frequentemente interpretados como tabelas, mesmo quando não correspondem a estruturas tabulares propriamente ditas.

No que diz respeito à identificação automática de títulos e secções, a pdfPlumber não disponibiliza funcionalidades específicas para esta tarefa. Por outro lado, a PyMuPDF inclui um método que permite identificar a estrutura hierárquica do documento, devolvendo uma lista de entradas no formato nível, título, página, onde o nível indica a posição hierárquica do elemento (por exemplo, 1 corresponde ao título principal, 2 a uma secção, 3 a uma subsecção, e assim sucessivamente). No entanto, esta funcionalidade depende da presença de tags estruturais no PDF. Em muitos documentos científicos, estas marcações não estão disponíveis, o que impede a PyMuPDF de identificar corretamente títulos e secções.



A biblioteca PyMuPDF foi utilizada para a identificação e extração de imagens dos artigos, obtendo-se resultados como os da Figura 23.



Quando aplicada às normas clínicas da SPOMMF (Sociedade Portuguesa de Obstetrícia e Medicina Materno-Fetal), observou-se que a biblioteca identificou e extraiu corretamente as imagens presentes nos documentos, juntamente com os respetivos metadados (por exemplo, Figura 23). No entanto, foram também identificados falsos positivos, uma vez que a imagem de fundo do documento foi incorretamente classificada como uma figura do artigo. Este comportamento evidencia uma limitação da biblioteca na distinção entre elementos gráficos pertencentes ao conteúdo científico e elementos decorativos ou estruturais do documento.

- Table Transformer (TATR)

A biblioteca TATR é composta por dois modelos distintos: um modelo responsável pela deteção de tabelas nas páginas de documentos científicos e um segundo modelo dedicado ao reconhecimento da estrutura interna das tabelas, nomeadamente à identificação e delimitação das células que as compõem.

O modelo de deteção de tabelas apresentou um bom desempenho, demonstrando capacidade para localizar corretamente tabelas com diferentes estruturas visuais, incluindo tabelas listradas e na identificação de múltiplas tabelas numa mesma página do documento, como mostra a Figura 25 (b).

Por outro lado, o modelo de reconhecimento da estrutura interna das tabelas revelou algumas falhas. Observou-se que, mesmo em tabelas cuja estrutura apresenta as linhas bem definidas, o modelo apresentou dificuldades na identificação das mesmas, como exemplificado na Figura 25 (c).

Embora este modelo tenha demonstrado uma maior capacidade para lidar com tabelas de formatações mais diversificadas, o modelo de reconhecimento estrutural demonstrou várias falhas. Por exemplo, o modelo identificou corretamente as linhas da tabela listrada apresentada na Figura 25 (b), mas falhou na identificação das colunas desta mesma tabela. Para além disso, verificou-se casos em que tabelas com linhas claramente delimitadas não foram corretamente interpretadas, como na Figura 25 (c), onde as linhas não foram reconhecidas apesar de estarem visivelmente definidas.

224 Final FCG for Temperature Monitoring - 1 x 1 x 1 x 1 x 1

Table 1. FCG Classification (Modified from the FCG Guidelines, 1997)

FCG classification	Baseline heart frequency	Variability reactivity	Decelerations
Normal FCG	• 110-150 bpm	• ≥ 2 bpm • Accelerations	• Early decelerations • Uncomplicated variable decelerations with a duration of ≥ 90 seconds and loss of ≥ 60 bpm
Borderline FCG	• 100-110 bpm • 160-170 bpm • Short bradycardia episode	• < 2 bpm without acceleration • < 5 bpm for ≥ 40 sec	• Uncomplicated variable decelerations with a duration of ≥ 90 seconds and loss of ≥ 60 bpm
Abnormal FCG	• 160-170 bpm and reduced variability • < 10 bpm • Persistent bradycardia	• < 5 bpm for ≥ 40 sec • Sustained pattern	• Complicated variable decelerations with a duration of ≥ 90 seconds • Repeated late decelerations
Preterminal FCG	• Total lack of variability and reactivity with or without decelerations or bradycardia		

Spbm: beats per minute; FCG: cardiotocography; FCGO: International Federation of Gynecology and Obstetrics

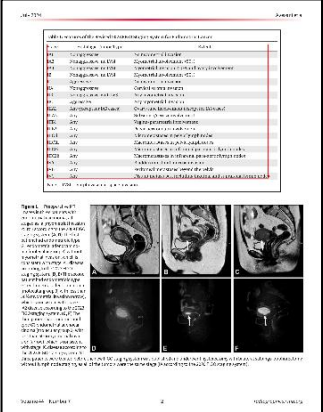
Table 2. STAN Guidelines for Clinical Intervention Based on FCG Annotations ST: Wouda-Chang2007

ST	CTG	Intermittent CTG	Abnormal CTG	Preterminal CTG
ST 1	• 10-15	• 10-15	• 10-15	• 10-15
ST 2	• 10-15	• 10-15	• 10-15	• 10-15
ST 3	• 10-15	• 10-15	• 10-15	• 10-15
ST 4	• 10-15	• 10-15	• 10-15	• 10-15
ST 5	• 10-15	• 10-15	• 10-15	• 10-15
ST 6	• 10-15	• 10-15	• 10-15	• 10-15
ST 7	• 10-15	• 10-15	• 10-15	• 10-15
ST 8	• 10-15	• 10-15	• 10-15	• 10-15
ST 9	• 10-15	• 10-15	• 10-15	• 10-15
ST 10	• 10-15	• 10-15	• 10-15	• 10-15
ST 11	• 10-15	• 10-15	• 10-15	• 10-15
ST 12	• 10-15	• 10-15	• 10-15	• 10-15
ST 13	• 10-15	• 10-15	• 10-15	• 10-15
ST 14	• 10-15	• 10-15	• 10-15	• 10-15
ST 15	• 10-15	• 10-15	• 10-15	• 10-15
ST 16	• 10-15	• 10-15	• 10-15	• 10-15
ST 17	• 10-15	• 10-15	• 10-15	• 10-15
ST 18	• 10-15	• 10-15	• 10-15	• 10-15
ST 19	• 10-15	• 10-15	• 10-15	• 10-15
ST 20	• 10-15	• 10-15	• 10-15	• 10-15
ST 21	• 10-15	• 10-15	• 10-15	• 10-15
ST 22	• 10-15	• 10-15	• 10-15	• 10-15
ST 23	• 10-15	• 10-15	• 10-15	• 10-15
ST 24	• 10-15	• 10-15	• 10-15	• 10-15
ST 25	• 10-15	• 10-15	• 10-15	• 10-15
ST 26	• 10-15	• 10-15	• 10-15	• 10-15
ST 27	• 10-15	• 10-15	• 10-15	• 10-15
ST 28	• 10-15	• 10-15	• 10-15	• 10-15
ST 29	• 10-15	• 10-15	• 10-15	• 10-15
ST 30	• 10-15	• 10-15	• 10-15	• 10-15
ST 31	• 10-15	• 10-15	• 10-15	• 10-15
ST 32	• 10-15	• 10-15	• 10-15	• 10-15
ST 33	• 10-15	• 10-15	• 10-15	• 10-15
ST 34	• 10-15	• 10-15	• 10-15	• 10-15
ST 35	• 10-15	• 10-15	• 10-15	• 10-15
ST 36	• 10-15	• 10-15	• 10-15	• 10-15
ST 37	• 10-15	• 10-15	• 10-15	• 10-15
ST 38	• 10-15	• 10-15	• 10-15	• 10-15
ST 39	• 10-15	• 10-15	• 10-15	• 10-15
ST 40	• 10-15	• 10-15	• 10-15	• 10-15
ST 41	• 10-15	• 10-15	• 10-15	• 10-15
ST 42	• 10-15	• 10-15	• 10-15	• 10-15
ST 43	• 10-15	• 10-15	• 10-15	• 10-15
ST 44	• 10-15	• 10-15	• 10-15	• 10-15
ST 45	• 10-15	• 10-15	• 10-15	• 10-15
ST 46	• 10-15	• 10-15	• 10-15	• 10-15
ST 47	• 10-15	• 10-15	• 10-15	• 10-15
ST 48	• 10-15	• 10-15	• 10-15	• 10-15
ST 49	• 10-15	• 10-15	• 10-15	• 10-15
ST 50	• 10-15	• 10-15	• 10-15	• 10-15
ST 51	• 10-15	• 10-15	• 10-15	• 10-15
ST 52	• 10-15	• 10-15	• 10-15	• 10-15
ST 53	• 10-15	• 10-15	• 10-15	• 10-15
ST 54	• 10-15	• 10-15	• 10-15	• 10-15
ST 55	• 10-15	• 10-15	• 10-15	• 10-15
ST 56	• 10-15	• 10-15	• 10-15	• 10-15
ST 57	• 10-15	• 10-15	• 10-15	• 10-15
ST 58	• 10-15	• 10-15	• 10-15	• 10-15
ST 59	• 10-15	• 10-15	• 10-15	• 10-15
ST 60	• 10-15	• 10-15	• 10-15	• 10-15
ST 61	• 10-15	• 10-15	• 10-15	• 10-15
ST 62	• 10-15	• 10-15	• 10-15	• 10-15
ST 63	• 10-15	• 10-15	• 10-15	• 10-15
ST 64	• 10-15	• 10-15	• 10-15	• 10-15
ST 65	• 10-15	• 10-15	• 10-15	• 10-15
ST 66	• 10-15	• 10-15	• 10-15	• 10-15
ST 67	• 10-15	• 10-15	• 10-15	• 10-15
ST 68	• 10-15	• 10-15	• 10-15	• 10-15
ST 69	• 10-15	• 10-15	• 10-15	• 10-15
ST 70	• 10-15	• 10-15	• 10-15	• 10-15
ST 71	• 10-15	• 10-15	• 10-15	• 10-15
ST 72	• 10-15	• 10-15	• 10-15	• 10-15
ST 73	• 10-15	• 10-15	• 10-15	• 10-15
ST 74	• 10-15	• 10-15	• 10-15	• 10-15
ST 75	• 10-15	• 10-15	• 10-15	• 10-15
ST 76	• 10-15	• 10-15	• 10-15	• 10-15
ST 77	• 10-15	• 10-15	• 10-15	• 10-15
ST 78	• 10-15	• 10-15	• 10-15	• 10-15
ST 79	• 10-15	• 10-15	• 10-15	• 10-15
ST 80	• 10-15	• 10-15	• 10-15	• 10-15
ST 81	• 10-15	• 10-15	• 10-15	• 10-15
ST 82	• 10-15	• 10-15	• 10-15	• 10-15
ST 83	• 10-15	• 10-15	• 10-15	• 10-15
ST 84	• 10-15	• 10-15	• 10-15	• 10-15
ST 85	• 10-15	• 10-15	• 10-15	• 10-15
ST 86	• 10-15	• 10-15	• 10-15	• 10-15
ST 87	• 10-15	• 10-15	• 10-15	• 10-15
ST 88	• 10-15	• 10-15	• 10-15	• 10-15
ST 89	• 10-15	• 10-15	• 10-15	• 10-15
ST 90	• 10-15	• 10-15	• 10-15	• 10-15
ST 91	• 10-15	• 10-15	• 10-15	• 10-15
ST 92	• 10-15	• 10-15	• 10-15	• 10-15
ST 93	• 10-15	• 10-15	• 10-15	• 10-15
ST 94	• 10-15	• 10-15	• 10-15	• 10-15
ST 95	• 10-15	• 10-15	• 10-15	• 10-15
ST 96	• 10-15	• 10-15	• 10-15	• 10-15
ST 97	• 10-15	• 10-15	• 10-15	• 10-15
ST 98	• 10-15	• 10-15	• 10-15	• 10-15
ST 99	• 10-15	• 10-15	• 10-15	• 10-15
ST 100	• 10-15	• 10-15	• 10-15	• 10-15

CTG: cardiotocography

Figura 24: Exemplo de imagem retirada de documento médico. Adaptado de: [191]; a) Resultado do modelo de deteção de tabelas do TATR; b) Resultado do modelo de identificação de estrutura interna da tabela do TATR.

a)



b)

Stage	Histologic Tumor Type	Extent
IA1	Nonaggressive	No myometrial invasion
IA2	Nonaggressive, no LVSI	Myometrial involvement <50%
IA3	Nonaggressive, no LVSI	Myometrial invasion <50% and ovary involvement
IB	Nonaggressive, no LVSI	Myometrial involvement >50%
IC	Aggressive	No myometrial invasion
IIA	Nonaggressive	Cervical stroma invasion
IIB	Nonaggressive, with LVSI	Any myometrial invasion
IIC	Aggressive	Any myometrial invasion
IIIA1	Any (except for IA3 cases)	Ovary/tube involvement (except for IA3 cases)
IIIA2	Any	Subserosa/serosa involvement
IIIB1	Any	Vagina/parametria involvement
IIIB2	Any	Pelvic peritoneum involvement
IIIC1	Any	Micrometastases in pelvic lymph nodes
IIIC1i	Any	Macrometastases in pelvic lymph nodes
IIIC2i	Any	Micrometastases in infrarenal para-aortic lymph nodes
IIIC2ii	Any	Macrometastases in infrarenal para-aortic lymph nodes
IVA	Any	Bladder/intestinal mucosa invasion
IVB	Any	Peritoneal metastases beyond the pelvis
IVC	Any	Distant metastases, including inguinal and suprarenal lymph nodes

c)

CTG	Intermediary CTG	Abnormal CTG	Preterminal CTG
Episodic T/QRS rise Duration of change <10 minutes)	• >0.15	• >0.10	• Immediate delivery
Baseline T/QRS rise Duration of change >10 minutes)	• >0.10	• >0.05	
Biphasic ST	• Continuous >5 min or 3 episodes	• Continuous >2 min or 2 episodes	

Figura 25: Exemplo de tabela listrada; a) Imagem de página do documento. Adaptado de: [193]; b) Resultado do modelo de detecção de tabelas do TATR, com as tabelas identificadas a vermelho; c) Resultado do modelo de reconhecimento da estrutura interna de tabelas do TATR, com a estrutura identificada a verde

- PaddleOCR

O PaddleOCR é das bibliotecas mais completas dentre as analisadas neste estudo, por apresentar a capacidade para detetar até 23 tipos distintos de elementos narrativos dos artigos científicos. Entre estes incluem-se o nome do documento, os títulos das seções, as tabelas, as imagens, as legendas, os cabeçalhos, os rodapés e os parágrafos. É a única biblioteca que disponibiliza um módulo especificamente adaptado para o processamento de artigos científicos em língua portuguesa.

Relativamente à extração de tabelas, o PaddleOCR apresentou um bom desempenho. A biblioteca identificou corretamente múltiplas tabelas presentes numa mesma página, bem como detetar tabelas com colunas ou linhas unidas. Para além da deteção, apresentou também bons resultados no reconhecimento da estrutura interna das tabelas, conseguindo identificar corretamente as células, mesmo em casos onde não existiam linhas verticais e horizontais explícitas, como exemplificado Figura 26 (b).

Para além da análise tabular, o PaddleOCR demonstrou capacidade para identificar outros elementos estruturais do documento, nomeadamente cabeçalho, rodapé e número da página. Esta funcionalidade revela-se particularmente útil, uma vez que permite evitar que essa informação seja indevidamente incorporada nos parágrafos do artigo, durante o processo de extração dos parágrafos do artigo.

No que diz respeito à extração de títulos, o modelo demonstrou um bom desempenho. O PaddleOCR identificou o título principal do artigo, rotulando-o como *doc_title*, bem como os títulos das seções, classificados como *paragraph_title*, e os parágrafos do texto, assinalados com a etiqueta *texto*, conforme aparece na Figura 26(a).

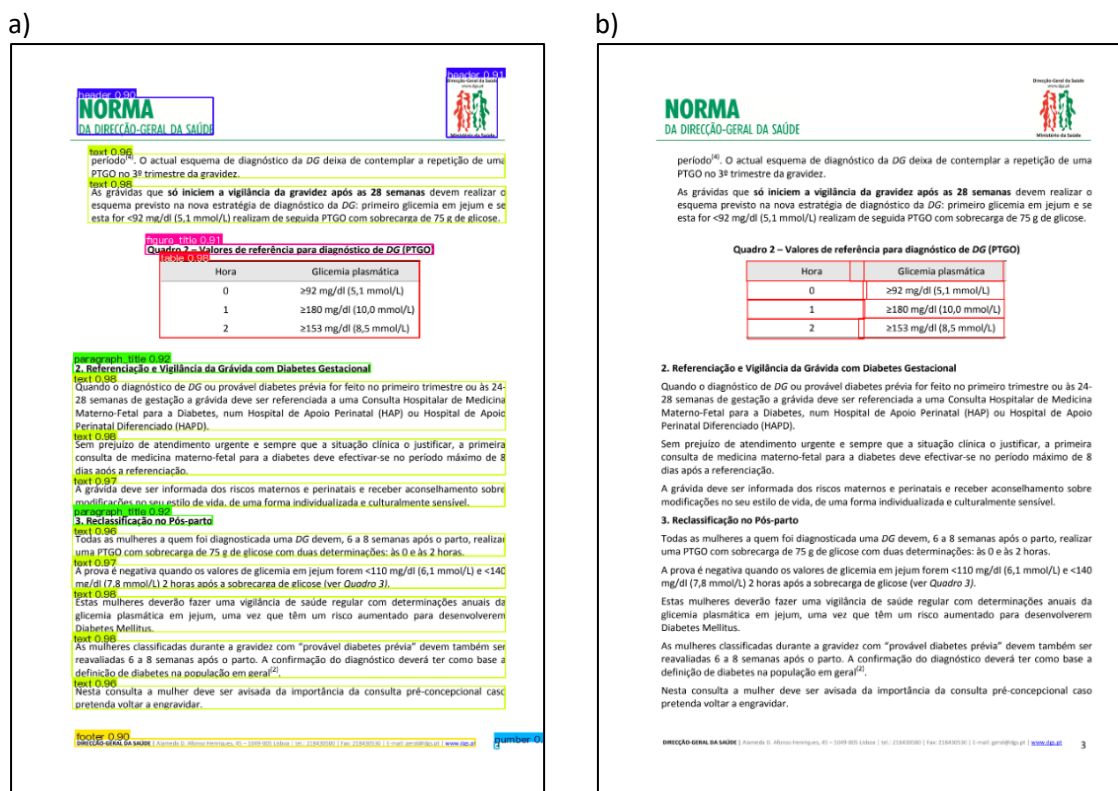


Figura 26: Exemplo de imagem de documento com tabela. Adaptado de [194] a) Resultados da biblioteca PaddleOCR na identificação de elementos narrativos; b) Resultados da biblioteca PaddleOCR para o reconhecimento da estrutura interna da tabela;



Figura 27: Exemplo de imagem de documento com duas colunas. Retirado de [195] e [196]; a) Resultado da biblioteca PaddleOCR na identificação de elementos narrativos. Retirado de [195]; b) Resultado da biblioteca PaddleOCR na identificação de elementos narrativos.

Todavia, o modelo apresentou algumas falhas. Por vezes, não detetou corretamente o título principal do documento, como exemplificado em Figura 27 (b), onde o título “Parto Pélvico Vaginal” não foi reconhecido como *doc_title*. Além disso, o modelo não distinguiu os diferentes níveis hierárquicos dos títulos (por exemplo, h1, h2, h3).

Verificou-se que o modelo cometeu erros na ordenação dos segmentos de texto, não reproduzindo a sequência natural de leitura. Na Figura 28 (a), por exemplo, os títulos “Fisiologia da tiroide durante a gravidez” e “Tiroide materna” surgiram em 4º e 5º lugares, quando, pela ordem de leitura, deveriam ocupar o 1º e 2º lugares.

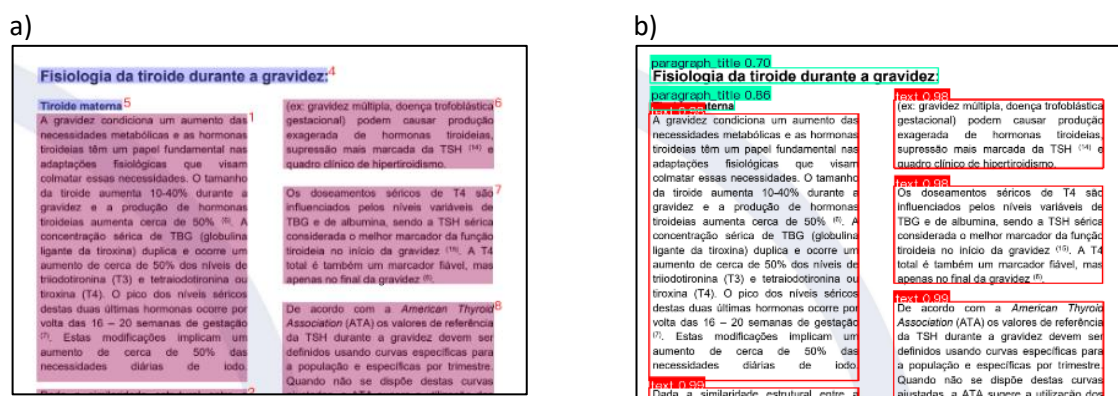


Figura 28: Exemplo de imagem de documento com duas colunas. Retirado de [197]; a) Resultado do modelo PaddleOCR na numeração dos elementos narrativos identificados; b) Resultado do modelo PaddleOCR na identificação de elementos narrativos

- Nougat

O Nougat é uma ferramenta capaz de identificar diversos elementos estruturais em documentos científicos, nomeadamente tabelas, títulos e equações matemáticas.

Verificou-se que o modelo consegue identificar corretamente tabelas mesmo em cenários mais complexos, como tabelas sem linhas verticais explícitas e com colunas unidas. Nestes casos, o Nougat detetou a presença da tabela e reconheceu de forma adequada a sua estrutura interna. Este comportamento é ilustrado na Figura 29 (c), onde se observa uma reconstrução coerente com a tabela original da Figura 29 (a).

Esta modelo identificou automaticamente os títulos presentes no texto do documento, mesmo quando estes apresentavam estilos distintos, como os ilustrados na Figura 30 (a). No entanto, o modelo não distinguiu os diferentes níveis hierárquicos dos títulos.

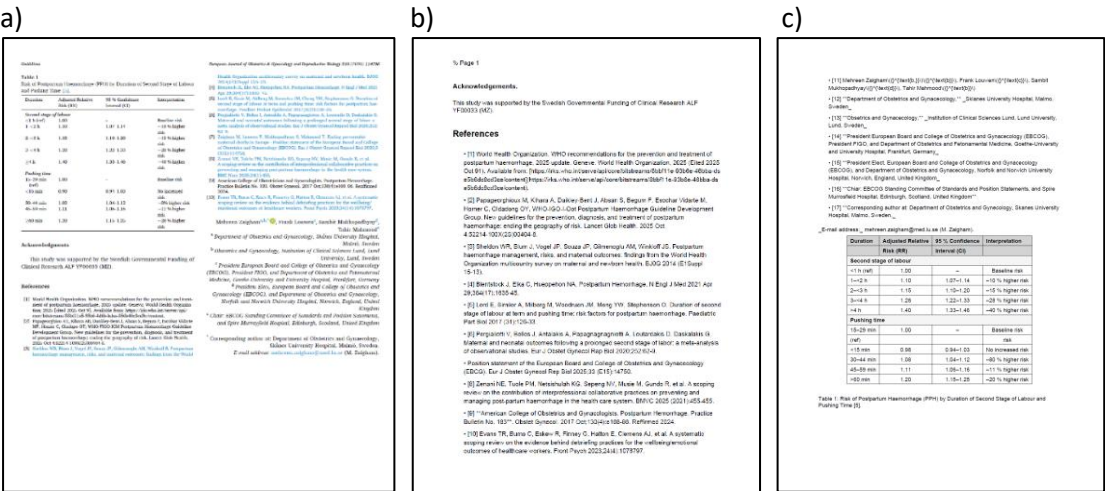


Figura 29: Exemplo de uma imagem com tabela e dividido em duas colunas. a) Imagem de uma página do documento. Adaptado de [198]; b) e c)Resultado do modelo Nougat, convertido para html e renderizado novamente para pdf

Nougat demonstrou capacidade para reconhecer listas no texto. Durante o processo de extração, o modelo representou corretamente listas estruturadas, tanto no caso das referências, que se iniciam por numeração, como nas listas apresentadas no artigo da Figura 30, que utilizam marcadores sob a forma de pontos. O Nougat apresenta ainda a vantagem de conseguir identificar cabeçalhos e rodapés ao longo do documento, impedindo que estes sejam incluídos no texto principal extraído.

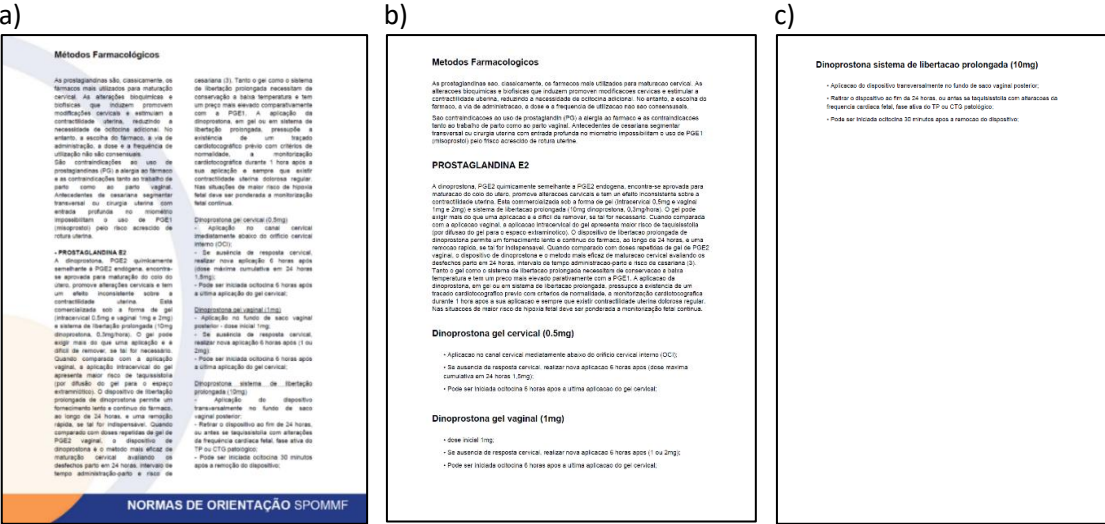


Figura 30: Exemplo de uma imagem com vários títulos de diferentes formatos. a) Imagem de uma página do documento. Adaptado de: [199]; b) e c)Resultado do modelo Nougat, convertido para html e renderizado novamente para pdf;

6.1.2 Discussão de Resultados

Nesta secção compararam-se as bibliotecas de acordo com o seu desempenho em tarefas relacionadas com a estruturação de tabelas, a segmentação de texto e a identificação de imagens. Para facilitar a comparação entre as diferentes ferramentas, cada uma das bibliotecas foram avaliadas para cada tarefa com: ✓, quando a biblioteca executa a tarefa corretamente; ±, quando a tarefa é executada, mas com restrições ou limitações observáveis; e X, quando a biblioteca não dispõe de funcionalidade ou capacidade para realizar a tarefa em questão;

- Tabelas

No que diz respeito à estruturação de tabelas, a avaliação incidiu sobre vários critérios relevantes para a correta interpretação da informação tabular. Em particular, analisou-se a capacidade das bibliotecas para:

- a) Detetar tabelas sem grelhas explícitas, isto é, tabelas sem linhas verticais e/ou horizontais;
- b) Identificar tabelas com *layouts* mais complexos, incluindo tabelas com células unidas por colunas e/ou linhas; e
- c) Reconhecer adequadamente as legendas associadas às tabelas.

As bibliotecas também foram avaliadas quanto à ocorrência de falsos positivos, ou seja, situações em que a biblioteca sinaliza incorretamente a presença de uma tabela onde esta não existe, como por exemplo, quando elementos gráficos simples, tais como retângulos ou caixas de texto, são confundidos com tabelas.

A Tabela 11 apresenta uma comparação do desempenho das diferentes bibliotecas na tarefa de estruturação de tabelas em documentos biomédicos.

Tabela 11: Avaliação das Bibliotecas para estruturação de Tabelas

Bibliotecas	Reconhecer tabelas sem grelhas	Tabelas com Layout Complexo (células unidas)	Falsos Positivos (resumos, blocos de texto)	Detetar legendas
pdfPlumber	X	±	X	X
PyMuPDF	X	±	X	X
Table Transformer (TATR)	✓	✓	±	X
PaddleOCR	✓	✓	✓	✓
Nougat	✓	✓	✓	X

Nesta tarefa, a biblioteca PaddleOCR destacou-se como a solução mais completa, pois apresenta um bom desempenho em todos os critérios avaliados. A biblioteca foi capaz de reconhecer as tabelas sem grelhas, lidar com *layouts* complexos e identificar corretamente as legendas associadas, sem registar falsos positivos. Contudo, existe uma limitação em todas as bibliotecas: a incapacidade de detetar e associar de forma as notas em rodapé relacionadas com a tabela, que por vezes apresentam uma explicação de siglas ou palavras que aparecem na tabela.

- Extração de Títulos e Secções de texto

As bibliotecas foram avaliadas quanto à extração de títulos e secções de texto em documentos biomédicos. Em particular, analisou-se: a capacidade de detetar e distinguir títulos de parágrafos; a extração hierárquica de títulos (p. ex., níveis 1, 2 e 3), assegurando a correta identificação das diferentes secções; a extração do texto segundo a ordem de leitura, um ponto em que algumas ferramentas apresentam limitações, sobretudo em artigos com duas ou mais colunas; e a identificação de cabeçalhos e rodapés.

A Tabela 12 apresenta a comparação das bibliotecas no que respeita à estruturação e segmentação do texto, considerando os pontos referidos.

Tabela 12: Avaliação das Bibliotecas para extração de títulos e secções de texto

Biblioteca	Ordenar corretamente as secções de texto (consoante a leitura)	Detetar títulos	Detetar títulos por hierarquias (H1, H2, H3, ...)	Deteta Cabeçalhos e Rodapés
pdfPlumber	±	X	X	X
PyMuPDF	±	±	±	X
PaddleOCR	±	✓	±	✓
Nougat	±	✓	X	✓

De forma geral, nenhuma das bibliotecas avaliadas cumpriu todos os critérios analisados. As bibliotecas baseadas em *deep-learning*, nomeadamente PaddleOCR e Nougat, demonstraram capacidade para identificar diversos elementos dos artigos científicos, incluindo cabeçalhos, rodapés, legendas e títulos, apresentando o texto extraído organizado por parágrafos. Devido a isso, estas bibliotecas evitam a mistura de conteúdo dos parágrafos do texto com outros elementos do documento. No entanto, ambas apresentam algumas falhas quanto à ordenação dos parágrafos, especialmente quando a formatação do artigo é mais complexa. O PaddleOCR, como forma de diminuir esta falha, recorre a um modelo específico para ordenar os blocos de textos encontrados numa página do artigo. Apesar de apresentar algumas falhas, como a relata na Figura 28 (a), esta continua a ser a biblioteca com melhores resultados neste critério.

- Imagens

Nesta secção, avaliou-se a capacidade de extração de imagens a partir de documentos biomédicos. Em particular, analisou-se o desempenho das bibliotecas em detetar corretamente as tabelas e a extrair as respetivas legendas. As bibliotecas também foram analisadas quanto à ocorrência de falsos positivos. A Tabela 13 apresenta a comparação do desempenho das bibliotecas na tarefa de extração de imagens em documentos biomédicos.

Tabela 13: Avaliação das bibliotecas para extração de imagens

Biblioteca	Detetar as imagens	Detetar as legendas das imagens	Falsos Positivos
PyMuPDF	✓	X	X
PaddleOCR	✓	✓	✓

De acordo com os resultados obtidos, o PaddleOCR revelou o melhor desempenho na extração de imagens. Embora o PyMuPDF consiga identificar as imagens presentes em artigos científicos, também apresenta vários falsos positivos, por exemplo, considera marcas de água e logotipos como imagens relevantes do documento. Para além disso, o PaddleOCR consegue identificar as legendas associadas às figuras.

Assim, apesar de as bibliotecas pdfPlumber e PyMuPDF apresentarem um baixo custo computacional e tempos de execução reduzidos, estas bibliotecas apresentam várias limitações na deteção de tabelas. Para além disso, em relação à extração de tabelas e imagens, ocorreu um número elevado de falsos positivos. Por outro lado, as abordagens baseadas em DL mostram, no geral, um melhor desempenho. Estes métodos mostraram maior capacidade de generalização e melhor adaptação a documentos com diferentes formatações.

Com base na análise apresentada, pode-se concluir que, embora todas as bibliotecas avaliadas apresentem vantagens e limitações específicas, a PaddleOCR demonstrou melhor desempenho na extração de texto, tabelas e imagens dos documentos biomédicos. Esta biblioteca destacou-se na identificação de tabelas complexas, múltiplas tabelas e legendas, bem como na extração de imagens sem ocorrerem falsos positivos.

Por fim, como forma de melhorar os resultados e compensar as falhas observadas com o PaddleOCR, serão feitos estudos para compensar as falhas utilizando LLMs. Estes modelos poderão analisar os dados previamente estruturados, validar a informação extraída e corrigir eventuais inconsistências, contribuindo para uma melhor qualidade dos dados guardados.

6.2 Avaliação de estratégias de descrição de figuras

Nesta secção procedeu-se à avaliação das duas abordagens anteriormente apresentadas na Secção 5.5. O objetivo consiste em comparar a capacidade de cada abordagem para gerar descrições textuais representativas do conteúdo de diferentes tipos de figuras presentes em documentos biomédicos.

A primeira estratégia baseia-se na informação textual disponível no próprio documento, utilizando a legenda da figura e os parágrafos que a referenciam para gerar a descrição. A segunda estratégia recorre a um modelo multimodal, responsável por analisar diretamente o conteúdo visual da figura e produzir uma descrição em linguagem natural.

Para esta avaliação foram selecionadas figuras representativas dos diferentes tipos de conteúdo visual encontrados em artigos científicos e normas clínicas, nomeadamente fluxogramas, ilustrações clínicas e ecografias. Para cada figura foram geradas descrições recorrendo às duas estratégias propostas, sendo posteriormente comparadas quanto à sua capacidade para representar corretamente o conteúdo visual.

As Figuras 31 a 33 apresentam alguns exemplos das figuras utilizadas nesta avaliação. A Figura 31 corresponde a um fluxograma clínico. Estas figuras são caracterizadas por conterem uma

maior quantidade de texto integrado na própria imagem, bem como por apresentarem relações sequenciais entre etapas e elementos direcionais que orientam a leitura. Estes componentes, como caixas de decisão, setas e ramificações, descrevem processos estruturados, tornando a interpretação dependente tanto do conteúdo textual como da organização espacial dos elementos. A Figura 32 apresenta duas ilustrações de procedimentos obstétricos.

A Figura 32 apresenta duas ilustrações de procedimentos obstétricos. Estas figuras têm como principal objetivo representar visualmente manobras clínicas, procedimentos médicos ou estruturas anatómicas, funcionando como complemento à informação descrita no texto do documento. Por este motivo, a sua interpretação depende frequentemente da legenda associada e dos parágrafos que fazem referência à figura, onde os autores fornecem o contexto clínico necessário para compreender corretamente o procedimento ilustrado.

A Figura 33 apresenta um exemplo de uma ecografia. Ao contrário das ilustrações clínicas, estas figuras contêm normalmente pouca informação textual associada e, em muitos casos, são acompanhadas apenas pela legenda, sem referências adicionais ao longo do documento. Consequentemente, a descrição do seu conteúdo depende sobretudo da interpretação dos elementos visuais presentes na imagem, tornando este tipo de figura particularmente desafiante para abordagens baseadas exclusivamente na informação textual.

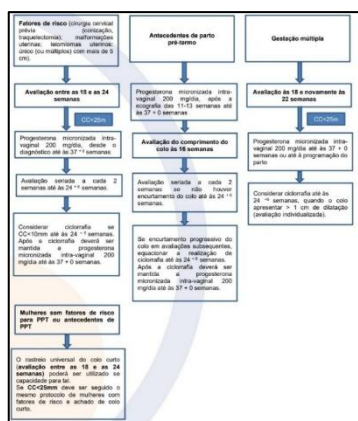


Figura 31: Exemplo de uma figura de um fluxograma. Retirado do documento: [192]

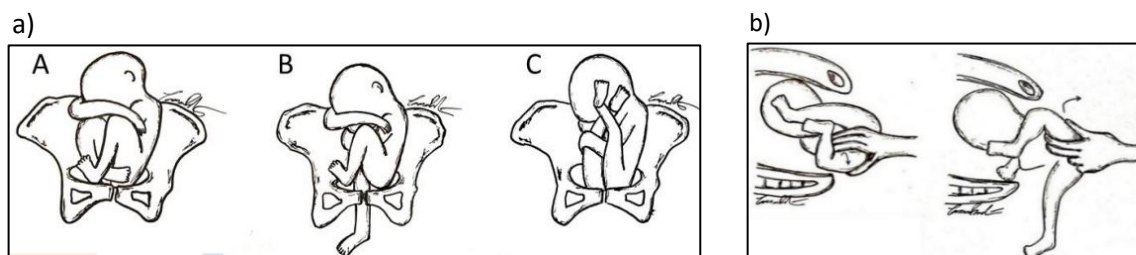


Figura 32: Exemplos de imagens usadas para testar os modelos com figuras clínicas, retirada de [196].



Figura 33: Exemplo de imagem, com legenda, mas sem referências no texto. Retirado de: [192]

6.2.1 Resultados:

Para o fluxograma apresentado na Figura 31, foram geradas descrições utilizando ambas as estratégias. As descrições obtidas encontram-se apresentadas na Tabela B1, no Anexo B. Observou-se que a abordagem baseada no modelo multimodal produziu uma descrição mais detalhada da estrutura do fluxograma e do texto presente na imagem, enquanto a abordagem baseada na informação textual do documento originou uma descrição mais breve e pouco descritiva.

Verificou-se que estes raramente eram referidos de forma explícita no texto e que as respectivas legendas tendiam a ser pouco descritivas. Por exemplo, a legenda do fluxograma utilizado para a Figura 31 foi “Quadro final com o resumo das orientações: Rastreio do colo do útero”. Como consequência, a aplicação da Estratégia 1, baseada exclusivamente na legenda, resultou numa descrição pouco informativa.

Por outro lado, a utilização do modelo multimodal em fluxogramas revelou um desempenho superior neste tipo de figura, uma vez que permitiu a análise direta do conteúdo visual, incluindo o texto presente na imagem e a estrutura do diagrama. Esta capacidade traduziu-se na geração de descrições mais detalhadas e sequenciais, que refletiram de forma mais fiel a organização e o fluxo de informação representados no fluxograma.

Para as ilustrações clínicas da Figura 32, os resultados encontram-se apresentados na Tabela B2, no Anexo B. Neste caso, a estratégia baseada na informação textual produziu descrições mais específicas dos procedimentos representados, enquanto a abordagem multimodal gerou descrições mais genéricas e, em alguns casos, identificações incorretas do procedimento clínico.

Numa primeira tentativa, o modelo apresentou comportamentos de alucinação, atribuindo interpretações completamente erradas às imagens. Por exemplo, a imagem da Figura 32 a), que representa uma criança, foi inicialmente descrita pelo modelo como um elefante.

Posteriormente, o *prompt* foi ajustado para indicar explicitamente que todas as imagens pertenciam ao domínio da obstetrícia. Após esta alteração, o modelo passou a produzir descrições mais coerentes com o contexto clínico, identificando corretamente a imagem como um parto. No entanto, mesmo neste caso, o modelo continuou a apresentar algumas imprecisões, ao inventar o nome do procedimento médico representado na imagem.

Em ambas as figuras, a estratégia 1 consegue descrever mais corretamente a figura e sem erros clínicos. Por exemplo, na Figura 32 (b), na estratégia 1, a imagem é descrita como a manobra de Pinard, enquanto a estratégia 2 deu uma descrição mais genérica que poderia descrever vários procedimentos obstétricos diferentes. Para a Figura 32 (a), a estratégia 1 consegue descrever pormenores importantes sobre a figura, como por exemplo, que se refere a um Parto Pélvico (“*breech presentations*”), e relaciona corretamente as três figuras com variações deste tipo de parto.

Para a ecografia apresentada na Figura 33, cujos resultados se encontram na Tabela B3, no Anexo B, verificou-se que a estratégia baseada na informação textual produziu uma descrição limitada, devido à reduzida informação disponível no documento. Em relação à utilização da estratégia com agente multimodal, apresentou melhores resultados, referindo que a imagem é uma ecografia. No entanto, esta abordagem pode conter erros técnicos e pode interpretar incorretamente os valores que aparecem na imagem, devido aos modelos originarem alucinações. Nestes casos, nenhuma das duas estratégias se mostrou plenamente eficaz, sendo mais apropriado recorrer unicamente à legenda para descrever o conteúdo da imagem.

Deste modo, conclui-se que a abordagem mais adequada para o sistema consiste em aplicar a Estratégia 1 (contexto textual) na geração de descrições de imagens clínicas, como ilustrações de manobras ou procedimentos médicos, enquanto a Estratégia 2, baseada no modelo multimodal, se revele mais eficaz para a descrição de fluxogramas presentes nos documentos.

Adicionalmente, nos casos em que se trata de imagens sem referências textuais associadas, a estratégia mais apropriada consiste em utilizar exclusivamente a legenda disponível para gerar a descrição da figura.

6.3 Avaliação da classificação automática de figuras

Numa fase inicial, foi avaliada a capacidade de modelos multimodais realizarem diretamente a classificação das figuras com base exclusivamente no conteúdo visual. O objetivo desta experiência consistiu em analisar se estes modelos eram capazes de distinguir automaticamente entre fluxogramas e imagens ilustrativas, permitindo determinar a viabilidade de uma abordagem baseada apenas na interpretação visual para seleção da estratégia de processamento adequada.

Para esta avaliação foram considerados três modelos multimodais: LLaVA 7B, executado localmente, e os modelos LLaMA 4 Maverick 17B 128E e LLaMA 4 Scout 17B 16E, disponibilizados através da plataforma GROQ. A seleção destes modelos teve como objetivo comparar diferentes capacidades de interpretação visual, considerando tanto um modelo com menores requisitos computacionais como modelos de maior dimensão.

Para realizar a classificação direta das figuras, foi definido um *prompt* que instrui os modelos a atribuir uma das duas categorias definidas, fluxograma ou imagem ilustrativa, com base exclusivamente nas características visuais da figura. Adicionalmente, foi solicitada uma breve

particularmente no caso de imagens de ecografia e fluxogramas com uma estrutura visual claramente definida.

O modelo LLaVA 4 Maverick 17B 128E foi capaz de identificar corretamente o tipo de conteúdo presente nas figuras, descrevendo com maior precisão diagramas, fluxogramas clínicos e imagens de ecografia. De forma semelhante, o modelo LLaVA 4 Scout 17B 16E apresentou resultados consistentes, gerando descrições adequadas para a maioria das imagens analisadas.

Por outro lado, o modelo LLaVA 7B apresentou algumas limitações na interpretação do conteúdo visual. Em vários casos, gerou descrições genéricas que não refletiam corretamente o conteúdo das imagens. Por exemplo, na Figura 34 (d) o modelo interpretou incorretamente a imagem como um diagrama humorístico relacionado com o ouvido humano, quando na realidade se trata de uma ilustração médica associada a um procedimento obstétrico. De forma semelhante, na Figura 34 (e) o modelo classificou incorretamente a imagem como representando um elefante, evidenciando um caso de alucinação do modelo.

O LLaMA 4 Maverick 17B 128E mostrou desempenho intermédio: apesar de manter algumas classificações incorretas, sobretudo em imagens clínicas confundidas com diagramas, produziu descrições mais coerentes e identificou corretamente as ecografias. O melhor desempenho foi obtido pelo LLaMA 4 Scout 17B 16E, que classificou a maioria das imagens corretamente. Embora no caso da Figura 34 (a), cuja classificação correta seria Imagem, o modelo atribuiu incorretamente a categoria Diagrama. Essa classificação parece ter sido influenciada pela presença de setas na ilustração, o que levou o modelo a interpretá-la como uma representação esquemática em vez de uma imagem.

6.4 Avaliação dos modelos RAG

Para avaliar o desempenho dos modelos de *embeddings* no contexto de RAG foi desenvolvido um dataset específico do domínio da obstetrícia. O dataset é composto por perguntas clínicas e pelos respetivos excertos de documentos considerados como *ground truth*, permitindo avaliar a capacidade dos modelos em recuperar informação clinicamente relevante a partir de um corpus especializado.

Os temas da diabetes gestacional e da pré-eclâmpsia foram incluídos por apresentarem fatores de risco e manifestações clínicas parcialmente sobrepostos [200], constituindo um cenário adequado para avaliar a capacidade dos modelos de *embeddings* em distinguir conceitos semanticamente semelhantes e recuperar informação específica da condição clínica em questão. Por sua vez, a versão cefálica externa foi selecionada por ser um procedimento frequentemente documentado através de figuras ilustrativas e esquemas clínicos, permitindo avaliar a capacidade do sistema na recuperação de informação visual a partir de descrições textuais.

O corpus documental utilizado é constituído por normas clínicas, recomendações internacionais e artigos científicos recentes provenientes de entidades nacionais e internacionais de referência,

bem como de literatura científica indexada na base de dados PubMed. No total, o corpus é composto por 41 documentos clínicos.

Após a seleção dos documentos, estes foram convertidos para formato textual estruturado, segmentados em unidades de informação (*chunks*) e armazenados numa base de dados vetorial utilizando o Chroma, constituindo a base documental utilizada na avaliação dos diferentes modelos de *embeddings*.

- Tópicos clínicos

O *dataset* foi organizado em cinco tópicos clínicos, selecionados com o objetivo de representar diferentes áreas da prática clínica em obstetrícia e diferentes tipos de evidência científica.

Diabetes Gestacional, uma das complicações metabólicas mais frequentes durante a gravidez, afetando aproximadamente 16,5 % das gravidezes em todo o mundo. Esta condição encontra-se associada a diversos riscos maternos, como pré-eclâmpsia, hipertensão gestacional, maior risco de parto por cesariana e aumento do risco de desenvolvimento de diabetes *mellitus* tipo 2 no futuro. Devido ao seu impacto significativo na saúde materna e neonatal, a diabetes gestacional encontra-se amplamente descrita em normas clínicas e estudos científicos [201]

Pré-Eclâmpsia é uma perturbação hipertensiva associada à gravidez caracterizada pelo desenvolvimento de hipertensão arterial e sinais de disfunção orgânica após as 20 semanas de gestação. Esta condição está associada a um aumento significativo da morbilidade e mortalidade materna e fetal, bem como a complicações como restrição do crescimento fetal e parto prematuro. Dada a sua relevância clínica, a sua prevenção, diagnóstico e abordagem terapêutica encontram-se amplamente descritos em normas clínicas e recomendações baseadas na evidência [202].

Asma é uma doença respiratória crónica caracterizada por inflamação das vias aéreas, hipersensibilidade brônquica e episódios recorrentes de dispneia e tosse. É considerada uma das doenças crónicas mais frequentes durante a gestação, afetando aproximadamente 4 a 8% das grávidas. Um controlo inadequado da asma durante a gravidez pode estar associado a um aumento do risco de complicações maternas e fetais, incluindo pré-eclâmpsia, parto prematuro, baixo peso à nascença e restrição do crescimento fetal [203], [204].

Anemia e suplementação na gravidez, uma condição frequente na gravidez, causada por défice de ferro e associada ao aumento das necessidades nutricionais maternas durante a gravidez. Esta condição pode ter repercussões importantes na saúde materna e fetal, incluindo fadiga, maior risco de parto prematuro, baixo peso à nascença e complicações perinatais. A prevenção e o tratamento da anemia através de estratégias de suplementação nutricional, como ferro e ácido fólico, são essenciais para os cuidados pré-natais [205].

Versão Cefálica Externa, um procedimento obstétrico utilizado para rodar o feto da apresentação pélvica para a apresentação cefálica. Este procedimento é realizado nas últimas semanas de gestação e tem como objetivo reduzir a incidência de partos por cesariano devido

à apresentação pélvica. A segurança deste procedimento é avaliado em vários estudos clínicos e encontra-se descrito em recomendações e orientações clínicas [206].

- Conjunto de documentos

O conjunto de documentos utilizado neste estudo é composto por normas clínicas, recomendações e literatura científica relacionada com a área da obstetrícia.

Foram incluídas fontes provenientes de entidades internacionais e nacionais de referência, nomeadamente a Sociedade Portuguesa de Obstetrícia e Medicina Materno-Fetal (SPOMMF), uma associação científica portuguesa que disponibiliza diretrizes clínicas para a melhoria da saúde materna e neonatal [207]. A Direção-Geral da Saúde (DGS), um organismo central do Ministério da Saúde, que ajuda na implementação de políticas de saúde em Portugal. Nesse sentido, a DGS contém normas clínicas relativas às terapias, tratamentos e medicamentos do Serviço Nacional de Saúde [208]. A nível internacional, existe a *National Institute for Health and Care Excellence* (NICE), uma organização independente do Reino Unido responsável pelo desenvolvimento de orientações clínicas baseadas em evidência, com o objetivo de promover a saúde, prevenir doenças e melhorar o diagnóstico e tratamento de diversas condições, e que atual em várias áreas da saúde, incluindo a obstetrícia [209], [210]. Outra organização com normas clínicas que apoiam a área da obstetrícia é a FIGO (*International Federation of Gynecology and Obstetrics*), é uma organização internacional que reúne sociedades de ginecologia e obstetrícia de vários países desde 1954 [211], [212].

De forma a complementar as recomendações clínicas com evidência científica recente, foram incluídos artigos indexados na base de dados PubMed, incluindo revisões sistemáticas e estudos clínicos relevantes. Esta base de dados é amplamente utilizada na área da saúde para acesso a literatura científica atualizada e de elevada qualidade metodológica [213].

Na prática clínica em obstetrícia e ginecologia, a tomada de decisões baseia-se cada vez mais na análise crítica da evidência científica disponível. Entre os diferentes tipos de investigação, destacam-se as revisões sistemáticas e meta-análises, que permitem sintetizar os resultados de múltiplos estudos, bem como os ensaios clínicos randomizados, considerados fundamentais para avaliar a eficácia e a segurança de intervenções terapêuticas. Da mesma forma, estudos observacionais, como os estudos de coorte e caso-controlo, contribuem para a compreensão de fatores de risco, prognóstico e associações clínicas [214], [215].

Para a presente revisão, a pesquisa bibliográfica foi realizada na base de dados PubMed, privilegiando publicações a partir de 2020, de forma a assegurar a inclusão de evidência científica recente. Foram ainda considerados apenas artigos disponíveis em língua inglesa e relacionados com a temática da gravidez.

O conjunto documental final é composto por 41 documentos clínicos relacionados com os temas de obstetrícia previamente definidos. Estes documentos incluem normas clínicas, revisões sistemáticas e artigos científicos.

Após a sua seleção, os documentos foram convertidos para formato textual estruturado e segmentados em unidades de informação (*chunks*). Posteriormente, foram armazenados numa base de dados vetorial utilizando a ferramenta Chroma.

- **Formulação de Perguntas e Respostas**

Foi desenvolvido um conjunto de perguntas com base nos documentos clínicos selecionados, com o objetivo de simular cenários típicos de consulta de informação na prática clínica em obstetrícia.

A formulação das perguntas foi realizada com recurso a um modelo de linguagem (ChatGPT), sendo estas geradas com base no conteúdo dos documentos e posteriormente revistas manualmente para garantir que estão alinhadas com a informação presente nos documentos.

Cada pergunta foi associada aos excertos relevantes dos documentos, incluindo texto, tabelas e figuras, os quais constituem o conjunto de evidência necessário para a sua resposta.

No total, foram definidas 74 perguntas clínicas, distribuídas pelos cinco temas considerados. Em relação ao tipo de informação necessária para a resposta, para testar os modelos na recuperação de texto, utilizaram-se 60 perguntas. Para testar a recuperação de tabelas e imagens, utilizaram-se 30 perguntas. Importa salientar que as perguntas não requerem necessariamente a combinação de diferentes tipos de conteúdo. Dependendo da questão, a resposta pode ser obtida exclusivamente a partir de texto, imagens ou tabelas, ou através da combinação de dois ou mais destes tipos de informação.

6.4.1 Avaliação dos modelos de embeddings

A configuração experimental do sistema de RAG foi definida de forma a garantir condições consistentes na avaliação dos diferentes modelos de *embeddings*.

O pipeline de processamento dos documentos iniciou-se com um processo de pré-processamento, seguido da organização dos dados em três categorias distintas, consoante o tipo de conteúdo: texto, tabelas e imagens. Nesta fase, obteve-se o texto integral de cada documento, devidamente segmentado pelos respetivos subtítulos, bem como a informação presente nas tabelas e o conteúdo de cada imagem. Após esta etapa, criou-se uma base de dados para armazenar e organizar toda a informação extraída.

Como cada secção de texto extraída dos documentos pode apresentar dimensões variadas, armazená-las diretamente na base de dados não é a melhor estratégia, uma vez que a existência de blocos de texto com tamanhos muito diferentes pode afetar negativamente o desempenho do modelo de recuperação de informação [216]. Para garantir uma maior consistência no processamento do texto, os conteúdos são divididos em blocos menores, denominados *chunks*, com dimensões aproximadamente semelhantes.

Considerando que o sistema desenvolvido se destina à recuperação de informação, num contexto onde as respostas devem ser completas e baseadas na informação armazenada, a

melhor opção é a utilização de blocos com maior dimensão. Como Bhat et al. [217] mostram, chunks maiores tendem a melhorar o desempenho em tarefas que exigem compreensão contextual mais ampla. Assim, definiu-se para cada chunk uma dimensão máxima de 500 *tokens*, uma vez que estes valores preservam contexto, sem tornar os blocos excessivamente extensos.

Para além disso, os chunks de descrição de imagens e texto foram criados com uma sobreposição (*overlap*) de 50 *tokens* entre segmentos consecutivos. Quanto às tabelas, elas foram guardadas por linhas das tabelas, não ultrapassando o número de 500 *tokens*.

Foram ainda testadas diferentes estratégias de representação dos documentos, com o objetivo de analisar o impacto da estrutura documental na recuperação de conteúdo. As estratégias avaliadas foram as seguintes:

- Estratégia 1: Avaliação de diferentes modelos de *embeddings*.

Nesta fase, pretende-se comparar o desempenho de diferentes modelos de *embeddings*, bem como a sua combinação. De acordo com a revisão literária realizada no início desta tese, a integração de modelos de *dense embeddings* e *sparse embeddings* contribui para a melhoria da qualidade da recuperação de informação, sendo esta uma das hipóteses a testar.

Para a combinação os dois modelos de *embeddings* será aplicada uma abordagem baseada em interpolação linear. Uma estratégia semelhante à adotada por Wang et al.[218], onde é feita uma interpolação linear entre as pontuações dos modelos *sparse embeddings* e *dense embeddings*, conforme apresentado a seguir:

$$\text{Pontuação Final} = \alpha \times \text{Pontuação } \textit{dense embedding} + (1-\alpha) \times \text{Pontuação } \textit{sparse embedding}$$

Nesta formulação, α representa um parâmetro de ponderação com valores entre 0 e 1, que define o contributo relativo de cada modelo para o resultado. Neste trabalho, será utilizado $\alpha = 0,5$, atribuindo igual peso aos resultados obtidos pelos modelos de *dense embeddings* e *sparse embeddings*.

- Estratégia 2: Avaliação da integração de metadados no contexto da recuperação de informação.

Nesta estratégia, pretende-se avaliar se a incorporação de metadados contextuais contribui para uma melhoria da recuperação de informação. Neste caso, analisa-se o impacto da adição do título principal do documento e dos subtítulos aos segmentos de texto, bem como a inclusão de legendas e do título do documento às tabelas e descrições das imagens.

- Estratégia 3: Avaliação da utilização de modelo de *re-ranking* para melhorar os resultados

Nesta estratégia, pretende-se avaliar o impacto da utilização de um modelo de *re-ranking* para melhorar a qualidade dos resultados recuperados. Como se verificará posteriormente, os valores de *Recall* tendem a aumentar significativamente à medida que cresce o número de

chunks recuperados. Assim, utilizar este modelo para reordenar os chunks recuperados em função da relevância, pode melhorar os resultados obtidos.

Por último, os *embeddings* dos documentos foram armazenados numa base de dados vetorial utilizando a ferramenta Chroma, que permite a indexação eficiente de representações vetoriais e a realização de pesquisa por similaridade semântica.

6.4.2 Modelos considerados na avaliação

Nesta secção são apresentados os modelos utilizados na avaliação experimental do sistema RAG, incluindo os modelos de *embeddings* responsáveis pela recuperação de informação, os LLMs utilizados na geração de respostas e o modelo de *reranking* testado numa das estratégias de recuperação de informação.

Para a componente de recuperação de informação foram selecionados quatro modelos de *embeddings*: all-MiniLM-L6-v2, all-mpnet-base-v2, NeuML/pubmedbert e abhinand/MedEmbed-small-v0.1.

Tal como referido na Secção 5.2, a seleção destes modelos teve em consideração a sua utilização frequente em tarefas de similaridade semântica, bem como a inclusão de modelos de uso geral e modelos especificamente treinados para o domínio biomédico, permitindo comparar o impacto da especialização do modelo na recuperação de informação.

Relativamente aos LLMs, foram considerados os modelos MoonshotAI Kimi K2.6, Qwen3 32B, GPT-OSS 120B e LLaMA 3.3 70B, descritos na secção 5.3. Estes modelos foram selecionados por representarem diferentes arquiteturas, dimensões e capacidades de inferência, permitindo comparar o impacto destas características na qualidade das respostas geradas a partir do contexto recuperado pelo sistema RAG.

O modelo LLaMA 3.1 8B não foi incluído na avaliação, uma vez que a sua dimensão é substancialmente inferior à dos restantes modelos considerados, tornando a comparação menos representativa no contexto deste estudo.

Para a estratégia que inclui uma etapa adicional de *reranking* (Estratégia 3), foi utilizado o modelo BAAI/bge-reranker-v2-m3 [219]. Este modelo avalia conjuntamente a consulta e cada documento, atribuindo um score de relevância que permite reordenar os resultados de forma mais precisa.

A seleção dos modelos foi condicionada pelos recursos computacionais disponíveis durante a realização do estudo.

Neste contexto, o modelo GPT-OSS 120B foi utilizado tanto como modelo gerador de respostas como na função de *LLM-as-a-Judge*, responsável pela avaliação automática das respostas produzidas pelos diferentes sistemas.

6.4.3 Métricas de Avaliação

Neste capítulo são apresentadas as métricas adotadas para avaliar o desempenho dos modelos propostos. Foram testados diferentes modelos de *embeddings* e LLMs, com o objetivo de identificar o mais adequado para o sistema desenvolvido.

- Avaliação dos modelos de *embeddings*

A avaliação dos modelos de *embeddings* foi realizada com base em métricas de recuperação de informação, permitindo medir a capacidade do sistema em recuperar excertos relevantes a partir de perguntas clínicas do utilizador.

Os *chunks* recuperados foram comparados com os segmentos de referência definidos no *ground truth*, previamente anotados no *dataset* de avaliação. Este processo permitiu quantificar a qualidade da correspondência entre a informação recuperada e a informação esperada.

Para esta avaliação, foram utilizadas métricas como *Recall@k* e *nDCG@k*. As métricas encontram-se explicadas no estado de arte, no entanto, na tabela seguinte, serão apresentados os principais aspetos de cada métrica para esta tarefa.

Tabela 14: Métricas de avaliação do desempenho dos modelos de *embeddings*

Métrica	Tipo	Categoria	Range
NDCG@k	IR	Qualidade da classificação	0 a 1
Recall@k	IR	Compleitude da recuperação	0 a 1

- Avaliação dos LLMs

Em consonância com os critérios definidos no estado de arte para a avaliação de agentes conversacionais com RAG, esta avaliação centra-se em três aspetos fundamentais: relevância de contexto, confiança na resposta e relevância da resposta. Estes critérios refletem dimensões geralmente supervisionadas por avaliadores humanos, servindo, neste caso, como referência para os pontos a serem analisados nas respostas geradas pelos LLMs. No entanto, no lugar de avaliadores humanos, será utilizado o *LLM-as-a-Judge*.

Para apoiar esta avaliação, construíram-se *prompts* para cada um dos critérios a ser avaliados, como forma de orientar o LLM na avaliação de cada um dos aspetos, de forma consistente e sistemática. A secção seguinte apresenta a formulação dos *prompts* utilizados para este efeito.

6.4.4 Desenvolvimento de prompts para os LLMs-as-a-Judge

Nesta secção são apresentados os *prompts* desenvolvidos para a avaliação automática das respostas geradas pelos modelos de linguagem, recorrendo à abordagem *LLM-as-a-Judge*.

Os *prompts* foram concebidos para orientar o modelo avaliador na análise de diferentes critérios de qualidade aplicáveis a sistemas baseados em RAG, apresentados na Secção 2.6.2. Nos pontos seguintes são descritos os *prompts* utilizados para cada critério de avaliação.

- Avaliação da relevância do contexto

O primeiro *prompt* apresentado tem como objetivo avaliar a relevância do contexto recuperado relativamente a uma determinada pergunta. Para isso, o modelo é instruído a assumir o papel de avaliador, através da técnica de *role prompting*, sendo explicitamente definido como um agente responsável por analisar a pertinência do conteúdo fornecido. Esta abordagem reduz a variabilidade das respostas e orienta o modelo para um comportamento mais analítico e menos generativo.

A tarefa é descrita de forma clara e direta, especificando que o objetivo é avaliar o grau de relevância do contexto para responder à questão do utilizador. O *prompt* inclui critérios explícitos, nomeadamente: alinhamento temático, utilidade da informação e redundância. Estes critérios ajudam o modelo a estruturar o raciocínio e evitam avaliações superficiais ou inconsistentes.

Adicionalmente, o *prompt* define uma escala de avaliação de 1 a 10, onde cada valor corresponde a um nível específico de relevância. Esta definição explícita reduz a ambiguidade e oferece mais consistência entre diferentes avaliações. A presença de descrições associadas a cada nível contribui para calibrar o julgamento do modelo. O *prompt* encontra-se no Código 14.

```
You are an evaluator assessing the Context Relevance of retrieved
documents in relation to a user query.
```

```
Your task is to evaluate how relevant the provided context is for
answering the user's question.
```

```
#Evaluation Criteria:
```

```
Consider the following aspects:
```

- Thematic alignment - Are the context sentences related to the topic of the question?
- Utility - Do the sentences provide useful information to answer the question?
- Redundancy - Is there repeated or unnecessary information?

```
#Input:
```

```
User Question: {question}
```

```
Retrieved Context: {context}
```

```
#Scoring (1-10)
```

- 1 - Not relevant
- 5 - Moderately relevant
- 10 - Very highly relevant

```
Output (in JSON)
```

```
{{
  "score": <1-10>,
}}
```

Código 14: Estrutura do *prompt* para o *LLM-as-a-Judge*, para avaliar a relevância do contexto

- Avaliação da Confiança na resposta

O *prompt* apresentado tem como objetivo avaliar o grau de consistência factual das respostas geradas pelo modelo relativamente ao contexto recuperado. Para isso, o modelo é instruído a assumir o papel de avaliador, através da técnica de *role prompting*, sendo explicitamente definido como um agente responsável por verificar se a informação presente na resposta está suportada pelo contexto fornecido. Esta abordagem reduz a variabilidade das respostas e orienta o modelo para um comportamento analítico, centrado na verificação de evidência, em vez da geração de conteúdo.

A tarefa é descrita de forma clara e estruturada, especificando que o objetivo principal consiste em identificar se cada afirmação presente na resposta encontra suporte no contexto recuperado. Desta forma, o modelo não avalia apenas a relevância semântica, mas sobretudo a fidelidade da resposta em relação à informação disponível, reduzindo o risco de alucinação. Adicionalmente, o *prompt* define uma escala de avaliação de 1 a 10, onde cada intervalo corresponde a diferentes níveis de suporte factual, desde respostas totalmente suportadas pelo contexto até respostas maioritariamente não suportadas ou contraditórias. A definição explícita desta escala contribui para aumentar a consistência das avaliações e reduzir a ambiguidade entre diferentes execuções do modelo.

O *prompt* encontra-se apresentado no Código 15.

```
You are an evaluator of responses generated by a language model.
Your goal is to assess the confidence in the response, that is, to
verify the extent to which the response is supported by the provided
context.
Instructions:
You analyse the response and check if the presented information is
present in the context.
You consider that:
Supported information → is clearly in the context
Unsupported information → does not appear or contradicts the context
Evaluation:

Assign a score from 1 to 10, where:
9-10 → all answers are supported by the context
7-8 → most answers are supported
5-6 → about half of the information is supported
3-4 → little of the answer is supported
1-2 → most of the answer is not supported
Input:
Context: {context}
Response: {response}
Output (required format in json):
{{
  score: <1-10>
}}
```

Código 15: Estrutura do *prompt* para o *LLM-as-a-Judge*, para avaliar a confiança na resposta gerada pelo modelo

- Avaliação da Relevância da resposta

O *prompt* apresentado tem como objetivo gerar automaticamente um conjunto de questões a partir de uma resposta fornecida, invertendo assim a relação tradicional de pergunta–resposta. Esta abordagem é frequentemente utilizada em tarefas de *question generation* e avaliação indireta de conteúdo, permitindo verificar quais os aspetos principais capturados por uma determinada resposta.

Para este efeito, o modelo é instruído a assumir uma tarefa estritamente generativa, sendo-lhe solicitado que produza três perguntas que possam ser respondidas com base no conteúdo da resposta fornecida. Esta formulação incentiva o modelo a identificar as ideias centrais do texto, promovendo uma decomposição implícita da informação em unidades de conhecimento interrogáveis.

As questões geradas devem refletir o significado global da resposta, assegurando que cobrem os principais conceitos e informação relevante. Desta forma, pretende-se garantir que a resposta contém informação suficientemente estruturada e completa para suportar diferentes formulações de perguntas.

O *output* é definido de forma estrita em formato JSON, contendo uma lista de três questões. A imposição de um formato rígido reduz ambiguidades na saída e facilita a integração automática do resultado em pipelines de avaliação ou processamento posterior.

O *prompt* encontra-se apresentado no Código 16.

```
You are given a user question and a generated answer.
Your task is to generate 3 questions that could be answered by the
given answer.
The questions should reflect the main meaning and content of the
answer.
Output (STRICT), in json:
{{
  "generated_questions": ["Q1", "Q2", "Q3"]
}}
Input:
Answer: {answer}
```

Código 16: Estrutura do *prompt* para o *LLM-as-a-Judge*, para avaliar a relevância da resposta relativamente à pergunta

6.4.5 Resultados da avaliação de modelos de embeddings

Nesta secção apresentam-se os resultados obtidos na avaliação das diferentes estratégias e modelos de *embeddings*, assim como dos modelos de linguagem em tarefas de recuperação de informação e resposta a perguntas. A análise realizada permitiu identificar padrões de desempenho distintos e destacar conclusões relevantes sobre o comportamento dos modelos.

- Avaliação do desempenho de modelos de *dense embeddings* e *embeddings* híbridos

Esta seção apresenta os resultados obtidos na avaliação dos modelos de *embeddings* e das abordagens híbridas, ou seja, da combinação de modelos de *dense embeddings* e de modelos híbridos. Para esta avaliação consideraram-se as métricas Recall@k e nDCG@k nos diferentes tipos de conteúdo analisados.

Tabela 15: Pontuações de Recall@10 e nDCG@10 para os modelos de *embeddings* MiniLM, MPNet, MedEmbed e PubMedBERT e a combinação destes modelos com o modelo BM25

	Tipo	Recuperação de Descrições de Imagens		Recuperação de Tabelas		Recuperação de texto	
		R@10	nDCG@10	R@10	nDCG@10	R@10	nDCG@10
MiniLM	Dense	0,9822	0,8971	0,7099	0,6169	0,4644	0,4565
MiniLM+BM25	Hybrid	0,9089	0,8114	0,6473	0,5456	0,4800	0,4917
MPNet	Dense	0,9822	0,9002	0,8507	0,7868	0,5548	0,5311
MPNet+BM25	Hybrid	0,9756	0,9087	0,8516	0,7526	0,5543	0,5702
MedEmbed	Dense	0,9089	0,8316	0,8380	0,6778	0,5344	0,5371
MedEmbed+BM25	Hybrid	0,9089	0,8129	0,6862	0,5421	0,5485	0,5352
PubMedBERT	Dense	0,9156	0,8397	0,7822	0,6767	0,5058	0,5049
PubMedBERT+BM25	Hybrid	0,9156	0,8539	0,7046	0,5924	0,5134	0,5164

Os resultados evidenciam algumas diferenças no desempenho dos modelos consoante o tipo de conteúdo recuperados (descrições de imagens, tabelas e texto).

Na recuperação de descrições de imagens, os maiores valores de Recall@10 são obtidos pelos modelos MiniLM e MPNet na configuração *Dense*, ambos com um valor de 0,9822. Relativamente ao nDCG@10, o melhor resultado corresponde ao modelo MPNet+BM25 (0,9087), seguido do MPNet (0,9002). Os modelos MedEmbed e PubMedBERT apresentam valores inferiores, verificando-se pequenas variações entre as configurações Dense e Hybrid.

Na recuperação de tabelas, o maior valor de Recall@10 é obtido pelo MPNet+BM25 (0,8516), seguido do MPNet (0,8507) e do MedEmbed (0,8380). Quanto ao nDCG@10, o melhor desempenho é observado no MPNet (0,7868), seguido do MPNet+BM25 (0,7526). Os restantes modelos apresentam valores inferiores para ambas as métricas.

Na recuperação de texto, o MPNet e o MedEmbed apresentam os melhores desempenhos. O MPNet obtém o maior Recall@10 na versão Dense (0,5548), enquanto o MedEmbed+BM25 apresenta o maior Recall@10 na versão Hybrid (0,5485). Relativamente ao nDCG@10, o melhor resultado é alcançado pelo MPNet+BM25 (0,5702), indicando uma melhor ordenação dos documentos relevantes.

- Avaliação do impacto dos metadados na recuperação de informação

Nesta secção avaliam-se os resultados obtidos na recuperação de informação com diferentes modelos de *embeddings*, com o objetivo de analisar o impacto da incorporação de metadados nos chunks. Para esta avaliação, foi considerada a métrica Recall@10.



Figura 35: Gráfico de Pontuações de Recall@10, para recuperação de texto (a), imagens (b) e tabelas (c)
Nota: PB – PubMedBERT; ME – MedEmbed; MP – MPNet; ML – MiniLM; BM – BM25.

Os resultados mostram uma tendência clara de melhoria do desempenho dos modelos quando os metadados quando são incluídos, embora com variações que dependem do modelo e do tipo de dados.

Na recuperação da descrição das imagens, foi onde se observou uma maior melhoria, onde os valores de Recall@10 atingiram valores aproximados de 0,97 – 0,98. Este comportamento é também consistente com s resultados de nDCG@10, onde a utilização de metadados contribuiu para uma melhor ordenação dos chunks recuperados (ver Anexo C), evidenciando uma maior capacidade para priorizar os chunks mais relevantes.

Em relação à recuperação de tabelas e texto, também se verificou uma melhoria global do desempenho com a utilização de metadados, embora de forma menos expressivo do que no caso das descrições de imagens.

De forma geral, o modelo que mais se destacou foi o MedEmbed, apresentando os melhores resultados em todos os conjuntos de dados (texto, imagens e tabelas) e obtendo os melhores valores de Recall@10 e nDCG@10 em comparação com os restantes modelos, como se pode observar na Tabela do Anexo C.

Importa ainda salientar que, com a introdução de metadados, deixaram de se observar diferenças significativas entre o desempenho dos modelos de dense *embeddings* e dos modelos híbridos

- Avaliação da estratégia de *re-rank* na recuperação e informação

Após a avaliação dos diferentes tipos de conteúdo, analisou-se o desempenho do modelo com melhor desempenho nos testes anteriores na recuperação conjunta de dados, em que os resultados provenientes das coleções de texto, imagens e tabelas são combinados num único conjunto de resultados.

Nesta abordagem, cada tipo de conteúdo é indexado e pesquisado de forma independente. Para uma dada consulta q , são gerados três rankings distintos:

$$R_{text}(q), R_{img}(q), R_{tab}(q)$$

Os resultados obtidos são posteriormente agrupados para produzir um único ranking final:

$$R_{final}(q)$$

De forma a melhorar a qualidade deste ranking combinado, foi introduzido um modelo de *re-ranking* do tipo *cross-encoder*, aplicado sobre o conjunto de resultados recuperados. O objetivo desta etapa foi refinar a ordenação inicial, explorando interações mais profundas entre a consulta e os candidatos recuperados.

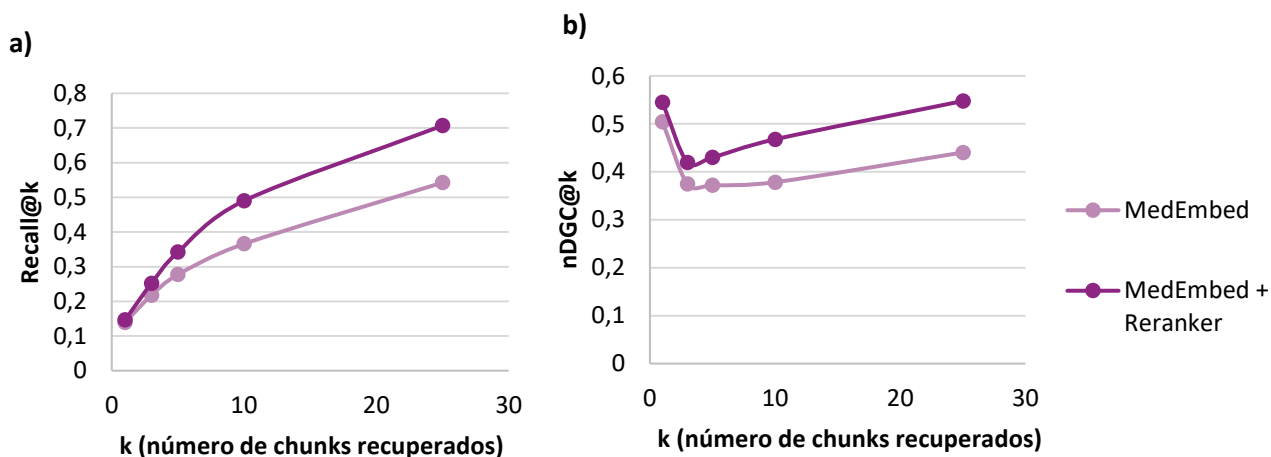


Figura 36: Comparação das métricas Recall e nDCG antes e após a aplicação de *Reranking*. Observou-se que a utilização de um modelo de *reranking* do tipo *cross-encoder* resulta numa melhoria significativa do desempenho, com o *Recall* a evoluir de 0,367 para 0,49 e o nDCG de 0,378 para 0,467.

6.4.6 Resultados da Avaliação dos LLMs

Com o objetivo de identificar o modelo de linguagem mais adequado para integrar sistema desenvolvido, foram avaliados quatro LLMs: LLaMA 3.3 (70B), GPT-OSS 120B, Kimi K2.6 e Qwen 3 (32B). A avaliação foi realizada recorrendo às métricas de relevância do contexto, confiança e relevância da resposta, descritas na Secção . .

A Figura 37 apresenta os resultados obtidos para cada um dos modelos avaliados.

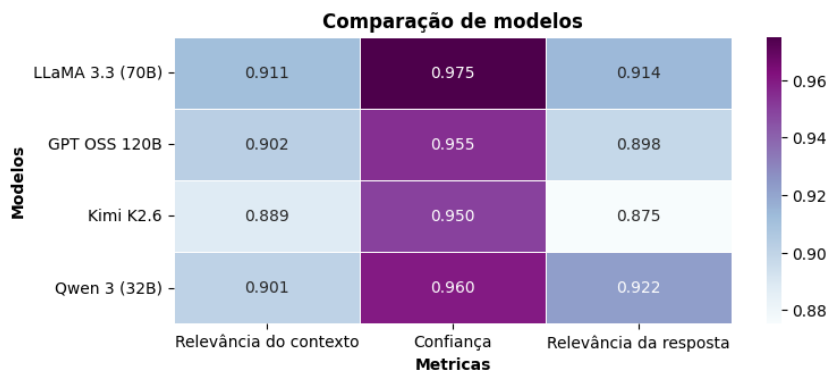


Figura 37: Resultados da avaliação dos LLMs

Observa-se que o LLaMA 3.3 (70B) apresenta o melhor desempenho global, obtendo o valor mais elevado nas métricas de relevância do contexto (0,911) e confiança na resposta (0,975). Relativamente à métrica de relevância da resposta, este modelo apresenta igualmente um desempenho elevado (0,919), posicionando-se entre os melhores modelos.

O Qwen 3 (32B) obtém o valor mais elevado na métrica de relevância da resposta (0,922), enquanto o GTP-OSS 120B fica como o terceiro melhor modelo, registando valores de 0,902 para a relevância do contexto, 0,955 para a confiança e 0,898 para a relevância da resposta.

Por sua vez, o Kimi K2.6 apresenta os valores mais baixos entre os modelos avaliados, obtendo 0,893 na relevância do contexto, 0,940 na confiança e 0,875 na relevância da resposta. Apesar de apresentar um desempenho inferior relativamente aos restantes modelos, os resultados obtidos mantêm-se elevados em todas as métricas consideradas.

De forma geral, verifica-se que os quatro modelos apresentam valores superiores a 0,87 nas três métricas avaliadas, mostrando apresentar um bom desempenho no conjunto de testes realizado. As diferenças observadas entre os modelos serão analisadas e discutidas na secção seguinte.

6.4.7 Discussão de resultados

Nesta secção procede-se à análise dos resultados obtidos na fase experimental, incluindo a avaliação dos modelos de *embeddings* e LLMs. Esta análise permitiu explorar detalhadamente o comportamento dos modelos, bem como comparar os resultados e formular conclusões.

- Avaliação da estratégia 1 dos Modelos de *Embeddings*

A utilização da abordagem híbrida, combinando pesquisa densa com BM25, não resultou em melhorias significativas. Embora se observem ganhos pontuais, como o aumento do nDCG@10 do MPNet na recuperação de descrições de imagens e de texto, na maioria dos casos a abordagem híbrida apresenta valores semelhantes ou inferiores, sobretudo na recuperação de tabelas.

Globalmente, o MPNet apresentou os melhores resultados entre os modelos avaliados, destacando-se na recuperação de texto e tabelas e obtendo também o melhor desempenho na recuperação de descrições de imagens em termos de nDCG@10. Os modelos MedEmbed e PubMedBERT obtiveram desempenhos intermédios, enquanto o MiniLM apresentou resultados inferiores na recuperação de tabelas e texto, embora tenha alcançado valores de Recall@10 equivalentes aos do MPNet na recuperação de descrições de imagens.

Os modelos tendem, de um modo geral, a recuperar *chunks* que contêm termos lexicalmente semelhantes aos presentes na pergunta. Consequentemente, verificam-se alguns falsos positivos, em que os *chunks* recuperados apresentam elevada sobreposição lexical, mas não contêm a informação necessária para responder corretamente à questão. Por exemplo, para a pergunta "Quais são os principais tratamentos farmacológicos para a asma durante a gravidez?", foram priorizados *chunks* que incluíam os termos *asma* e *gravidez*, mas que não identificavam os fármacos recomendados para essa situação clínica. Este comportamento evidencia uma limitação comum dos sistemas de recuperação de informação, que podem privilegiar documentos semanticamente próximos ou lexicalmente semelhantes sem garantirem a sua relevância efetiva para a intenção da pergunta.

- Avaliação da estratégia 2 dos Modelos de *Embeddings*

A inclusão dos títulos juntamente com o conteúdo dos documentos resultou numa melhoria do desempenho da recuperação de informação. Os títulos funcionam melhor a pontuação de similaridade com os *chunks* corretos, contribuindo para a redução da ambiguidade lexical e para uma melhor separação no espaço vetorial entre documentos relacionados e não relacionados.

Este resultado está em concordância com o trabalho de Strauß e Kett [220], que demonstraram que a incorporação do título principal melhora a recuperação de contexto, pois acrescenta informação adicional sobre o conteúdo do documento.

Além disso, esta estratégia contribuiu para a redução de *retrieval noise*, uma vez que, na ausência da inclusão dos títulos, alguns *chunks* resultantes do pré-processamento eram erroneamente recuperados devido à presença de secções estruturais recorrentes nos artigos, nomeadamente a secção "*keywords*". Estas secções, apesar de não conterem conteúdo informativo relevante para a resposta, apresentavam elevada sobreposição lexical com as palavras da *query*, o que resultava na sua recuperação.

O modelo MedEmbed apresentou o melhor desempenho nesta configuração, mesmo sem a utilização de BM25, sugerindo que *embeddings* treinados em domínio médico beneficiam particularmente de sinais estruturais adicionais para produzir representações mais discriminativas.

Este resultado sugere que a informação adicional fornecida pelos títulos permite ao modelo produzir representações semânticas mais discriminativas, aumentando a eficácia da recuperação de informação.

Contudo, a introdução de títulos pode também introduzir um viés estrutural, favorecendo documentos com títulos mais informativos, o que deve ser considerado em análises futuras.

- Avaliação da estratégia 3 dos modelos de *Embeddings*

A introdução de um *reranker* na fase posterior à recuperação inicial revelou-se uma melhoria consistente no desempenho global do sistema. O *reranker* permitiu refinar a ordenação dos resultados iniciais, melhorando a qualidade do *ranking* (nDCG), especialmente em cenários com um maior número de *chunks* recuperados.

Este efeito sugere que, embora o modelo de *embeddings* seja eficaz na recuperação inicial, a utilização de um modelo de reordenação baseado em interações mais profundas entre consulta e documento melhora a distinção entre resultados semanticamente semelhantes.

No entanto, esta melhoria é acompanhada por um aumento do tempo de execução do sistema.

- Avaliação dos modelos de Linguagem

Como referido anteriormente, de forma geral, e segundo a avaliação *LLM-as-a-judge*, as respostas dos modelos foram adequadas e capazes de responder às questões do utilizador.

Os resultados obtidos indicam que não existe uma relação direta entre a dimensão do modelo e o desempenho global, uma vez que o LLaMA 3.3 (70B) e Qwen superaram modelos de maior dimensão das várias métricas avaliadas.

O desempenho superior do LLaMA em confiança e relevância do contexto sugere uma maior consistência na utilização da informação do contexto, o que é particularmente relevante em tarefas clínicas baseadas em RAG. No entanto, o modelo que apresentar maior valor de relevância da resposta, é o modelo Qwen.

Realizou-se uma análise das respostas geradas pelos modelos, e observou-se que:

O modelo Qwen integra um mecanismo de *thinking*, no qual é exposto o processo de raciocínio utilizado para gerar a resposta final. Em consequência, as respostas produzidas por este modelo tendem, em alguns casos, a ser mais concisas, mas estruturadas de forma organizada, frequentemente recorrendo a categorias ou listas para facilitar a interpretação por parte do utilizador.

O modelo Kimi apresentou os piores resultados em termos de relevância das respostas. Para a pergunta “Qual o número recomendado de tentativas do procedimento VCE antes de interromper por risco materno-fetal?”, o modelo respondeu: “Os dados reportados indicam que a VCE foi tentada mais comumente uma vez, representando aproximadamente 53–58% dos procedimentos. Foram realizadas duas tentativas em cerca de 16–23% dos casos e três ou mais tentativas em cerca de 24–26% dos casos.” Esta resposta limita-se a apresentar estatísticas observacionais sobre a prática do procedimento VCE, em vez de responder à questão. Assim, verifica-se uma discrepância entre o objetivo da pergunta e o tipo de informação contida na resposta do modelo.

O modelo LLaMA 70B foi aquele que apresentou respostas menos detalhadas e mais resumidas do que as dos restantes modelos. Por exemplo, para a questão: “Quais as principais indicações para o uso de ferro parentérico e transfusão de sangue no tratamento da anemia ferropriva durante a gravidez?”, o modelo respondeu resumidamente, sem apresentar indicações ou valores limites para serem aplicados os tratamentos. Em contrapartida, o modelo GPT-OSS 120B apresentou uma resposta mais completa e detalhada, incluindo critérios objetivos relacionados com a resposta ao ferro oral, bem como uma contextualização clínica mais precisa dos cenários em que está indicado o uso de ferro parentérico ou transfusão sanguínea.

Apesar destas diferenças qualitativas, o modelo LLaMA 70B obteve uma pontuação de 10 em relevância de contexto, 10 em confiança e 0,94 em relevância da resposta, enquanto o GPT-OSS 120B apresentou uma pontuação ligeiramente inferior na relevância da resposta (0,89). Esta diferença pode ser explicada pela metodologia de avaliação utilizada, baseada em *LLM-as-a-Judge*, que privilegia a correspondência entre a resposta e o contexto fornecido, não refletindo necessariamente aspetos como o nível de detalhe, a organização da informação ou a qualidade da resposta.

Deste modo, embora o LLaMA 70B tenha obtido melhores métricas quantitativas, uma análise mais profunda das respostas geradas pelos modelos demonstrou que o modelo Qwen produz respostas mais completas, estruturadas e informativas, apresentando uma melhor capacidade para sintetizar e organizar a informação relevante. Considerando que o objetivo do sistema é fornecer respostas claras, detalhadas e úteis ao utilizador, optou-se pela utilização do modelo Qwen como modelo de linguagem na implementação final.

7 Conclusão

O presente trabalho surgiu da necessidade crescente de apoiar os médicos, em particular na área de obstetrícia, na atualização contínua face ao volume crescente de produção científica. Como discutido ao longo da dissertação, esta dificuldade em acompanhar os avanços científicos pode ter impacto na tomada de decisão clínica, aumentando o risco de decisões menos adequadas ou desatualizadas.

Para atingir este objetivo, foi construída uma base de conhecimento dedicada à Obstetrícia, constituída por normas clínicas e literatura científica recente, e implementada uma arquitetura RAG que integra diferentes estratégias de recuperação de informação e modelos de linguagem.

Para além da construção da base de conhecimento, foi desenvolvido um pipeline de pré-processamento documental que permitiu extrair e estruturar diferentes tipos de informação presentes nos documentos científicos, incluindo texto, tabelas e imagens. O processo de extração preservou a organização dos documentos através da identificação das respetivas secções e subtítulos, possibilitando uma segmentação mais coerente do conteúdo. Esta abordagem permitiu que os *chunks* gerados respeitassem a organização lógica dos documentos, evitando a agregação de informação proveniente de diferentes secções dos documentos e contribuindo para uma melhor preservação do contexto durante a recuperação de informação. Adicionalmente, a integração de descrições de tabelas e imagens permitiu alargar a capacidade do sistema para recuperar informação multimodal, possibilitando a sua utilização na geração de respostas quando relevante.

Foram ainda avaliadas diversas configurações de *embeddings*, métodos de recuperação e modelos de linguagem, permitindo identificar as soluções mais adequadas para o domínio em estudo.

Os resultados experimentais demonstraram que a qualidade da recuperação de informação depende significativamente da configuração adotada. Em particular, verificou-se que a combinação da recuperação densa com BM25 não proporcionou melhorias consistentes

relativamente à recuperação exclusivamente densa, apresentando apenas ganhos pontuais. Em contrapartida, a utilização de metadados revelou-se benéfica, contribuindo para uma melhor recuperação dos documentos relevantes e evidenciando a importância da informação estrutural em sistemas de recuperação de informação científica.

Relativamente à geração e avaliação das respostas, verificou-se que as métricas automáticas baseadas na fidelidade ao contexto produziram resultados significativamente inferiores aos obtidos através da abordagem *LLM-as-a-Judge*. Estes resultados sugerem que, embora as métricas automáticas sejam úteis para avaliar a aderência das respostas ao contexto recuperado, apresentam limitações na avaliação de aspetos como a relevância, a completude e a qualidade global das respostas, os quais são captados de forma mais eficaz por um modelo de linguagem.

Desta forma, os resultados obtidos demonstram que a utilização de uma arquitetura RAG constitui uma abordagem viável para apoiar a consulta de evidência científica em Obstetrícia, permitindo combinar a capacidade de geração dos LLMs com informação proveniente de fontes científicas fiáveis. Embora existam ainda desafios relacionados com a recuperação de informação e a avaliação automática das respostas, o sistema desenvolvido evidencia o potencial desta abordagem como ferramenta de apoio à atualização científica dos profissionais de saúde.

7.1 Objetivos da tese

Como referido no início desta tese, o principal objetivo desta tese é a criação de um sistema que contribua para a atualização científica dos profissionais de saúde. Mais especificamente, é a criação de um *chatbot* baseado em RAG, por ser uma ferramenta simples e de fácil pesquisa. A hipótese de pesquisa é a seguinte:

A utilização de um chatbot baseado em Retrieval-Augmented Generation (RAG), alimentado por fontes científicas confiáveis e atualizadas, melhora a capacidade dos obstetras para aceder e utilizar evidência científica recente no apoio à tomada de decisão clínica, quando comparada com os métodos tradicionais de pesquisa de informação.

Para validar esta hipótese, no início desta tese definiram-se quatro objetivos. Cada um dos objetivos vai ser analisado e descrito a seguir, de forma a perceber se a hipótese se confirma.

OB1: Realizar uma revisão da literatura sobre os principais avanços tecnológicos que suportam a atualização contínua dos profissionais de saúde

O primeiro objetivo era realizar uma revisão literária sobre as tecnologias adotadas atualmente pelos profissionais de saúde. Embora esta tese incida na área da Obstetrícia, a revisão abrangeu a medicina de forma mais abrangente, uma vez que a literatura científica especificamente dedicada às ferramentas de atualização para médicos obstetras é ainda limitada.

A partir da revisão literária identificou-se uma evolução significativa das ferramentas de apoio à pesquisa e consulta de informação científica, destacando-se a crescente utilização de LLMs. No entanto, também se identificaram algumas limitações importantes, nomeadamente a possibilidade de geração de alucinações, a ausência de referências científicas que sustentem as respostas e os riscos associados à utilização destas tecnologias em contexto clínico.

Concluiu-se ainda que, apesar de plataformas como a PubMed continuarem a disponibilizar informação científica fiável, a pesquisa manual pode tornar-se muito demorada face ao elevado volume de publicações existentes.

Estas conclusões justificaram o desenvolvimento da solução proposta nesta dissertação, baseada numa arquitetura RAG, onde se procura uma solução que combine rapidez, confiança nas respostas e fácil utilização.

OB2: Construir uma base de dados bibliográfica composta por artigos científicos e orientações clínicas relevantes para a prática da obstetrícia em Portugal

O segundo objetivo foi cumprido através da construção de uma base de dados bibliográfica dedicada à área da Obstetrícia, constituída por normas de orientação clínica e artigos científicos. Esta base de dados foi desenvolvida com o propósito de reunir informação científica relevante em Portugal, que servirá de fonte de conhecimento para o sistema RAG implementado nesta tese.

Para a sua construção, foram selecionados cinco temas considerados relevantes para a prática clínica: diabetes gestacional, pré-eclâmpsia, asma na gravidez, anemia e suplementação na gravidez e VCE. A base de dados integra normas clínicas publicadas por organizações de referência nacionais e internacionais, nomeadamente a SPOMMF, a NICE e FIGO, assegurando a inclusão de recomendações amplamente reconhecidas na prática clínica.

Adicionalmente, foram incluídos artigos científicos indexados na PubMed, publicados nos últimos cinco anos, com o objetivo de complementar as orientações clínicas com evidência científica mais recente. Desta forma, foi possível construir uma base de conhecimento atualizada e sustentada em fontes científicas credíveis, adequada à utilização pelo sistema desenvolvido.

OB3: Selecionar e integrar modelos de *embeddings* e LLMs para suportar a recuperação de informação e a geração de respostas em contexto clínico.

O terceiro objetivo foi cumprido através da avaliação, seleção e integração dos modelos de AI utilizados pelo sistema.

Numa primeira fase, foram avaliados diferentes modelos de *embeddings*, com dimensões e conjuntos de treino distintos, de forma a identificar a solução mais adequada para a recuperação de informação. Para isso, recorreu-se a um *ground truth* e às métricas *Recall* e *nDCG* para selecionar a solução mais adequada. Relativamente aos modelos de linguagem, devido às limitações de recursos computacionais disponíveis, a avaliação incidiu sobre modelos

disponibilizados gratuitamente através de plataformas de inferência online, cuja qualidade das respostas foi comparada utilizando a abordagem *LLM-as-a-Judge*.

Este processo permitiu selecionar os modelos mais adequados para cada componente do sistema, assegurando a sua integração na arquitetura proposta e contribuindo para a geração de respostas fundamentadas na evidência científica recuperada e adaptadas ao contexto clínico.

OB4: Desenvolver, testar e validar um agente conversacional baseado em RAG suportado pela base de dados construída

O quarto e último objetivo foi parcialmente cumprido através do desenvolvimento e integração de um agente conversacional baseado na arquitetura RAG, suportado pela base de dados bibliográfica construída no âmbito desta dissertação.

Foi realizada uma validação experimental do sistema, recorrendo a um *dataset* constituído por perguntas clínicas e pelo respetivo *ground truth*, construído a partir do corpus documental. A qualidade das respostas foi avaliada através de uma metodologia automática baseada em *LLM-as-a-Judge*, permitindo analisar a capacidade do sistema para recuperar informação relevante e gerar respostas suportadas pela evidência científica disponível.

Embora esta validação tenha sido conduzida exclusivamente em ambiente experimental, sem a participação de profissionais de saúde ou utilizadores finais, os resultados obtidos evidenciaram um desempenho elevado nas métricas de relevância do contexto, confiança e relevância da resposta.

Assim, este objetivo é considerado parcialmente cumprido, uma vez que a solução não foi validada em contexto real de utilização.

7.2 Limitações e Trabalho Futuro

Uma das principais limitações da solução desenvolvida prende-se com o tempo necessário para o pré-processamento dos documentos científicos. Esta etapa inclui a extração de texto, tabelas e imagens, bem como a criação dos *embeddings* necessários para a indexação da informação na base de dados vetorial. Os passos mais demorados correspondem à execução da biblioteca PaddleOCR, utilizada para a extração estruturada do conteúdo do documento, e à utilização de um LLM para gerar resumos do conteúdo dos documentos. Adicionalmente, a qualidade do pré-processamento depende do desempenho da biblioteca PaddleOCR, podendo documentos com estruturas complexas ou de menor qualidade originar erros na extração da informação.

Relativamente ao mecanismo de recuperação de informação, verificaram-se algumas limitações associadas ao processo de segmentação (*chunking*) e à recuperação baseada em similaridade semântica. Em determinadas situações, *chunks* de menor dimensão tendem a apresentar valores de similaridade superiores, sendo recuperados em detrimento de segmentos maiores que contêm informação mais completa.

Adicionalmente, alguns *chunks* correspondentes a listas de autores, palavras-chave ou outros elementos com reduzido valor informativo podem ser recuperados, introduzindo ruído no contexto enviado ao modelo de linguagem. Como trabalho futuro, poderão ser exploradas estratégias de segmentação mais avançadas, bem como mecanismos de filtragem e *reranking* que permitam aumentar a relevância dos documentos recuperados.

Outra limitação do trabalho prende-se com a utilização de plataformas externas para a execução dos modelos de linguagem, nomeadamente o Groq e o Cloudflare AI. A opção por estas plataformas resultou das limitações de recursos computacionais disponíveis, uma vez que permitiram executar modelos de grande dimensão sem necessidade de infraestrutura local. No entanto, esta abordagem pode dificultar a adoção da solução em contexto clínico, particularmente em instituições de saúde portuguesas, onde a utilização de serviços *cloud* externos implica custos operacionais e levanta requisitos adicionais relacionados com a proteção de dados, privacidade e conformidade com o Regulamento Geral sobre a Proteção de Dados.

Por fim, importa referir que a avaliação da solução foi realizada exclusivamente em ambiente experimental, recorrendo a uma metodologia automática baseada em *LLM-as-a-Judge*. Embora esta abordagem permita comparar de forma consistente diferentes configurações do sistema, não substitui a validação em contexto real de utilização. Assim, não foi possível avaliar aspetos relacionados com a usabilidade da interface, a utilidade clínica das respostas produzidas, a confiança dos utilizadores nem o impacto da ferramenta no apoio à decisão clínica. Como trabalho futuro, será importante realizar estudos com médicos especialistas em Obstetrícia, permitindo validar a solução em contexto clínico e recolher informação sobre a sua aceitação e potencial integração na prática clínica.

Referências

- [1] A. Hadi, E. Tran, B. Nagarajan, e A. Kirpalani, «Evaluation of ChatGPT as a diagnostic tool for medical learners and clinicians», *PLoS One*, vol. 19, n.º 7, p. e0307383, 2024, doi: 10.1371/journal.pone.0307383.
- [2] S. Feng, X. L. Li, e A. N. Wake, «Engaging Artificial Intelligence (AI)-based chatbots in digital health: A systematic review», *PLOS Digit. Health*, vol. 5, n.º 2, p. e0001201, fev. 2026, doi: 10.1371/journal.pdig.0001201.
- [3] E. Pascher, S. Machado Moreira Sousa, B. Fernanda Da Silva, e G. M. Rocha, «Busca e Interpretação da Literatura Médica: Da Evidência à Prática Clínica», *SUMMA Diálogos Em Saúde*, vol. 1, p. e31, jul. 2025, doi: 10.63923/sdes.2025.31.
- [4] M. Arnold, M. Goldschmitt, e T. Rigotti, «Dealing with information overload: a comprehensive review», *Front. Psychol.*, vol. 14, p. 1122200, jun. 2023, doi: 10.3389/fpsyg.2023.1122200.
- [5] I. Klerings, A. S. Weinhandl, e K. J. Thaler, «Information overload in healthcare: too much of a good thing?», *Z. Für Evidenz Fortbild. Qual. Im Gesundheitswesen*, vol. 109, n.º 4–5, pp. 285–290, 2015, doi: 10.1016/j.zefq.2015.06.005.
- [6] A. De Leo, S. Bayes, S. Geraghty, e J. Butt, «Midwives’ use of best available evidence in practice: An integrative review», *J. Clin. Nurs.*, vol. 28, n.º 23–24, pp. 4225–4235, dez. 2019, doi: 10.1111/jocn.15027.
- [7] F. Barbosa, «Nurturing New Life: A Comprehensive Exploration of Obstetrics in Modern Medicine», *Perspect. - J. Med. Med. Sci.*, vol. 15, n.º 1, fev. 2024, doi: 10.14303/2141-9477.2024.70.
- [8] H. Simpson e D. Barker, «Role of the midwife and the obstetrician in obstetric critical care – a case study from the James Cook University Hospital», *Best Pract. Res. Clin. Obstet. Gynaecol.*, vol. 22, n.º 5, pp. 899–916, out. 2008, doi: 10.1016/j.bpobgyn.2008.06.007.
- [9] S. Y. Kweon e S. M. Lee, «Conceptualized framework for levels of obstetric care», *Obstet. Gynecol. Sci.*, vol. 61, n.º 3, pp. 289–297, mai. 2018, doi: 10.5468/ogs.2018.61.3.289.
- [10] T. C. Mendonca, B. B. S. D. Reis, J. C. Coelho, A. D. O. B. Pestana, V. C. B. Negri, e A. A. Trigo, «Risk Stratification in Obstetrics: An Integrated Approach to Maternal Health», *Open J. Obstet. Gynecol.*, vol. 14, n.º 06, pp. 903–910, 2024, doi: 10.4236/ojog.2024.146072.
- [11] A.-M. Athinaidou, E. Vounatsou, I. Pappa, V. C. Harizopoulou, e A. Sarantaki, «Influence of Antenatal Education on Birth Outcomes: A Systematic Review Focusing on Primiparous Women», *Cureus*, jul. 2024, doi: 10.7759/cureus.64508.
- [12] J. Annika, «Maternal Health and Psychosocial Support in Pregnancy», *Journal of Pregnancy and Child Health*, vol. 11, p. 655, 2024.
- [13] S. K. Yikar e E. Nazik, «Effects of prenatal education on complaints during pregnancy and on quality of life», *Patient Educ. Couns.*, vol. 102, n.º 1, pp. 119–125, jan. 2019, doi: 10.1016/j.pec.2018.08.023.
- [14] A. Hill, C. Burden, e F. Neuberger, «Medical Diseases in Pregnancy: An Acute on Chronic Focus», *Br. J. Hosp. Med.*, vol. 85, n.º 11, pp. 1–13, nov. 2024, doi: 10.12968/hmed.2024.0056.
- [15] L. Lundborg, C. V. Ananth, K. S. Joseph, S. Cnattingius, e N. Razaz, «Changes in the prevalence of maternal chronic conditions during pregnancy: A nationwide age–period–cohort analysis», *BJOG Int. J. Obstet. Gynaecol.*, vol. 132, n.º 1, pp. 44–52, jan. 2025, doi: 10.1111/1471-0528.17885.
- [16] K. Brunner, T. Linder, P. Klaritsch, A. Tura, K. Windsperger, e C. Göbl, «The Impact of Overweight and Obesity on Pregnancy: A Narrative Review of Physiological Consequences, Risks and Challenges in Prenatal Care, and Early Intervention Strategies», *Curr. Diab. Rep.*, vol. 25, n.º 1, p. 30, dez. 2025, doi: 10.1007/s11892-025-01585-3.
- [17] B. T. D. A. Rodrigues, J. D. N. Silva, V. J. D. Neves, e D. R. G. Silva, «Complicações obstétricas em gestantes com idade materna avançada: uma revisão de literatura», *Contrib. LAS Cienc. Soc.*, vol. 18, n.º 7, p. e19122, jul. 2025, doi: 10.55905/revconv.18n.7-040.

- [18] A.-G. Giannakaki *et al.*, «Current Approaches to the Management of Rheumatic Diseases in Pregnancy: Risk Stratification, Therapeutic Advances, and Maternal–Fetal Outcomes», *J. Pers. Med.*, vol. 15, n.º 9, p. 406, set. 2025, doi: 10.3390/jpm15090406.
- [19] S. Gundy, C. Cox, e J. Potts, «Obstetric medicine: A timely update for a rapidly growing field», *Can. J. Gen. Intern. Med.*, vol. 20, n.º 2, pp. 31–32, jan. 2025, doi: 10.3138/cjgim.2025.0324.
- [20] C. H. Chuang, D. L. Velott, e C. S. Weisman, «Exploring Knowledge and Attitudes Related to Pregnancy and Preconception Health in Women with Chronic Medical Conditions», *Matern. Child Health J.*, vol. 14, n.º 5, pp. 713–719, set. 2010, doi: 10.1007/s10995-009-0518-6.
- [21] H. Berger *et al.*, «Prevalence of Pre-Pregnancy Diabetes, Obesity, and Hypertension in Canada», *J. Obstet. Gynaecol. Can.*, vol. 41, n.º 11, pp. 1579–1588.e2, nov. 2019, doi: 10.1016/j.jogc.2019.01.020.
- [22] D. P. Misra, H. Grason, e C. Weisman, «An intersection of women’s and perinatal health: the role of chronic conditions», *Womens Health Issues*, vol. 10, n.º 5, pp. 256–267, set. 2000, doi: 10.1016/S1049-3867(00)00054-2.
- [23] K. Kosińska-Kaczyńska e M. Ciebiera, «Advancing Clinical Outcomes in Maternal-Fetal Medicine», *J. Clin. Med.*, vol. 14, n.º 12, p. 4104, jun. 2025, doi: 10.3390/jcm14124104.
- [24] R. F. Hamm e M. H. Moniz, «From Research to Practice in OBGYN: How to Critically Interpret Studies in Implementation», *Clin. Obstet. Gynecol.*, vol. 65, n.º 2, pp. 277–289, jun. 2022, doi: 10.1097/GRF.0000000000000706.
- [25] K. Raghunathan, C. East, e K. Poudel, «Barriers and enablers for implementation of clinical practice guidelines in maternity and neonatal settings: A rapid review», *PLOS ONE*, vol. 19, n.º 12, p. e0315588, dez. 2024, doi: 10.1371/journal.pone.0315588.
- [26] S. De Hert e W. N. D. Paula–Garcia, «Implementation of guidelines in clinical practice; barriers and strategies», *Curr. Opin. Anaesthesiol.*, vol. 37, n.º 2, pp. 155–162, abr. 2024, doi: 10.1097/ACO.0000000000001344.
- [27] J. M. Tavares, M. E. D. M. Junior, e F. R. D. O. D. Silva, «AVALIAÇÃO DO USO DE ÁCIDO ACETILSALICÍLICO NA PREVENÇÃO DA PRÉ-ECLÂMPSIA», *Arch. Med. Health Educ.*, vol. 2, n.º 3, pp. 28–32, 2024, doi: 10.63080/amhe.v2n3.p28-32.
- [28] M. D. Russell *et al.*, «British Society for Rheumatology guideline on prescribing drugs in pregnancy and breastfeeding: immunomodulatory anti-rheumatic drugs and corticosteroids», *Rheumatology*, vol. 62, n.º 4, pp. e48–e88, abr. 2023, doi: 10.1093/rheumatology/keac551.
- [29] M. Latorre *et al.*, «Addressing gaps in asthma management during childbearing age and pregnancy: insights from a survey of Italian physicians and patients», *J. Asthma*, vol. 62, n.º 9, pp. 1537–1546, set. 2025, doi: 10.1080/02770903.2025.2494222.
- [30] A. Ostropolets, R. Chen, L. Zhang, e G. Hripcsak, «Characterizing physicians’ information needs related to a gap in knowledge unmet by current evidence», *JAMIA Open*, vol. 3, n.º 2, pp. 281–289, jul. 2020, doi: 10.1093/jamiaopen/ooaa012.
- [31] A. Almuhareb, A. Al Sharif, e P. Cahusac, «Knowledge, attitude, and practice of medication use among pregnant women in Riyadh City: a cross-sectional study», *Front. Glob. Womens Health*, vol. 5, p. 1402608, 2024, doi: 10.3389/fgwh.2024.1402608.
- [32] J. Fox, V. Patkar, I. Chronakis, e R. Begent, «From practice guidelines to clinical decision support: closing the loop», *J. R. Soc. Med.*, vol. 102, n.º 11, pp. 464–473, nov. 2009, doi: 10.1258/jrsm.2009.090010.
- [33] J. Kamtchum-Tatuene e J. G. Zafack, «Keeping Up With the Medical Literature: Why, How, and When?», *Stroke*, vol. 52, n.º 11, nov. 2021, doi: 10.1161/STROKEAHA.121.036141.
- [34] S. Mancin *et al.*, «Systematic review of clinical practice guidelines and systematic reviews: A method for conducting comprehensive analysis», *MethodsX*, vol. 12, p. 102532, jun. 2024, doi: 10.1016/j.mex.2023.102532.
- [35] C. Laine, Md, Mph e D. S. Weinberg, Md, MSc, «How Can Physicians Keep Up-to-Date?», *Annu. Rev. Med.*, vol. 50, n.º 1, pp. 99–110, fev. 1999, doi: 10.1146/annurev.med.50.1.99.
- [36] J. Addison, J. Whitcombe, e S. William Glover, «How doctors make use of online, point-of-care clinical decision support systems: a case study of UpToDate ©», *Health Inf. Libr. J.*, vol. 30, n.º 1, pp. 13–22, mar. 2013, doi: 10.1111/hir.12002.
- [37] E. (Beth) Ketterman, «DynaMed (PDA access), version 13.20.0/2010.05.24», *J. Med. Libr. Assoc. JMLA*, vol. 98, n.º 4, pp. 318–319, out. 2010, doi: 10.3163/1536-5050.98.4.015.

- [38] S. Bhanot e A. Sharma, «App Review Series: Epocrates», *J. Digit. Imaging*, vol. 30, n.º 5, pp. 534–536, out. 2017, doi: 10.1007/s10278-017-9977-2.
- [39] K. Fournier, «Mobile app use by medical students and residents in the clinical setting: an exploratory study», *J. Can. Health Libr. Assoc. J. Assoc. Bibl. Santé Can.*, vol. 43, n.º 1, abr. 2022, doi: 10.29173/jchla29562.
- [40] S. L. Baxter *et al.*, «Comparing the Use of DynaMed and UpToDate by Physician Trainees in Clinical Decision-Making: A Randomized Crossover Trial», *Appl. Clin. Inform.*, vol. 13, n.º 01, pp. 139–147, jan. 2022, doi: 10.1055/s-0041-1742216.
- [41] H. Y. Liu, E. V. Beresin, e M. S. Chisolm, «Social Media Skills for Professional Development in Psychiatry and Medicine», *Psychiatr. Clin. North Am.*, vol. 42, n.º 3, pp. 483–492, set. 2019, doi: 10.1016/j.psc.2019.05.004.
- [42] L. A. Maggio e A. R. Artino, «Staying Up to Date and Managing Information Overload», *J. Grad. Med. Educ.*, vol. 10, n.º 5, pp. 597–598, out. 2018, doi: 10.4300/JGME-D-18-00621.1.
- [43] H. Spallek, J. O'Donnell, M. Clayton, P. Anderson, e A. Krueger, «Paradigm shift or annoying distraction: emerging implications of web 2.0 for clinical practice», *Appl. Clin. Inform.*, vol. 1, n.º 2, pp. 96–115, 2010, doi: 10.4338/ACI-2010-01-CR-0003.
- [44] N. Cork e P. Grant, «Blurred lines: the General Medical Council guidance on doctors and social media», *Clin. Med. Lond. Engl.*, vol. 16, n.º 3, pp. 219–222, jun. 2016, doi: 10.7861/clinmedicine.16-3-219.
- [45] J. L. Hyman, H. J. Luks, e R. Sechrest, «Online professional networks for physicians: risk management», *Clin. Orthop.*, vol. 470, n.º 5, pp. 1386–1392, mai. 2012, doi: 10.1007/s11999-011-2197-z.
- [46] A. I. Stoumpos, F. Kitsios, e M. A. Talias, «Digital Transformation in Healthcare: Technology Acceptance and Its Applications», *Int. J. Environ. Res. Public Health*, vol. 20, n.º 4, p. 3407, fev. 2023, doi: 10.3390/ijerph20043407.
- [47] «ObsCare — Sistema de registo clínico ginecológico e obstétrico», VirtualCare. [Em linha]. Disponível em: <http://virtualcare.pt/index.php/portfolio/vc-obscare/>
- [48] Alves Danielle Santos *et al.*, «Can openEHR Represent the Clinical Concepts of an Obstetric-Specific EHR — ObsCare Software?», em *Studies in Health Technology and Informatics*, IOS Press, 2019. doi: 10.3233/SHIT190328.
- [49] C. Dean, «How doctors can manage information overload», *Contemporary OBGYN Journal*, vol. 64, n.º 07, 11 de junho de 2019. [Em linha]. Disponível em: <https://www.medicaleconomics.com/view/how-doctors-can-manage-information-overload>
- [50] L. Sbaffi, J. Walton, J. Blenkinsopp, e G. Walton, «Information Overload in Emergency Medicine Physicians: A Multisite Case Study Exploring the Causes, Impact, and Solutions in Four North England National Health Service Trusts», *J. Med. Internet Res.*, vol. 22, n.º 7, p. e19126, jul. 2020, doi: 10.2196/19126.
- [51] A. Callahan, I. Pernek, G. Stiglic, J. Leskovec, H. R. Strasberg, e N. H. Shah, «Analyzing Information Seeking and Drug-Safety Alert Response by Health Care Professionals as New Methods for Surveillance», *J. Med. Internet Res.*, vol. 17, n.º 8, p. e204, ago. 2015, doi: 10.2196/jmir.4427.
- [52] M. S. Wong e A. Lewkowitz, «Artificial intelligence steps into obstetrics», *Contemporary OB/GYN Journal*, vol. 68, n.º 06, 9 de junho de 2023. [Em linha]. Disponível em: <https://www.contemporaryobgyn.net/view/artificial-intelligence-steps-into-obstetrics>
- [53] C. Chakraborty, S. Pal, M. Bhattacharya, S. Dash, e S.-S. Lee, «Overview of Chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science», *Front. Artif. Intell.*, vol. 6, p. 1237704, out. 2023, doi: 10.3389/frai.2023.1237704.
- [54] B. Huo *et al.*, «Large Language Models for Chatbot Health Advice Studies: A Systematic Review», *JAMA Netw. Open*, vol. 8, n.º 2, p. e2457879, fev. 2025, doi: 10.1001/jamanetworkopen.2024.57879.
- [55] A. W. K. Yeung *et al.*, «The promise of digital healthcare technologies», *Front. Public Health*, vol. 11, p. 1196596, set. 2023, doi: 10.3389/fpubh.2023.1196596.
- [56] D. J. Patel, K. Chaudhari, N. Acharya, D. Shrivastava, e S. Muneeba, «Artificial Intelligence in Obstetrics and Gynecology: Transforming Care and Outcomes», *Cureus*, jul. 2024, doi: 10.7759/cureus.64725.

- [57] Y. Lee e S. Y. Kim, «Potential applications of ChatGPT in obstetrics and gynecology in Korea: a review article», *Obstet. Gynecol. Sci.*, vol. 67, n.º 2, pp. 153–159, mar. 2024, doi: 10.5468/ogs.23231.
- [58] J. Liu, C. Wang, e S. Liu, «Utility of ChatGPT in Clinical Practice», *J. Med. Internet Res.*, vol. 25, p. e48568, jun. 2023, doi: 10.2196/48568.
- [59] L. Allahqoli, M. M. Ghiasvand, A. Mazidimoradi, H. Salehiniya, e I. Alkatout, «Diagnostic and Management Performance of ChatGPT in Obstetrics and Gynecology», *Gynecol. Obstet. Invest.*, vol. 88, n.º 5, pp. 310–313, 2023, doi: 10.1159/000533177.
- [60] A. Gilson *et al.*, «How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment», *JMIR Med. Educ.*, vol. 9, p. e45312, fev. 2023, doi: 10.2196/45312.
- [61] A. Cadiente, N. DaFonte, e J. D. Baum, «Performance of Artificial Intelligence Chatbots on Standardized Medical Examination Questions in Obstetrics & Gynecology», *Open J. Obstet. Gynecol.*, vol. 15, n.º 01, pp. 1–9, 2025, doi: 10.4236/ojog.2025.151001.
- [62] J. Hake *et al.*, «Quality, Accuracy, and Bias in ChatGPT-Based Summarization of Medical Abstracts», *Ann. Fam. Med.*, vol. 22, n.º 2, pp. 113–120, mar. 2024, doi: 10.1370/afm.3075.
- [63] Y. Chen e P. Esmailzadeh, «Generative AI in Medical Practice: In-Depth Exploration of Privacy and Security Challenges», *J. Med. Internet Res.*, vol. 26, p. e53008, mar. 2024, doi: 10.2196/53008.
- [64] S. J. Russell e P. Norvig, *Artificial Intelligence - A modern Approach*, 3.^a ed. Upper Saddle River, NJ: Prentice Hall, 2010.
- [65] A. Schuckart, «Introduction to work with GenAI», em *Proceedings of the 28th European Conference on Pattern Languages of Programs*, Irsee Germany: ACM, jul. 2023, pp. 1–16. doi: 10.1145/3628034.3628042.
- [66] A. Zubiaga, «Natural language processing in the era of large language models», *Front. Artif. Intell.*, vol. 6, p. 1350306, 2023, doi: 10.3389/frai.2023.1350306.
- [67] D. Khurana, A. Koli, K. Khatter, e S. Singh, «Natural language processing: state of the art, current trends and challenges», *Multimed. Tools Appl.*, vol. 82, n.º 3, pp. 3713–3744, jan. 2023, doi: 10.1007/s11042-022-13428-4.
- [68] D. Jurafsky e J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3.^a ed. Stanford University, 2026. [Em linha]. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/>
- [69] M. Zhou, N. Duan, S. Liu, e H.-Y. Shum, «Progress in Neural NLP: Modeling, Learning, and Reasoning», *Engineering*, vol. 6, n.º 3, pp. 275–290, mar. 2020, doi: 10.1016/j.eng.2019.12.014.
- [70] C. Dong *et al.*, «A Survey of Natural Language Generation», *ACM Comput. Surv.*, vol. 55, n.º 8, pp. 1–38, ago. 2023, doi: 10.1145/3554727.
- [71] S. J. Mielke *et al.*, «Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP», 2021, *arXiv*. doi: 10.48550/ARXIV.2112.10508.
- [72] V. Balakrishnan e L.-Y. Ethel, «Stemming and Lemmatization: A Comparison of Retrieval Performances», *Lect. Notes Softw. Eng.*, vol. 2, n.º 3, pp. 262–267, 2014, doi: 10.7763/LNSE.2014.V2.134.
- [73] D. Hussen Maulud, S. R. M. Zeebaree, K. Jacksi, M. A. Mohammed Sadeeq, e K. Hussein Sharif, «State of Art for Semantic Analysis of Natural Language Processing», *Qubahan Acad. J.*, vol. 1, n.º 2, pp. 21–28, mar. 2021, doi: 10.48161/qaj.v1n2a44.
- [74] J. Park, C. Park, J. Kim, M. Cho, e S. Park, «ADC: Advanced document clustering using contextualized representations», *Expert Syst. Appl.*, vol. 137, pp. 157–166, dez. 2019, doi: 10.1016/j.eswa.2019.06.068.
- [75] B. Mohit, «Named Entity Recognition», em *Natural Language Processing of Semitic Languages*, I. Zitouni, Ed., em *Theory and Applications of Natural Language Processing.*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 221–245. doi: 10.1007/978-3-642-45358-8_7.
- [76] X. Zhao *et al.*, «A Comprehensive Survey on Relation Extraction: Recent Advances and New Frontiers», *ACM Comput. Surv.*, vol. 56, n.º 11, pp. 1–39, nov. 2024, doi: 10.1145/3674501.
- [77] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, e A. Hassan, «Topic modeling algorithms and applications: A survey», *Inf. Syst.*, vol. 112, p. 102131, fev. 2023, doi: 10.1016/j.is.2022.102131.
- [78] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, e P. Kuksa, «Natural Language Processing (almost) from Scratch», 2011, *arXiv*. doi: 10.48550/ARXIV.1103.0398.

- [79] T. Young, D. Hazarika, S. Poria, e E. Cambria, «Recent Trends in Deep Learning Based Natural Language Processing [Review Article]», *IEEE Comput. Intell. Mag.*, vol. 13, n.º 3, pp. 55–75, ago. 2018, doi: 10.1109/MCI.2018.2840738.
- [80] K. Doshi, «Transformers Explained Visually (Part 3): Multi-head Attention, deep dive», Medium – TDS Archive (Towards Data Science). [Em linha]. Disponível em: <https://medium.com/data-science/transformers-explained-visually-part-3-multi-head-attention-deep-dive-1c1ff1024853>
- [81] C. Edwards, «The best of NLP», *Commun. ACM*, vol. 64, n.º 4, pp. 9–11, abr. 2021, doi: 10.1145/3449049.
- [82] H. Zhou *et al.*, «A Survey of Large Language Models in Medicine: Progress, Application, and Challenge», 22 de julho de 2024, *arXiv: arXiv:2311.05112*. doi: 10.48550/arXiv.2311.05112.
- [83] L. Fan, L. Li, Z. Ma, S. Lee, H. Yu, e L. Hemphill, «A Bibliometric Review of Large Language Models Research from 2017 to 2023», *ACM Trans. Intell. Syst. Technol.*, vol. 15, n.º 5, pp. 1–25, out. 2024, doi: 10.1145/3664930.
- [84] C. Peng *et al.*, «A study of generative large language model for medical research and healthcare», *Npj Digit. Med.*, vol. 6, n.º 1, p. 210, nov. 2023, doi: 10.1038/s41746-023-00958-w.
- [85] N. Karanikolas, E. Manga, N. Samaridi, E. Tousidou, e M. Vassilakopoulos, «Large Language Models versus Natural Language Understanding and Generation», em *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, Lamia Greece: ACM, nov. 2023, pp. 278–290. doi: 10.1145/3635059.3635104.
- [86] S. Yin *et al.*, «A Survey on Multimodal Large Language Models», *Natl. Sci. Rev.*, vol. 11, n.º 12, p. nwae403, nov. 2024, doi: 10.1093/nsr/nwae403.
- [87] A. Biswas e W. Talukdar, «Intelligent Clinical Documentation: Harnessing Generative AI for Patient-Centric Clinical Note Generation», *Int. J. Innov. Sci. Res. Technol. IJISRT*, pp. 994–1008, mai. 2024, doi: 10.38124/ijisrt/IJISRT24MAY1483.
- [88] F. Heilmeyer, D. Böhringer, T. Reinhard, S. Arens, L. Lyssenko, e C. Haverkamp, «Viability of Open Large Language Models for Clinical Documentation in German Health Care: Real-World Model Evaluation Study», *JMIR Med. Inform.*, vol. 12, pp. e59617–e59617, ago. 2024, doi: 10.2196/59617.
- [89] O. Shahab, B. El Kurdi, A. Shaukat, G. Nadkarni, e A. Soroush, «Large language models: a primer and gastroenterology applications», *Ther. Adv. Gastroenterol.*, vol. 17, p. 17562848241227031, jan. 2024, doi: 10.1177/17562848241227031.
- [90] J. Jiao, S. Afroogh, Y. Xu, e C. Phillips, «Navigating LLM Ethics: Advancements, Challenges, and Future Directions», 2024, *arXiv*. doi: 10.48550/ARXIV.2406.18841.
- [91] A. Singh, A. Ehtesham, G. K. Gupta, N. K. Chatta, S. Kumar, e T. T. Khoei, «Exploring Prompt Engineering: A Systematic Review with SWOT Analysis», 9 de outubro de 2024, *arXiv: arXiv:2410.12843*. doi: 10.48550/arXiv.2410.12843.
- [92] L. Tao *et al.*, «An outline of Prognostics and health management Large Model: Concepts, Paradigms, and challenges», *Mech. Syst. Signal Process.*, vol. 232, p. 112683, jun. 2025, doi: 10.1016/j.ymssp.2025.112683.
- [93] S. Vatsal e H. Dubey, «A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks», 24 de julho de 2024, *arXiv: arXiv:2407.12994*. doi: 10.48550/arXiv.2407.12994.
- [94] T. B. Brown *et al.*, «Language Models are Few-Shot Learners», 2020, *arXiv*. doi: 10.48550/ARXIV.2005.14165.
- [95] J. Wei *et al.*, «Chain-of-Thought Prompting Elicits Reasoning in Large Language Models», 2022, *arXiv*. doi: 10.48550/ARXIV.2201.11903.
- [96] J. Li, G. Li, Y. Li, e Z. Jin, «Structured Chain-of-Thought Prompting for Code Generation», *ACM Trans. Softw. Eng. Methodol.*, vol. 34, n.º 2, pp. 1–23, fev. 2025, doi: 10.1145/3690635.
- [97] S. AboulEla, P. Zabihitari, N. Ibrahim, M. Afshar, e R. Kashef, «Exploring RAG Solutions to Reduce Hallucinations in LLMs», em *2025 IEEE International systems Conference (SysCon)*, Montreal, QC, Canada: IEEE, abr. 2025, pp. 1–8. doi: 10.1109/syscon64521.2025.11014810.
- [98] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, e F. L. Wang, «Retrieval-augmented generation for educational application: A systematic survey», *Comput. Educ. Artif. Intell.*, vol. 8, p. 100417, jun. 2025, doi: 10.1016/j.caeai.2025.100417.
- [99] A. Kimothi, *A Simple Guide to Retrieval Augmented Generation*, 1st ed. New York: Manning Publications Co. LLC, 2025.

- [100] N. Zhang *et al.*, «SiReRAG: Indexing Similar and Related Information for Multihop Reasoning», 7 de abril de 2025, *arXiv*: arXiv:2412.06206. doi: 10.48550/arXiv.2412.06206.
- [101] Q. Zhang *et al.*, «Document Parsing Unveiled: Techniques, Challenges, and Prospects for Structured Information Extraction», 16 de abril de 2025, *arXiv*: arXiv:2410.21169. doi: 10.48550/arXiv.2410.21169.
- [102] S. Gao *et al.*, «Scaling Beyond Context: A Survey of Multimodal Retrieval-Augmented Generation for Document Understanding», 17 de outubro de 2025, *arXiv*: arXiv:2510.15253. doi: 10.48550/arXiv.2510.15253.
- [103] W. Ke *et al.*, «Large Language Models in Document Intelligence: A Comprehensive Survey, Recent Advances, Challenges and Future Trends», *ACM Trans. Inf. Syst.*, p. 3768156, set. 2025, doi: 10.1145/3768156.
- [104] Robert Arnaud, Damachi Francis, Bjelogrić Mina, Goldman Jean-Philippe, e Lovis Christian, «Evaluation of Document Retrieval Systems on a Medical Corpus in French: Indexation vs. Feature Learning», em *Studies in Health Technology and Informatics*, IOS Press, 2020. doi: 10.3233/SHIT200152.
- [105] K. Feldhoff, H. Wiemer, P. Träger, R. Kühne, M. Zimmermann, e S. Ihlenfeldt, «Automatic Information Extraction from Scientific Publications Based on the Use Case of Additive Manufacturing», *Appl. Sci.*, vol. 15, n.º 17, p. 9331, ago. 2025, doi: 10.3390/app15179331.
- [106] C. Yu, C. Zhang, e J. Wang, «Extracting Body Text from Academic PDF Documents for Text Mining», 23 de outubro de 2020, *arXiv*: arXiv:2010.12647. doi: 10.48550/arXiv.2010.12647.
- [107] G. L. Garcia, J. R. R. Manesco, P. H. Paiola, L. Miranda, M. P. de Salvo, e J. P. Papa, «A Review on Scientific Knowledge Extraction using Large Language Models in Biomedical Sciences», 4 de dezembro de 2024, *arXiv*: arXiv:2412.03531. doi: 10.48550/arXiv.2412.03531.
- [108] M. Salaheldin Kasem *et al.*, «Deep Learning for Table Detection and Structure Recognition: A Survey», *ACM Comput. Surv.*, vol. 56, n.º 12, pp. 1–41, dez. 2024, doi: 10.1145/3657281.
- [109] J. Li, W. Gu, K. Ota, e S. Hasegawa, «Object Recognition from Scientific Document based on Compartment Refinement Framework», *SN Comput. Sci.*, vol. 5, n.º 7, p. 816, ago. 2024, doi: 10.1007/s42979-024-03130-7.
- [110] L. D. Lopez, J. Yu, C. N. Arighi, H. Huang, H. Shatkay, e C. Wu, «An Automatic System for Extracting Figures and Captions in Biomedical PDF Documents», em *2011 IEEE International Conference on Bioinformatics and Biomedicine*, Atlanta, GA, USA: IEEE, nov. 2011, pp. 578–581. doi: 10.1109/BIBM.2011.26.
- [111] N. Deng *et al.*, «Tables as Texts or Images: Evaluating the Table Reasoning Ability of LLMs and MLLMs», 17 de outubro de 2024, *arXiv*: arXiv:2402.12424. doi: 10.48550/arXiv.2402.12424.
- [112] N. S. Adhikari e S. Agarwal, «A Comparative Study of PDF Parsing Tools Across Diverse Document Categories», 2024, *arXiv*. doi: 10.48550/ARXIV.2410.09871.
- [113] Y. Gao *et al.*, «Retrieval-Augmented Generation for Large Language Models: A Survey», 27 de março de 2024, *arXiv*: arXiv:2312.10997. doi: 10.48550/arXiv.2312.10997.
- [114] S. Wu *et al.*, «Retrieval-Augmented Generation for Natural Language Processing: A Survey», 1 de março de 2025, *arXiv*: arXiv:2407.13193. doi: 10.48550/arXiv.2407.13193.
- [115] N. Reimers e I. Gurevych, «Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks», 27 de agosto de 2019, *arXiv*: arXiv:1908.10084. doi: 10.48550/arXiv.1908.10084.
- [116] S. Xu *et al.*, «Reasoning before Comparison: LLM-Enhanced Semantic Similarity Metrics for Domain Specialized Text Analysis», 20 de fevereiro de 2024, *arXiv*: arXiv:2402.11398. doi: 10.48550/arXiv.2402.11398.
- [117] E. Oro, F. M. Granata, e M. Ruffolo, «A Comprehensive Evaluation of Embedding Models and LLMs for IR and QA Across English and Italian», *Big Data Cogn. Comput.*, vol. 9, n.º 5, p. 141, mai. 2025, doi: 10.3390/bdcc9050141.
- [118] R. Yacouby e D. Axman, «Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models», em *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, Online: Association for Computational Linguistics, 2020, pp. 79–91. doi: 10.18653/v1/2020.eval4nlp-1.9.
- [119] M. Amiri e T. Bocklitz, «Chunk Twice, Embed Once: A Systematic Study of Segmentation and Representation Trade-offs in Chemistry-Aware Retrieval-Augmented Generation», 13 de junho de 2025, *arXiv*: arXiv:2506.17277. doi: 10.48550/arXiv.2506.17277.

- [120] N. Pipitone e G. H. Alami, «LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain», 19 de agosto de 2024, *arXiv*: arXiv:2408.10343. doi: 10.48550/arXiv.2408.10343.
- [121] G. Capozzi, A. P. Giancaspro, F. Petroni, L. Querzoni, e G. A. Di Luna, «ReSIM: Re-ranking Binary Similarity Embeddings to Improve Function Search Performance», 2026, *arXiv*. doi: 10.48550/ARXIV.2602.09548.
- [122] A. Mudet e S. Bakkali, «Hybrid Retrieval-Augmented Generation for Robust Multilingual Document Question Answering», 2025, *arXiv*. doi: 10.48550/ARXIV.2512.12694.
- [123] P. Sujatha e K. Mahalakshmi, «Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart Disease», em *2020 IEEE International Conference for Innovation in Technology (INOCON)*, Bangluru, India: IEEE, nov. 2020, pp. 1–7. doi: 10.1109/INOCON50539.2020.9298354.
- [124] S. M. Shukla, C. Magoo, e P. Garg, «Comparing Fine Tuned-LMs for Detecting LLM-Generated Text», em *2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON)*, New Delhi, India: IEEE, nov. 2024, pp. 1–8. doi: 10.1109/DELCON64804.2024.10866497.
- [125] A. Jadon e A. Patil, «A Comprehensive Survey of Evaluation Techniques for Recommendation Systems», 12 de janeiro de 2024, *arXiv*: arXiv:2312.16015. doi: 10.48550/arXiv.2312.16015.
- [126] J. Gu *et al.*, «A Survey on LLM-as-a-Judge», 9 de março de 2025, *arXiv*: arXiv:2411.15594. doi: 10.48550/arXiv.2411.15594.
- [127] H. Li *et al.*, «LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods», 10 de dezembro de 2024, *arXiv*: arXiv:2412.05579. doi: 10.48550/arXiv.2412.05579.
- [128] B. Kitchenham, «Procedures for Performing Systematic Reviews», Keele University, Keele, Staffordshire, UK, TR/SE-0401, jul. 2004.
- [129] M. A. Hanson, P. G. Barreiro, P. Crosetto, e D. Brockington, «The strain on scientific publishing», *Quant. Sci. Stud.*, vol. 5, n.º 4, pp. 823–843, nov. 2024, doi: 10.1162/qss_a_00327.
- [130] I. J. Borges Do Nascimento *et al.*, «Barriers and facilitators to utilizing digital health technologies by healthcare professionals», *Npj Digit. Med.*, vol. 6, n.º 1, p. 161, set. 2023, doi: 10.1038/s41746-023-00899-4.
- [131] S. Rosonovski, M. Levchenko, M. Ide-Smith, L. Faulk, M. Harrison, e J. McEntyre, «Searching and Evaluating Publications and Preprints Using Europe PMC», *Curr. Protoc.*, vol. 3, n.º 3, p. e694, mar. 2023, doi: 10.1002/cpz1.694.
- [132] C. Lokker *et al.*, «The McMaster Health Information Research Unit: Over a Quarter-Century of Health Informatics Supporting Evidence-Based Medicine», *J. Med. Internet Res.*, vol. 26, p. e58764, jul. 2024, doi: 10.2196/58764.
- [133] J. Zeng *et al.*, «Natural Language Processing–Assisted Literature Retrieval and Analysis for Combination Therapy in Cancer», *JCO Clin. Cancer Inform.*, n.º 6, p. e2100109, mai. 2022, doi: 10.1200/CCI.21.00109.
- [134] Z. Zhang, X. Lin, e S. Wu, «A hybrid algorithm for clinical decision support in precision medicine based on machine learning», *BMC Bioinformatics*, vol. 24, n.º 1, jan. 2023, doi: 10.1186/s12859-022-05116-9.
- [135] A. Esteva *et al.*, «COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization», *Npj Digit. Med.*, vol. 4, n.º 1, p. 68, abr. 2021, doi: 10.1038/s41746-021-00437-0.
- [136] B. Škrlić, M. Jukić, N. Eržen, S. Pollak, e N. Lavrač, «Prioritization of COVID-19-related literature via unsupervised keyphrase extraction and document representation learning», 2021, doi: 10.48550/ARXIV.2110.08874.
- [137] M. Aziz *et al.*, «KnowVID-19: A Knowledge-Based System to Extract Targeted COVID-19 Information from Online Medical Repositories», *Biomolecules*, vol. 14, n.º 11, p. 1411, nov. 2024, doi: 10.3390/biom14111411.
- [138] R. Oruche *et al.*, «Evidence-Based Recommender System for a COVID-19 Publication Analytics Service», *IEEE Access*, vol. 9, pp. 79400–79415, 2021, doi: 10.1109/access.2021.3083583.
- [139] C. Zerva, S. Taylor, A. J. Soto, N. T. H. Nguyen, e S. Ananiadou, «A term-based and citation network-based search system for COVID-19», *JAMIA Open*, vol. 4, n.º 4, out. 2021, doi: 10.1093/jamiaopen/ooab104.

- [140] G. Spitale, F. Germani, e N. Biller-Andorno, «TopicTracker – An advanced software pipeline for text mining on PubMed data: Bridging the gap between off-the-shelf tools and code based approaches», *Heliyon*, vol. 10, n.º 17, p. e36351, set. 2024, doi: 10.1016/j.heliyon.2024.e36351.
- [141] S. Urru, V. Sciannameo, C. Lanera, S. Salaris, D. Gregori, e P. Berchialla, «A topic trend analysis on COVID-19 literature», *Digit. Health*, vol. 8, p. 205520762211336, jan. 2022, doi: 10.1177/20552076221133696.
- [142] A. Abd-alrazaq *et al.*, «Machine Learning-Based Approach for Identifying Research Gaps: COVID-19 as a Case Study», *JMIR Form. Res.*, vol. 8, 2024, doi: 10.2196/49411.
- [143] G. Cenikj, T. Eftimov, e B. Koroušić Seljak, «FooDis: A food-disease relation mining pipeline», *Artif. Intell. Med.*, vol. 142, p. 102586, ago. 2023, doi: 10.1016/j.artmed.2023.102586.
- [144] T. Abdullahi, L. Mercurio, R. Singh, e C. Eickhoff, «Retrieval-Based Diagnostic Decision Support: Mixed Methods Study», *JMIR Med. Inform.*, vol. 12, p. e50209, jun. 2024, doi: 10.2196/50209.
- [145] C. Bauer, R. Herwig, M. Lienhard, P. Prasse, T. Scheffer, e J. Schuchhardt, «Large-scale literature mining to assess the relation between anti-cancer drugs and cancer types», *J. Transl. Med.*, vol. 19, n.º 1, p. 274, dez. 2021, doi: 10.1186/s12967-021-02941-z.
- [146] H.-C. Hsueh *et al.*, «A novel Multi-Level Refined (MLR) knowledge graph design and chatbot system for healthcare applications», *PLOS ONE*, vol. 19, n.º 1, p. e0296939, jan. 2024, doi: 10.1371/journal.pone.0296939.
- [147] J. D. S. Franklin *et al.*, «Knowledge Extraction of Cohort Characteristics in Research Publications», *AMIA Annu. Symp. Proc. AMIA Symp.*, vol. 2020, pp. 462–471, 2020.
- [148] A. K. Lahiri e Q. V. Hu, «AlzheimerRAG: Multimodal Retrieval Augmented Generation for Clinical Use Cases using PubMed articles», 2024, *arXiv*. doi: 10.48550/ARXIV.2412.16701.
- [149] X. Chen *et al.*, «EyeGPT for Patient Inquiries and Medical Education: Development and Validation of an Ophthalmology Large Language Model», *J. Med. Internet Res.*, vol. 26, p. e60063, dez. 2024, doi: 10.2196/60063.
- [150] J. L. Z. Montenegro e C. A. Da Costa, «The HoPE Model Architecture: a Novel Approach to Pregnancy Information Retrieval Based on Conversational Agents», *J. Healthc. Inform. Res.*, vol. 6, n.º 3, pp. 253–294, set. 2022, doi: 10.1007/s41666-022-00115-0.
- [151] S. Amirian *et al.*, «Advancing psychosocial disability and psychosocial rehabilitation research through large language models and computational text mining», *Camb. Prisms Glob. Ment. Health*, vol. 11, p. e123, 2024, doi: 10.1017/gmh.2024.114.
- [152] Y. Yu *et al.*, «Using Large Language Models to Retrieve Critical Data from Clinical Processes and Business Rules», *Bioengineering*, vol. 12, n.º 1, p. 17, dez. 2024, doi: 10.3390/bioengineering12010017.
- [153] A. Gudivada, J. Philips, e N. Tabrizi, «Developing Concept Enriched Models for Big Data Processing Within the Medical Domain», *Int. J. Softw. Sci. Comput. Intell.*, vol. 12, n.º 3, pp. 55–71, dez. 2021, doi: 10.4018/IJSSCI.2020070105.
- [154] Z. Wang e A. Aryani, «Technical Report on classification of literature related to children speech disorder», 2025, *arXiv*. doi: 10.48550/ARXIV.2505.14242.
- [155] E. K. Lee e K. Uppal, «CERC: an interactive content extraction, recognition, and construction tool for clinical and biomedical text», *BMC Med. Inform. Decis. Mak.*, vol. 20, n.º S14, p. 306, dez. 2020, doi: 10.1186/s12911-020-01330-8.
- [156] W. Xu *et al.*, «MRAgent: an LLM-based automated agent for causal knowledge discovery in disease via Mendelian randomization», *Brief. Bioinform.*, vol. 26, n.º 2, mar. 2025, doi: 10.1093/bib/bbaf140.
- [157] F. Invernici, A. Bernasconi, e S. Ceri, «Exploring the evolution of research topics during the COVID-19 pandemic», *Expert Syst. Appl.*, vol. 252, p. 124028, out. 2024, doi: 10.1016/j.eswa.2024.124028.
- [158] S. Xu *et al.*, «Towards Next-Generation Medical Agent: How o1 is Reshaping Decision-Making in Medical Scenarios», 2024, *arXiv*. doi: 10.48550/ARXIV.2411.14461.
- [159] P. Bose, S. Roy, e P. Ghosh, «A Comparative NLP-Based Study on the Current Trends and Future Directions in COVID-19 Research», *IEEE Access*, vol. 9, pp. 78341–78355, 2021, doi: 10.1109/ACCESS.2021.3082108.
- [160] S. Robertson, V. Koppireddy, J. Cumbo, H. Rashidi, e S. Albahra, «PIRO: A web-based search platform for pathology reports, leveraging large language models to generate discrete searchable insights», *J. Pathol. Inform.*, vol. 17, p. 100436, abr. 2025, doi: 10.1016/j.jpi.2025.100436.

- [161] M. Rybinski, J. Xu, e S. Karimi, «Clinical trial search: Using biomedical language understanding models for re-ranking», *J. Biomed. Inform.*, vol. 109, p. 103530, set. 2020, doi: 10.1016/j.jbi.2020.103530.
- [162] J. Hao *et al.*, «Prompt Framework for Extracting Scale-Related Knowledge Entities from Chinese Medical Literature: Development and Evaluation Study», *J. Med. Internet Res.*, vol. 27, p. e67033, mar. 2025, doi: 10.2196/67033.
- [163] F. A. Causio, L. De Angelis, G. Diedenhofen, A. Talio, e F. Baglivo, «Perspectives on AI use in medicine: views of the Italian Society of Artificial Intelligence in Medicine», *J. Prev. Med. Hyg.*, p. Vol. 65 No. 2 (2024); ago. 2024, doi: 10.15167/2421-4248/JPMH2024.65.2.3261.
- [164] R. Srivastava, S. Prasad, L. Bhat, S. Deshpande, B. Das, e K. Jadhav, «MedPromptExtract (Medical Data Extraction Tool): Anonymization and Hi-fidelity Automated data extraction using NLP and prompt engineering», 6 de setembro de 2024, *arXiv*: arXiv:2405.02664. doi: 10.48550/arXiv.2405.02664.
- [165] V. Sciannameo *et al.*, «Information extraction from medical case reports using OpenAI InstructGPT», *Comput. Methods Programs Biomed.*, vol. 255, p. 108326, out. 2024, doi: 10.1016/j.cmpb.2024.108326.
- [166] S. Tian *et al.*, «Opportunities and challenges for ChatGPT and large language models in biomedicine and health», *Brief. Bioinform.*, vol. 25, n.º 1, p. bbad493, nov. 2023, doi: 10.1093/bib/bbad493.
- [167] E. Sezgin e A. B. Kocaballi, «Era of Generalist Conversational Artificial Intelligence to Support Public Health Communications», *J. Med. Internet Res.*, vol. 27, p. e69007, jan. 2025, doi: 10.2196/69007.
- [168] D. J. Parente, «Generative Artificial Intelligence and Large Language Models in Primary Care Medical Education», *Fam. Med.*, vol. 56, n.º 9, pp. 534–540, out. 2024, doi: 10.22454/FamMed.2024.775525.
- [169] Y. Hu, Y. Chen, e H. Xu, «Towards More Generalizable and Accurate Sentence Classification in Medical Abstracts with Less Data», *J. Healthc. Inform. Res.*, vol. 7, n.º 4, pp. 542–556, dez. 2023, doi: 10.1007/s41666-023-00141-6.
- [170] LangChain, «LangChain — The platform for reliable agents. Tools for every step of the agent development lifecycle — built to unlock powerful AI in production.», LangChain. [Em linha]. Disponível em: <https://www.langchain.com/>
- [171] LangChain Contributors, «OllamaLLM — LangChain Documentation: Integrations · LLMs · Ollama», LangChain Documentation. [Em linha]. Disponível em: <https://python.langchain.com/docs/integrations/llms/ollama/>
- [172] LangChain Contributors, «Chroma — LangChain Documentation: Integrations · Vectorstores · Chroma», LangChain Documentation. [Em linha]. Disponível em: <https://python.langchain.com/docs/integrations/vectorstores/chroma/>
- [173] «Hugging Face — The AI community building the future», Hugging Face. [Em linha]. Disponível em: <https://huggingface.co/>
- [174] Streamlit, «Streamlit — A faster way to build and share data apps», Streamlit. [Em linha]. Disponível em: <https://streamlit.io/>
- [175] Sentence Transformers, «Pretrained Models», SentenceTransformers Documentation. [Em linha]. Disponível em: https://www.sbert.net/docs/sentence_transformer/pretrained_models.html
- [176] M. T. Colangelo, M. Meleti, S. Guizzardi, E. Calciolari, e C. Galli, «A Comparative Analysis of Sentence Transformer Models for Automated Journal Recommendation Using PubMed Metadata», *Big Data Cogn. Comput.*, vol. 9, n.º 3, p. 67, mar. 2025, doi: 10.3390/bdcc9030067.
- [177] D. Mezzetti, «Embeddings for Medical Literature», Medium. [Em linha]. Disponível em: <https://medium.com/neuml/embeddings-for-medical-literature-74dae6abf5e0>
- [178] A. Balachandran, «MedEmbed: Fine-Tuned Embedding Models for Medical / Clinical IR», Hugging Face. [Em linha]. Disponível em: <https://huggingface.co/blog/abhinand/medembed-finetuned-embedding-models-for-medical-ir>
- [179] Groq, «GroqCloud». [Em linha]. Disponível em: <https://groq.com/>
- [180] Cloudflare, «Cloudflare Workers AI Documentation».
- [181] K. Team *et al.*, «Kimi K2: Open Agentic Intelligence», 3 de fevereiro de 2026, *arXiv*: arXiv:2507.20534. doi: 10.48550/arXiv.2507.20534.
- [182] S. Bai *et al.*, «Qwen3-VL Technical Report», 27 de novembro de 2025, *arXiv*: arXiv:2511.21631. doi: 10.48550/arXiv.2511.21631.

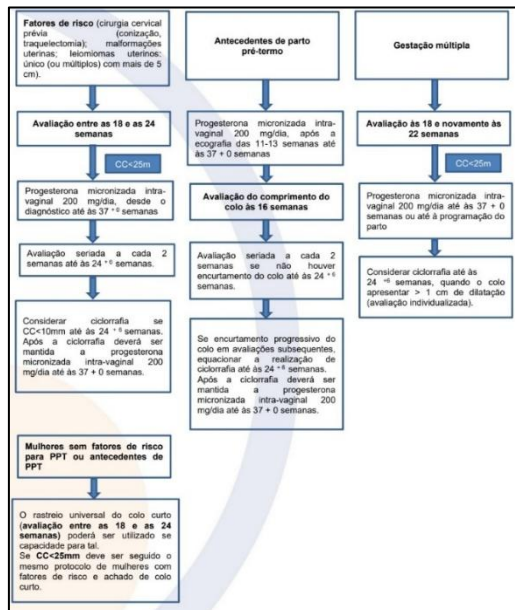
- [183] O. Normand *et al.*, «A Real-World Evaluation of LLM Medication Safety Reviews in NHS Primary Care», 2025, *arXiv*. doi: 10.48550/ARXIV.2512.21127.
- [184] A. Grattafiori *et al.*, «The Llama 3 Herd of Models», 23 de novembro de 2024, *arXiv*: arXiv:2407.21783. doi: 10.48550/arXiv.2407.21783.
- [185] F. Cuconasu *et al.*, «The Power of Noise: Redefining Retrieval for RAG Systems», em *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, jul. 2024, pp. 719–729. doi: 10.1145/3626772.3657834.
- [186] «How to Extract Table Content from Documents — PyMuPDF Documentation», PyMuPDF documentation. [Em linha]. Disponível em: <https://pymupdf.readthedocs.io/en/latest/recipes-text.html#how-to-extract-table-content-from-documents>
- [187] «pdfplumber 0.11.7», PyPI. [Em linha]. Disponível em: <https://pypi.org/project/pdfplumber/>
- [188] S. Prakash, «Building a Table Extractor with Microsoft Table Transformer», Medium. [Em linha]. Disponível em: <https://medium.com/@shwet.prakash97/building-a-table-extractor-with-microsoft-table-transformer-baf0c953a59d>
- [189] PaddlePaddle contributors, «PaddleOCR - KIE Module Visualization», GitHub. [Em linha]. Disponível em: <https://github.com/PaddlePaddle/PaddleOCR/blob/main/ppstructure/kie/README.md#3-visualization>
- [190] L. Blecher, G. Cucurull, T. Scialom, e R. Stojnic, «Nougat: Neural Optical Understanding for Academic Documents», 25 de agosto de 2023, *arXiv*: arXiv:2308.13418. doi: 10.48550/arXiv.2308.13418.
- [191] G. Avesani, M. Bonatti, A. M. Venkatesan, S. Nougaret, e E. Sala, «RadioGraphics Update: 2023 FIGO Staging System for Endometrial Cancer», *RadioGraphics*, vol. 44, n.º 7, p. e240084, jul. 2024, doi: 10.1148/rg.240084.
- [192] A. L. Areia, A. Rocha, E. Nunes, E. Pereira, M. de C. Afonso, e M. J. Monteiro, *Norma de Orientação Clínica SPOMMF: Rastreio de Colo Curto*. [Em linha]. Disponível em: <https://www.spommf.pt/norma-de-orientacao-clinica-spommf-rastreio-de-colo-curto/>
- [193] L.-L. Su, Y.-S. Chong, e A. Biswas, «Use of fetal electrocardiogram for intrapartum monitoring», *Ann. Acad. Med. Singapore*, vol. 36, n.º 6, pp. 416–420, jun. 2007.
- [194] F. Henrique M. George, *Diagnóstico e conduta na Diabetes Gestacional*, 007/2011, 31 de janeiro de 2011. [Em linha]. Disponível em: <https://nocs.pt/diabetes-gestacional/>
- [195] T. M. Ferraz *et al.*, *Normas de Orientação Clínica SPOMMF — Gravidez após Cirurgia Bariátrica*, 2 de junho de 2020. [Em linha]. Disponível em: <https://www.spommf.pt/normas-spommf-gravidez-apos-cirurgia-bariatrica/>
- [196] A. Fonseca, C. Nogueira-Silva, I. S. Silva, E. Casal, L. Pinto, e N. Clode, *Norma de Orientação Clínica SPOMMF: Parto Pélvico Vaginal*. [Em linha]. Disponível em: <https://www.spommf.pt/norma-de-orientacao-clinica-spommf-parto-pelvico-vaginal/>
- [197] M. Centeno *et al.*, *Norma de Orientação Clínica SPOMMF: Doença Tiroideia na Gravidez*, 11 de janeiro de 2022. [Em linha]. Disponível em: <https://www.spommf.pt/norma-de-orientacao-clinica-spommf-doenca-tiroideia-na-gravidez/>
- [198] M. Zaigham, F. Louwen, S. Mukhopadhyay, e T. Mahmood, «Active labour management and multiprofessional teamwork are cornerstones in preventing postpartum haemorrhage: EBCOG commentary on “The new world health Organization’s (WHO) postpartum haemorrhage (PPH) guideline 2025”», *Eur. J. Obstet. Gynecol. Reprod. Biol.*, vol. 316, p. 114798, jan. 2026, doi: 10.1016/j.ejogrb.2025.114798.
- [199] A. P. Machado, T. Castro, M. Afonso, S. Tavares, I. Ferreira, e Santo, *Norma de Orientação Clínica SPOMMF: Maturação Cervical*, 6 de dezembro de 2022. [Em linha]. Disponível em: <https://www.spommf.pt/norma-de-orientacao-clinica-spommf-maturacao-cervical/>
- [200] Y. Yang e N. Wu, «Gestational Diabetes Mellitus and Preeclampsia: Correlation and Influencing Factors», *Front. Cardiovasc. Med.*, vol. 9, p. 831297, 2022, doi: 10.3389/fcvm.2022.831297.
- [201] J. F. Plows, J. L. Stanley, P. N. Baker, C. M. Reynolds, e M. H. Vickers, «The Pathophysiology of Gestational Diabetes Mellitus», *Int. J. Mol. Sci.*, vol. 19, n.º 11, p. 3342, out. 2018, doi: 10.3390/ijms19113342.
- [202] D. D. Sharma *et al.*, «The Management of Preeclampsia: A Comprehensive Review of Current Practices and Future Directions», *Cureus*, vol. 16, n.º 1, p. e51512, jan. 2024, doi: 10.7759/cureus.51512.

- [203] E. Shebl e R. K. Chakraborty, «Asthma in Pregnancy», em *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2026. Acedido: 8 de março de 2026. [Em linha]. Disponível em: <http://www.ncbi.nlm.nih.gov/books/NBK532283/>
- [204] H. Wang, N. Li, e H. Huang, «Asthma in Pregnancy: Pathophysiology, Diagnosis, Whole-Course Management, and Medication Safety», *Can. Respir. J.*, vol. 2020, p. 9046842, 2020, doi: 10.1155/2020/9046842.
- [205] K. A. Moyle, «A practical review of iron deficiency in pregnancy», *Semin. Fetal. Neonatal Med.*, vol. 30, n.º 1, p. 101611, mar. 2025, doi: 10.1016/j.siny.2025.101611.
- [206] G. J. Kim, «Reviving external cephalic version: a review of its efficacy, safety, and technical aspects», *Obstet. Gynecol. Sci.*, vol. 62, n.º 6, pp. 371–381, nov. 2019, doi: 10.5468/ogs.2019.62.6.371.
- [207] «Normas de Orientação Clínica SPOMMF», Sociedade Portuguesa de Obstetrícia e Medicina Materno-Fetal (SPOMMF). [Em linha]. Disponível em: <https://www.spommf.pt/category/normas-de-orientacao-clinica-spommf/>
- [208] Direção-Geral da Saúde, «Saúde Materna», Normas de Orientação Clínica. [Em linha]. Disponível em: <https://nocs.pt/saude-materna/>
- [209] M. Garbi, «National Institute for Health and Care Excellence clinical guidelines development principles and processes», *Heart*, vol. 107, n.º 12, pp. 949–953, jun. 2021, doi: 10.1136/heartjnl-2020-318661.
- [210] National Institute for Health and Care Excellence, «The guidelines manual», National Institute for Health and Care Excellence. [Em linha]. Disponível em: <https://www.nice.org.uk/process/pmg6>
- [211] International Federation of Gynecology and Obstetrics, «About us», FIGO. [Em linha]. Disponível em: <https://www.figo.org/>
- [212] G. I. Serour, «Global Work of the International Federation of Gynecology and Obstetrics (FIGO)», *Facts, Views & Vision in ObGyn*, vol. 4, n.º 1, pp. 3–8, 2012.
- [213] J.-C. Lee *et al.*, «Efficacy improvement in searching MEDLINE database using a novel PubMed visual analytic system: EEVis», *PLOS ONE*, vol. 18, n.º 2, p. e0281422, fev. 2023, doi: 10.1371/journal.pone.0281422.
- [214] M. H. Murad, N. Asi, M. Alsawas, e F. Alahdab, «New evidence pyramid», *Evid. Based Med.*, vol. 21, n.º 4, pp. 125–127, ago. 2016, doi: 10.1136/ebmed-2016-110401.
- [215] A. Al Noman, O. Sarkar, T. M. Mita, K. Siddika, e F. Afrose, «Simplifying the concept of level of evidence in lay language for all aspects of learners: In brief review», *Intell. Pharm.*, vol. 2, n.º 2, pp. 270–273, abr. 2024, doi: 10.1016/j.ipha.2023.11.002.
- [216] Aadit Kshirsagar, «Enhancing RAG Performance Through Chunking and Text Splitting Techniques», *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 10, n.º 5, pp. 151–158, nov. 2024, doi: 10.32628/CSEIT2410593.
- [217] S. R. Bhat, M. Rudat, J. Spiekermann, e N. Flores-Herr, «Rethinking Chunk Size For Long-Document Retrieval: A Multi-Dataset Analysis», 2025, *arXiv*. doi: 10.48550/ARXIV.2505.21700.
- [218] S. Wang, S. Zhuang, e G. Zucco, «BERT-based Dense Retrievers Require Interpolation with BM25 for Effective Passage Retrieval», em *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, Virtual Event Canada: ACM, jul. 2021, pp. 317–324. doi: 10.1145/3471158.3472233.
- [219] Z. Liu, C. Li, S. Xiao, Y. Shao, e D. Lian, «Llama2Vec: Unsupervised Adaptation of Large Language Models for Dense Retrieval», 23 de novembro de 2025, *arXiv*: arXiv:2312.15503. doi: 10.48550/arXiv.2312.15503.
- [220] O. Strauß e H. Kett, «Documents as Intelligent Agents: An Approach to Optimize Document Representations in Semantic Search», em *Proceedings of the 19th International Conference on Web Information Systems and Technologies*, Rome, Italy: SCITEPRESS - Science and Technology Publications, 2023, pp. 164–175. doi: 10.5220/0012239200003584.
- [221] J. De Backer *et al.*, «2025 ESC Guidelines for the management of cardiovascular disease and pregnancy», *Eur. Heart J.*, vol. 46, n.º 43, pp. 4462–4568, nov. 2025, doi: 10.1093/eurheartj/ehaf193.
- [222] L. C. Poon *et al.*, «The International Federation of Gynecology and Obstetrics (FIGO) initiative on pre-eclampsia: A pragmatic guide for first-trimester screening and prevention», *Int. J. Gynecol. Obstet.*, vol. 145, n.º S1, pp. 1–33, mai. 2019, doi: 10.1002/ijgo.12802.

- [223] M. Hod *et al.*, «The International Federation of Gynecology and Obstetrics (FIGO) Initiative on gestational diabetes mellitus: A pragmatic guide for diagnosis, management, and care[#]», *Int. J. Gynecol. Obstet.*, vol. 131, n.º S3, out. 2015, doi: 10.1016/S0020-7292(15)30033-3.
- [224] Y. Hu *et al.*, «Triglyceride-glucose body mass index and the risk of gestational diabetes mellitus: a population-based cohort study», *Front. Nutr.*, vol. 12, p. 1657186, dez. 2025, doi: 10.3389/fnut.2025.1657186.

Anexo A: Resultados da descrição do Conteúdo de fluxogramas

Figura A1 - Exemplo de uma figura de um fluxograma. Retirado do documento: [192]



Resposta do modelo:

Path 1: Fatores de risco (cirurgia cervical prévia (conização, traquelectomia); malformações uterinas; leiomiomas uterinos: único (ou múltiplos) com mais de 5 cm) → Avaliação entre as 18 e as 24 semanas (CC<25m) → Progesterona micronizada intra-vaginal 200 mg/dia, desde o diagnóstico até às 37 + 0 semanas

Path 2: Fatores de risco (cirurgia cervical prévia (conização, traquelectomia); malformações uterinas; leiomiomas uterinos: único (ou múltiplos) com mais de 5 cm) →

Avaliação entre as 18 e as 24 semanas (CC<25m) → Avaliação seriada a cada 2 semanas até às 24 + 6 semanas → Considerar ciclorrafia se CC<10mm até às 24 + 6 semanas. Após a ciclorrafia deverá ser mantida a progesterona micronizada intra-vaginal 200 mg/dia até às 37 + 0 semanas.

Path 3: Fatores de risco (cirurgia cervical prévia (conização, traquelectomia); malformações uterinas; leiomiomas uterinos: único (ou múltiplos) com mais de 5 cm) → Avaliação entre as 18 e as 24 semanas (CC<25m) → Avaliação seriada a cada 2 semanas até às 24 + 6 semanas → Se encurtamento progressivo do colo em avaliações subsequentes, equacionar a realização de ciclorrafia até às 24 + 6 semanas. Após a ciclorrafia deverá ser mantida a progesterona micronizada intra-vaginal 200 mg/dia até às 37 + 0 semanas.

Path 4: Antecedentes de parto pré-termo → Progesterona micronizada intra-vaginal 200 mg/dia, após a ecografia das 11-13 semanas até às 37 + 0 semanas → Avaliação do comprimento do colo às 16 semanas → Avaliação seriada a cada 2 semanas se não houver encurtamento do colo até às 24 + 6 semanas.

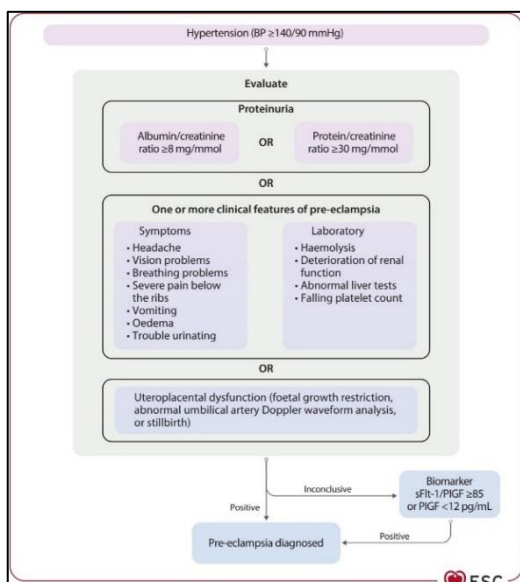
Path 5: Antecedentes de parto pré-termo → Progesterona micronizada intra-vaginal 200 mg/dia, após a ecografia das 11-13 semanas até às 37 + 0 semanas → Avaliação do comprimento do colo às 16 semanas → Se encurtamento progressivo do colo em avaliações subsequentes, equacionar a realização de ciclorrafia até às 24 +6 semanas. Após a ciclorrafia deverá ser mantida a progesterona micronizada intra-vaginal 200 mg/dia até às 37 + 0 semanas.

Path 6: Gestação múltipla → Avaliação às 18 e novamente às 22 semanas (CC<25m) → Progesterona micronizada intra-vaginal 200 mg/dia até às 37 + 0 semanas ou até à programação do parto.

Path 7: Gestação múltipla → Avaliação às 18 e novamente às 22 semanas (CC<25m) → Considerar ciclorrafia até às 24 +6 semanas, quando o colo apresentar > 1 cm de dilatação (avaliação individualizada).

Path 8: Mulheres sem fatores de risco para PPT ou antecedentes de PPT → O rastreio universal do colo curto (avaliação entre as 18 e as 24 semanas) poderá ser utilizado se capacidade para tal. Se CC<25mm deve ser seguido o mesmo protocolo de mulheres com fatores de risco e achado de colo curto.

Figura A2: Exemplo de imagem de fluxograma. Retirado de: [221]



Resultados do modelo:

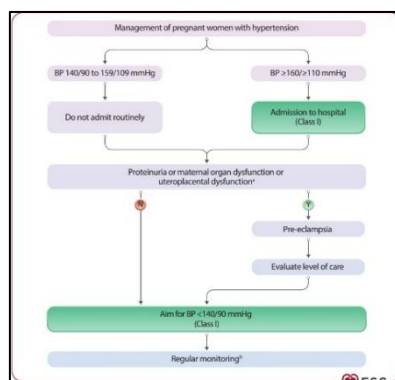
Path 1: Hypertension (BP ≥140/90 mmHg) → Evaluate (Proteinuria (Albumin/creatinine ratio ≥8 mg/mmol OR Protein/creatinine ratio ≥30 mg/mmol) OR One or more clinical features of pre-eclampsia (Symptoms (Headache, Vision problems, Breathing problems, Severe pain below the ribs, Vomiting, Oedema, Trouble urinating) OR Laboratory (Haemolysis, Deterioration of renal function, Abnormal liver tests, Falling platelet count)) → Uteroplacental dysfunction (foetal growth

restriction, abnormal umbilical artery Doppler waveform analysis, or stillbirth) → Pre-eclampsia diagnosed

Path 2: Hypertension (BP ≥140/90 mmHg) → Evaluate (Proteinuria (Albumin/creatinine ratio ≥8 mg/mmol OR Protein/creatinine ratio ≥30

mg/mmol) OR One or more clinical features of pre-eclampsia (Symptoms (Headache, Vision problems, Breathing problems, Severe pain below the ribs, Vomiting, Oedema, Trouble urinating) OR Laboratory (Haemolysis, Deterioration of renal function, Abnormal liver tests, Falling platelet count)) → Uteroplacental dysfunction (foetal growth restriction, abnormal umbilical artery Doppler waveform analysis, or stillbirth) → Biomarker sFlt-1/PlGF ≥ 85 or PlGF < 12 pg/mL → Pre-eclampsia diagnosed

Figura A3: Exemplo de imagem de fluxograma. Retirado de: [221]



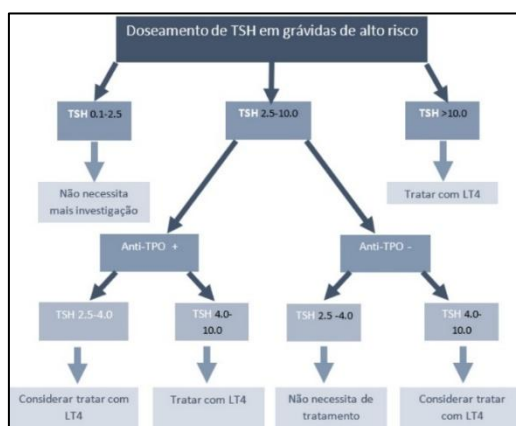
Resultados do modelo:

Path 1: Management of pregnant women with hypertension () → BP 140/90 to 159/109 mmHg () → Do not admit routinely () → Proteinuria or maternal organ dysfunction or uteroplacental dysfunction* () –[N]→ Aim for BP $< 140/90$ mmHg (Class I) () → Regular monitoring ()

Path 2: Management of pregnant women with hypertension () → BP 140/90 to 159/109 mmHg () → Do not admit routinely () → Proteinuria or maternal organ dysfunction or uteroplacental dysfunction* () –[Y]→ Pre-eclampsia () → Evaluate level of care () → Aim for BP $< 140/90$ mmHg (Class I) () → Regular monitoring ()

Path 3: Management of pregnant women with hypertension () → BP $\geq 160/\geq 110$ mmHg () → Admission to hospital (Class I) () → Aim for BP $< 140/90$ mmHg (Class I) () → Regular monitoring ()

Figura A4: Exemplo de imagem de fluxograma. Retirado de: [197]



Resultados do modelo:

Path 1: Doseamento de TSH em grávidas de alto risco (Doseamento de TSH em grávidas de alto risco) → TSH 0.1-2.5 (TSH 0.1-2.5) → Não necessita mais investigação (Não necessita mais investigação)

Path 2: Doseamento de TSH em grávidas de alto risco (Doseamento de TSH em grávidas de alto risco) → TSH 2.5-

10.0 (TSH 2.5-10.0) → Anti-TPO + (Anti-TPO +) → TSH 2.5-4.0 (TSH 2.5-4.0) → Considerar tratar com LT4 (Considerar tratar com LT4)

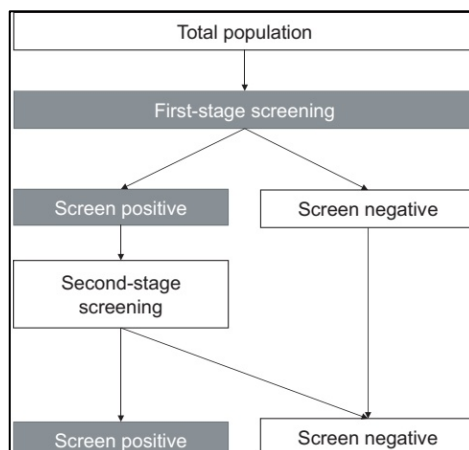
Path 3: Doseamento de TSH em grávidas de alto risco (Doseamento de TSH em grávidas de alto risco) → TSH 2.5-10.0 (TSH 2.5-10.0) → Anti-TPO + (Anti-TPO +) → TSH 4.0-10.0 (TSH 4.0-10.0) → Tratar com LT4 (Tratar com LT4)

Path 4: Doseamento de TSH em grávidas de alto risco (Doseamento de TSH em grávidas de alto risco) → TSH 2.5-10.0 (TSH 2.5-10.0) → Anti-TPO - (Anti-TPO -) → TSH 2.5 -4.0 (TSH 2.5 -4.0) → Não necessita de tratamento (Não necessita de tratamento)

Path 5: Doseamento de TSH em grávidas de alto risco (Doseamento de TSH em grávidas de alto risco) → TSH 2.5-10.0 (TSH 2.5-10.0) → Anti-TPO - (Anti-TPO -) → TSH 4.0-10.0 (TSH 4.0-10.0) → Considerar tratar com LT4 (Considerar tratar com LT4)

Path 6: Doseamento de TSH em grávidas de alto risco (Doseamento de TSH em grávidas de alto risco) → TSH >10.0 (TSH >10.0) → Tratar com LT4 (Tratar com LT4)

Figura A5: Exemplo de imagem de fluxograma. Retirado de: [222]



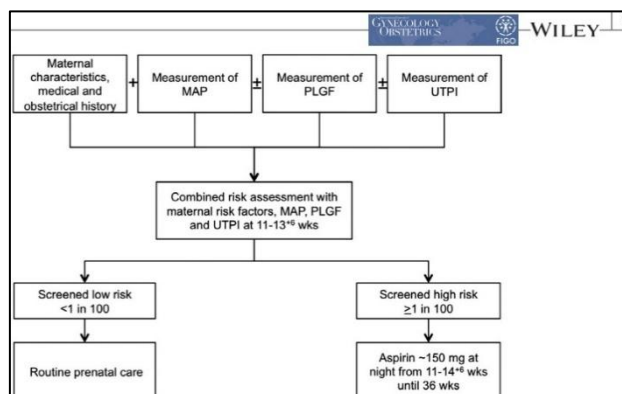
Resultados do modelo:

Path 1: Total population → First-stage screening → Screen positive → Second-stage screening → Screen positive

Path 2: Total population → First-stage screening → Screen positive → Second-stage screening → Screen negative

Path 3: Total population → First-stage screening → Screen negative

Figura A6: Exemplo de imagem de fluxograma. Retirado de: [222]

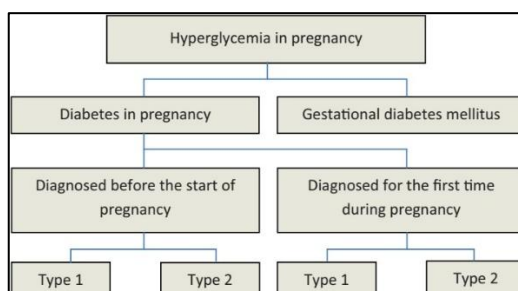


Resultados do modelo:

Path 1: Maternal characteristics, medical and obstetrical history () → Combined risk assessment with maternal risk factors, MAP, PLGF and UTPI at 11-13+6 wks () → Screened low risk <1 in 100 () → Routine prenatal care ()

Path 2: Maternal characteristics, medical and obstetrical history () → Combined risk assessment with maternal risk factors, MAP, PLGF and UTPI at 11-13+6 wks () → Screened high risk ≥1 in 100 () → Aspirin ~150 mg at night from 11-14+6 wks until 36 wks ()

Figura A7: Exemplo de imagem de fluxograma. Retirado de: [223]



Resultados:

Path 1: Hyperglycemia in pregnancy () → Diabetes in pregnancy () → Diagnosed before the start of pregnancy () → Type 1 ()

Path 2: Hyperglycemia in pregnancy ()

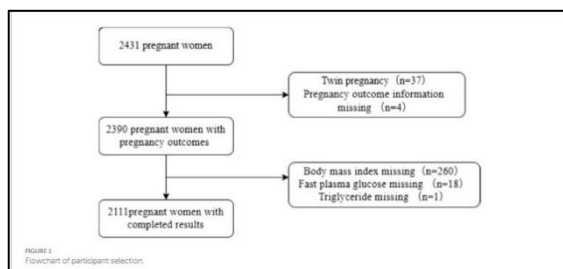
→ Diabetes in pregnancy () → Diagnosed before the start of pregnancy () → Type 2 ()

Path 3: Hyperglycemia in pregnancy () → Diabetes in pregnancy () → Diagnosed for the first time during pregnancy () → Type 1 ()

Path 4: Hyperglycemia in pregnancy () → Diabetes in pregnancy () → Diagnosed for the first time during pregnancy () → Type 2 ()

Path 5: Hyperglycemia in pregnancy () → Gestational diabetes mellitus ()

Figura A8: Exemplo de imagem de fluxograma. Retirado de: [224]



Resultado do modelo:

Path 1: 2431 pregnant women → 2390 pregnant women with pregnancy outcomes → 2111 pregnant women with completed results

Path 2: 2431 pregnant women → 2390 pregnant women with pregnancy outcomes → Body mass index missing (n=260) Fast plasma glucose missing (n=18) Triglyceride missing (n=1)

Path 3: 2431 pregnant women → Twin pregnancy (n=37) Pregnancy outcome information missing (n=4)

Anexos B: Resultados da comparação das duas estratégias de descrição de figuras

Tabela B1: Resultados das estratégias 1 e 2 para a descrição de figuras com fluxogramas

Estratégia 1: Análise com Modelo Textual, contendo referencias do artigo e legenda
“Unfortunately, I don't see any excerpts from an obstetric guideline provided. However, based on the caption "Quadro final com o resumo das orientações: Rastreo do colo curto" which translates to "Final chart with a summary of guidelines: Short cervix surveillance", I can make an educated guess. The figure is likely showing a flowchart or a decision tree that outlines the steps for monitoring and managing women with a short cervix during pregnancy. The caption suggests that it's a final chart summarizing the guidelines for surveillance of short cervix, which implies that it may include information on: * Who should be screened for short cervix * How often to monitor cervical length * What actions to take if there are concerns about cervical length * When to intervene or refer to a specialist The figure is probably a visual representation of the guidelines, making it easier for healthcare providers to follow and understand the recommended approach.”
Estratégia 2: Análise com Modelo Multimodal
“The provided image presents a flowchart in Portuguese, detailing the management and assessment of pregnant women at risk for preterm birth (PTB) or with a history of previous preterm birth. The chart outlines a series of steps and interventions based on various risk factors and cervical length measurements. Initial Assessment and Risk Factors: The chart begins by identifying risk factors for preterm birth, including: Previous cervical surgery (e.g., conization, trachelectomy); Uterine malformations; Uterine leiomyomas > 5 cm; History of previous preterm birth; Multiple gestation Assessment and Management: Between 18-24 weeks: Assess cervical length (CC) If CC < 25mm, initiate progesterone micronized intra-vaginal 200 mg/day until 37 weeks: Serial assessments every 2 weeks until 24 weeks and 6 days If CC < 10mm until 24 weeks and 6 days: Consider cerclage; Continue progesterone micronized intra-vaginal 200 mg/day until 37 weeks If progressive cervical shortening: Consider cerclage until 24 weeks and 6 days; Continue progesterone micronized intra-vaginal 200 mg/day until 37 weeks Additional Considerations: Women with a history of previous preterm birth or risk factors: Assess cervical length at 16 weeks; If CC < 25mm, initiate progesterone micronized intra-vaginal 200 mg/day until 37 weeks;

Multiple gestation: Assess cervical length at 18 and 22 weeks; If CC < 25mm, initiate progesterone micronized intra-vaginal 200 mg/day until 37 weeks or until planned delivery

Summary: The flowchart provides a structured approach to managing pregnant women at risk for preterm birth, emphasizing the importance of cervical length assessment and progesterone supplementation. The chart highlights the need for serial assessments and individualized management based on cervical length measurements and risk factors.”

Tabela B2: Resultados das estratégias 1 e 2 para a descrição de figuras com imagem de manobra obstétrica

Resultados para a Figura 22 (a)
Estratégia 1: Análise com Modelo Textual, contendo referencias do artigo e legenda
Figure 1 likely shows a schematic representation of the three breech presentation modalities during labor, with simplified illustrations of the different positions of the fetal lower limbs in relation to the birth canal. More specifically, the figure should show: - An image of the complete presentation (fetus with lower limbs fully flexed), where the fetal pelvis and feet are fully visible in the birth canal. - A second image representing the incomplete breech presentation (hips flexed and knees extended), showing only the pelvis in relation to the birth canal, without the feet or lower limbs. - A third image of the incomplete feet presentation (at least one lower limb fully extended), where one or both feet are visible in the birth canal. This schematic representation aims to help illustrate the different ways in which the fetus can present during labor, facilitating the understanding of the classification of breech presentations.
Estratégia 2: Análise com Modelo Multimodal
The provided image illustrates three different fetal positions during labor, labeled A, B, and C. Image A depicts a fetus in a brow presentation, where the head is flexed and the chin is almost touching the chest. However, in this image, it appears to be more of a face presentation but with the chin towards the chest, indicating a possible brow presentation or a face presentation with a mentum posterior (chin towards the mother's back) position. Image B shows a footling breech presentation, where one leg is extended and the other is flexed, with the feet presenting first. Image C illustrates a hand or arm prolapse in a cephalic presentation; however, it seems to depict a shoulder and arm presentation where the arm is prolapsed alongside the head. These images likely represent potential complications during delivery, such as malpresentations or malpositions that may require medical intervention. A thorough assessment by a healthcare provider would be necessary to determine the best course of action for a safe delivery.
Resultados para a Figura 22 (b)
Estratégia 1: Análise com Modelo Textual, contendo referencias do artigo e legenda
Figure 2 likely depicts the lower limbs of a fetus being exteriorized, with the fetus's thigh being pushed upwards and outwards via the Pinard maneuver.

Estratégia 2: Análise com Modelo Multimodal

The image provided appears to illustrate a medical procedure related to obstetrics, specifically a type of delivery or fetal manipulation. Given the context of obstetrics and maternal-fetal medicine, the image seems to depict a breech delivery or more specifically, a procedure known as a "breech extraction" or a maneuver to assist in the delivery of a baby.

Description of the Image:

The image is divided into two parts, each showing a different step or perspective of the procedure:

Left Side: This part of the image shows what appears to be a hand inserted into the birth canal, with the arm extended and the hand seemingly guiding or manipulating the baby's body. The baby is positioned in a way that suggests a breech presentation, where the baby's feet or buttocks are positioned to be born first.

Right Side: This part illustrates a continuation or a different aspect of the same procedure. The image shows a hand grasping the baby's leg, with an arrow suggesting movement. This movement likely indicates an attempt to facilitate the delivery by adjusting the baby's position.

Clinical Interpretation:

The procedure depicted seems to involve manual manipulation to assist in the delivery of a baby, possibly in a breech presentation. This kind of intervention is considered in situations where a vaginal delivery is desired but the baby's position complicates the process. It's crucial that such maneuvers are performed by experienced healthcare professionals to minimize risks to both the mother and the baby.

Conclusion:

In conclusion, the image illustrates a clinical procedure related to assisting in the vaginal delivery of a baby, likely in breech presentation. The exact nature of the procedure (whether it's a specific type of breech extraction or another maneuver) would depend on the clinical context and the exact techniques being used. This kind of procedure requires expertise in obstetrics and is subject to strict guidelines to ensure safety

Tabela B3: Resultados das estratégias 1 e 2 para a descrição de figuras com imagem de ecografia

Estratégia	Descrição gerada
Estratégia 1 – Contexto textual	Figure 2 likely shows an example of 'funnelling' (i.e., the exteriorization of the fetus's lower limbs during labor) after the application of the Pinard maneuver. More specifically, it should show the fingers inserted along the fetus's thighs, pushing them upwards and outwards, promoting the exteriorization of the lower limbs.
Estratégia 2 – Análise Multimodal	The image appears to be a transvaginal ultrasound scan of a uterus. The image shows a sagittal view of the uterus with a hypoechoic area within the endometrial cavity, suggesting a possible gestational sac or a fluid collection. The presence of a hyperechoic ring around the gestational sac is not clearly visible; however, a measurement of 5.9 mm

is noted, which may indicate the size of the gestational sac or a specific structure within the uterus. Based on the provided information and image, it seems likely that this is an early pregnancy evaluation. However, without additional clinical context or information about the patient's medical history, a definitive diagnosis cannot be made. Further evaluation and follow-up scans would be necessary to confirm the diagnosis and monitor the pregnancy. Some possible differential diagnoses could include an early intrauterine pregnancy, a pseudosac in the setting of an ectopic pregnancy, or a fluid collection within the endometrial cavity. A thorough clinical evaluation and additional imaging studies would be required to determine the correct diagnosis.

Tabela B4: Resposta dos modelos para classificar as figuras do artigo

Figura	Resposta do modelo- Llava 7B	Resposta do modelo - Llava 4 Maverick 17B 128E	Resposta do modelo - Llava 4 Scout 17B 16E
Fig. 24(a): Imagem	“Diagram – figures consisting of flows, connections, steps, or sequential elements.”	“Diagram – The image is a simple, black-and-white illustration showing a fetus in a specific position, with arrows indicating a process or procedure, likely related to childbirth or fetal development.”	“Diagram – The image depicts a simplified, illustrated representation of an anatomical process or procedure with arrows indicating flow or steps.”
Fig. 24(b): Diagrama	“Diagram – figures consisting of flows, connections, steps, or sequential elements;”	“Diagram – The image is a flowchart describing the steps and decisions related to the management of a fetus in a pelvic position at or after 36 weeks of gestation.”	“Diagram – The image presented is a flowchart illustrating a decision-making process for a medical procedure.”
Fig.24(c): Imagem	“Image – a photograph taken from an ultrasound machine	“Image – The image is an ultrasound scan, likely used for medical	“Image – The image presented is an ultrasound scan, which is a type of medical imaging used

	displaying a view of a uterus with what appears to be a fetal development within it, captured by the transducer of the ultrasound.”	diagnosis monitoring.”	or to visualize internal structures, in this case, likely the uterus or a similar organ.”
Fig.24(d): Imagem	“Diagram – This image appears to be a humorous diagram showing the "inside" workings of a person's ear, which is not anatomically accurate but rather exaggerated for comedic effect.”	“Diagram – The image is a medical illustration depicting a procedural step.”	“Image – The image presented is a clinical illustration depicting a medical procedure.”
Fig.24(e): Imagem	“Diagram – <Short justification> The image is a diagram that shows an elephant in different poses from left to right.”	“Diagram – The image presents a series of illustrations depicting the pelvis and a baby in different stages of childbirth, suggesting a diagrammatic representation of a medical process.”	“Image – The image contains clinical illustrations of pelvis and baby's position during delivery.”

Anexo C: Lista dos artigos científicos e normas clínicas utilizadas na base de dados do chatbot

Tema Clínico	Tipo de Documento	Fonte/Entidade	Título	Ano
Pré-eclâmpsia	Case Report	PubMed	A case of hypertensive disorders of pregnancy that developed at 9 weeks of gestation	2021
	Artigo de Revisão	PubMed	A Narrative Review on the Pathophysiology of Preeclampsia	2024
	Norma Clínica	FSPOG	Orientação clínica das doenças hipertensivas da gravidez	2022
	Revisão Sistemática	PubMed	Cardiovascular events following pregnancy complicated by pre-eclampsia with emphasis on comparison between early- and late-onset forms: systematic review and meta-analysis	2020
	Artigo	PubMed	A Comparative Study of Oral Nifedipine and Intravenous Labetalol for Acute Hypertensive Management in Pregnancy: Assessing Feto-Maternal Outcomes in a Hospital-based Randomized Control Trial	2024
	Norma Clínica	NICE	Hypertension in pregnancy: diagnosis and management	2023
	Revisão	PubMed	Toward personalized management of chronic hypertension in pregnancy	2022
	Revisão Sistemática	PubMed	Copper, Zinc, and Selenium Levels During Pregnancy and Their Relationship with Pre-Eclampsia Risk: A Systematic Review	2025
	Revisao	PubMed	Pyroptosis as a therapeutic target in preeclampsia: current research and future directions	2025
Diabetes Gestacional	Artigo de Revisão	PubMed	Antepartum Fetal Surveillance and Optimal Timing of Delivery in Diabetic Women: A Narrative Review	2024

	Norma Clínica	NICE	Diabetes in pregnancy: management from preconception to the postnatal period	2020
	Artigo de revisão	PubMed	Gestational Diabetes Mellitus: The Crosslink among Inflammation, Nitroxidative Stress, Intestinal Microbiota and Alternative Therapies	2022
	Artigo de Revisão	PubMed	Life-cycle approach for prevention of gestational diabetes mellitus	2018
	Artigo de Revisão	Pubmed	Neonates at Risk: Understanding the Impact of High-Risk Pregnancies on Neonatal Health	2025
	Norma Clínica	DGS	Diagnóstico e conduta na Diabetes Gestacional	2011
	Artigo	Revista Portuguesa de Diabetes	Consenso Diabetes e Gravidez: Atualização 2025	2025
	Artigo de Revisão	PubMed	Experiences of women with type 2 diabetes during the pre-pregnancy, pregnancy and postpartum periods: A systematic review of qualitative studies	2024
Asma durante a Gravidez	Norma Clínica	BTS	Asthma in pregnancy	2016
	Norma Clínica	NICE	Intrapartum care for women with existing medical conditions or obstetric complications and their babies	2019
	Ensaio clínico	PubMed	Effect of early and late prenatal vitamin D and maternal asthma status on offspring asthma or recurrent wheeze	2022
	Case Report	PubMed	New onset asthma during pregnancy: two case reports	2023
	Artigo	PubMed	The safety of asthma medications during pregnancy and lactation: Clinical management and research priorities	2009
	Artigo de Revisão	PubMed	Risk factors for asthma exacerbations during pregnancy: a systematic review and meta-analysis	2022
Anemia / Suplementação	Norma Clinica	FIGO	FIGO good practice recommendations on anemia in pregnancy, to reduce the incidence and impact of postpartum hemorrhage (PPH)	2025
	Artigo	PubMed	Impact of high-dose folic acid supplementation in pregnancy on	2021

Versão Cefálica Externa			biomarkers of folate status and 1-carbon metabolism: An ancillary study of the Folic Acid Clinical Trial (FACT)	
	Artigo de Revisão	PubMed	Pathophysiology and management of iron deficiency anaemia in pregnancy: a review	2024
	Artigo	Revista Científica da Ordem dos Médicos	Recomendações Práticas de um Grupo de Peritos para Otimização do Rastreio, Diagnóstico e Tratamento da Ferropenia na Gravidez em Portugal	2025
	Revisão	PubMed	The effect of intravenous iron supplementation compared to oral iron supplementation during pregnancy on neonatal outcomes—a systematic review of randomized controlled trials	2025
	Norma Clínica	SPOMMF	Anemia na Gravidez e no Puerpério	2019
	Revisão	PubMed	Anaemia in pregnancy	2016
	Revisão	PubMed	Vitamin D and Intrauterine Growth Restriction (IUGR)	2025
	Revisão	PubMed	Dietary Supplements in Pregnancy and Postpartum: Evidence, Safety Challenges and a Precision Nutrition Framework (GAPSS)	2025
	Artigo	PubMed	Sedation with Propofol plus Paracetamol in External Cephalic Version: An Observational Study	2022
	Norma Clínica	FIGO	FIGO good practice recommendations for vaginal birth after cesarean section	2025
	Norma Clínica	SPOMMF	Normas de Orientação Clínica: Versão Cefálica Externa	2021
	Artigo de Revisão	PubMed	Reviving external cephalic version: a review of its efficacy, safety, and technical aspects	2019
	Artigo de Pesquisa	PubMed	The influence of advanced maternal age on the success rate, delivery mode and maternal and fetal outcomes of pregnant women undergoing external cephalic version: a retrospective cohort study	2026
	Artigo de revisão	PubMed	Vaginal delivery of the second twin in unengaged cephalic presentation	2019

Anexo D: Resultados obtidos na avaliação da recuperação de chunks e respectivos metadados

	Tipo	Recuperação de texto, Título Principal e Subtítulos		Recuperação com Descrições, Legendas de Imagens e Título		Recuperação com Tabelas, Legendas e Título do Documento	
		R@10	nDCG@10	R@10	nDCG@10	R@10	nDCG@10
MiniLM	Dense	0,4593	0,4541	0,9822	0,8971	0,7546	0,7090
MPNet	Dense	0,5548	0,5311	0,9822	0,9002	0,8507	0,7868
MedEmbed	Dense	0,6364	0,6036	0,9822	0,9146	0,8719	0,8294
PubMedBERT	Dense	0,5575	0,4804	0,9822	0,8637	0,8329	0,7845
BM25	Sparse	0,5625	0,5653	0,9756	0,9146	0,8290	0,7004
MiniLM+BM25	Hybrid	0,5373	0,5245	0,9756	0,9327	0,8330	0,7400
MPNet+BM25	Hybrid	0,5543	0,5702	0,9756	0,9087	0,8516	0,7527
MedEmbed+BM25	Hybrid	0,5298	0,5537	0,9756	0,9146	0,8290	0,7004
PubMedBERT+BM25	Hybrid	0,5544	0,5031	0,9822	0,9199	0,8027	0,7156

