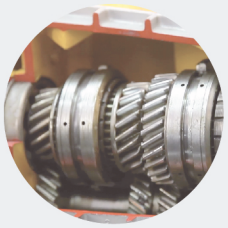
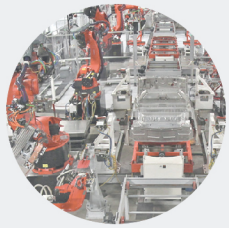
INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA E GESTÃO INDUSTRIAL



IA na Geração de Relatórios Técnico-Científicos de apoio à decisão em Farmacologia Clínica

JOÃO CARLOS MARQUES RIBEIRO
outubro de 2025

POLITÉCNICO
DO PORTO

ISEP INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

IA na Geração de Relatórios Técnico-Científicos de apoio à decisão em Farmacologia Clínica

Prova de Conceito

João Carlos Marques Ribeiro

Dissertação para obtenção do Grau de Mestre em Engenharia e Gestão Industrial

Orientador: Carlos Manuel Abreu Gomes Ferreira
Co-orientador: Davide Rua Carneiro

Porto, setembro 2025

Resumo

Esta dissertação apresenta uma prova de conceito (PoC) para a aplicação de técnicas de IAG, em particular *Large Language Models*, na automatização da elaboração de relatórios técnicos no domínio da Farmacologia Clínica (FC). O trabalho tem como motivação apoiar a Unidade de Farmacologia Clínica (UFC) do Hospital de São João, Porto, na produção de relatórios consistentes e cientificamente fundamentados para responder a pedidos clínicos complexos (ex.: implementação de terapêuticas *off-label*, avaliação do custo-benefício de um fármaco em determinado contexto clínico, utilização de um medicamento em detrimento de outro, entre outros), pedidos estes realizados pela Comissão de Farmácia e Terapêutica (CFT). O sistema proposto integra metodologias como *Retrieval-Augmented Generation* (RAG), *embeddings* semânticos e servidores MCP, procurando replicar, com a maior fidelidade possível, os processos seguidos pela UFC.

A investigação centrou-se em dois eixos principais: a recuperação de informação a partir de documentos regulamentares médicos, sobretudo, o "Resumo das Características do Medicamento" (RCM) dos fármacos, através de pesquisa semântica baseada em *embeddings*; e a revisão automatizada de literatura, recorrendo a um servidor MCP (BioMCP) para recuperar e sintetizar evidência relevante do PubMed e de fontes complementares. Diversos modelos de *embedding* e LLMs foram avaliados de forma sistemática, utilizando *datasets* especificamente construídos para o caso, *gold standards* validados por especialistas em farmacologia e métricas clássicas de recuperação de informação (MAP, MRR, NDCG, *Recall, Precision*, entre outras). Os resultados evidenciam o desempenho superior do *text-embedding-ada-002* da OpenAI na ordenação de documentos e do GPT-5 na estabilidade da pesquisa bibliográfica e na geração de relatórios, embora persistam limitações quanto à robustez, adequação clínica e consistência. O protótipo desenvolvido em Streamlit operacionaliza todo o processo, desde a receção estruturada do pedido (título, enquadramento e caso clínico) até à construção modular do relatório. As estratégias de avaliação combinaram análise de cobertura lexical, utilização de *LLM-as-a-judge* e aplicação de *checklists* estruturadas. Embora os resultados obtidos pelo sistema demonstrem a viabilidade técnica e o enorme potencial da automação baseada em IA neste processo em específico, os resultados sublinham que ainda são necessárias bastantes melhorias antes de sua adoção numa área científica tão crítica como a farmacologia/biomédica. Trabalhos futuros deverão centrar-se no *fine-tuning* de modelos específicos para português na área biomédica, na integração de arquiteturas híbridas de recuperação, em pipelines de validação mais robustos, em mecanismos de explicabilidade e na inclusão de especialistas humanos no processo de geração dos relatórios, algo que será crítico em fases posteriores. Em última análise, esta investigação contribui para definir padrões de conceção e metodologias de avaliação da aplicação de IA generativa em contextos médicos de elevada responsabilidade, abrindo caminho para sistemas de apoio à decisão clínica fiáveis e seguros.

Palavras-chave: *Large Language Models, Retrieval Augmented Generation, Model Context Protocol,* Farmacologia Clínica*,* Prova de Conceito

Abstract

This dissertation presents a proof of concept (PoC) for the application of Generative Artificial Intelligence techniques, particularly Large Language Models (LLMs), in the automation of technical report generation in the field of Clinical Pharmacology. The work is motivated by the need to support the Clinical Pharmacology Unit (UFC) of Hospital de São João, Porto, in producing timely, consistent, and scientifically grounded reports for complex clinical requests mad by the Pharmacy and Therapeutics Commission (CFT). The proposed system integrates methodologies such as *Retrieval-Augmented Generation* (RAG), semantic embeddings and MCP servers to replicate, as closely as possible, the processes followed by the hospital's Clinical Pharmacology Unit.

The research focused on two main pillars: information retrieval from regulatory documents, namely the Summary of Product Characteristics (SmPC), through embedding-based semantic search; and automated literature review, leveraging a biomedical MCP server (BioMCP) to retrieve and synthesize relevant evidence from PubMed and complementary sources. Multiple embedding models and LLMs were systematically evaluated using purpose-built datasets, gold standards validated by pharmacology experts, and classical IR metrics (MAP, MRR, NDCG, Recall, Precision). The results highlight the superior performance of OpenAI's *text-embedding-ada-002* in document ranking and GPT-5 in literature search stability and report generation, although limitations remain regarding robustness, clinical adequacy, and consistency.

The developed Streamlit-based prototype operationalizes the full pipeline, from structured input (title, context, clinical case) to the modular construction of a draft report. Evaluation strategies combined lexical coverage analysis, *LLM-as-a-judge* assessments, and structured checklists. While the system demonstrates the technical feasibility and potential of AI-driven automation in Clinical Pharmacology, the findings highlight that substantial improvements are still required before their full adoption in such a critical scientific domain as pharmacology/biomedicine.

Future work should address fine-tuning of domain-specific models in European Portuguese, integration of hybrid retrieval architectures, enhanced validation pipelines, explainability mechanisms and human in the loop within the report generation process. Ultimately, this research contributes to defining design patterns and evaluation methodologies for applying generative AI to high-stakes medical contexts, paving the way for trustworthy clinical decision support systems.

Keywords: Large Language Models, Retrieval Augmented Generation, Model Context Protocol, Clinical Pharmacology, Proof of Concept

Índice

Lista de Figuras e Gráficos

Orientador: Carlos Manuel Abreu Gomes Ferreira
Co-orientador: Davide Rua Carneiro

Porto, setembro 2025

Lista de Tabelas

Acrónimos

Lista de Acrónimos

CFT	Comissão de Farmácia e Terapêutica
CNN	*Convolutional Neural Network*
EMA	*European Medicines Agency*
FC	Farmacologia Clínica
FDA	*Food and Drugs Administration*
IA	Inteligência Artificial
IAG	Inteligência Artificial Generativa
LLM	*Large Language Model*
MLP	*Multilayer Perceptron*
NLP	*Natural Language Processing*
NN	*Neural Network*
RCM	Resumo das Caraterísticas do Medicamento
RNN	*Recurrent Neural Network*
SLM	*Small Language Model*
UFC	Unidade de Farmacologia Clínica

Orientador: Carlos Manuel Abreu Gomes Ferreira
Co-orientador: Davide Rua Carneiro

Porto, setembro 2025

1. Introdução

Neste capítulo faz-se uma breve descrição do contexto em que se realizou o projeto, bem como a pertinência do tema. Seguidamente, enuncia-se a questão de investigação e os objetivos global e específicos que se pretendem alcançar no decorrer do projeto, terminando com as opções metodológicas.

1.1. Problemas de investigação, enquadramento e pertinência

Falar de tecnologia sem falar de inteligência artificial (IA) é, nos dias que correm, inimaginável. Organizações têm procurado utilizar IA, não só nos seus produtos e serviços, mas também nos seus processos, com vista a aumentar produtividade e eficiência. A tendência tornou-se exponencialmente crescente com o lançamento da ferramenta ChatGPT, em novembro de 2022, que veio abalar e redefinir os padrões de IA (OpenAI, 2022). Atraindo um milhão de usuários em 5 dias, modelos como o GPT têm a capacidade de interpretar e gerar texto de uma forma bastante eloquente e natural, nos mais variados tópicos e formas, como se de um humano se tratasse, tudo isto em questões de segundos. Desde então, milhares de novas aplicações e modelos foram lançados, cada um com as suas características específicas, expandindo as capacidades na geração de novo conteúdo para outras modalidades, como imagem, vídeo, áudio e até CAD, com a criação de designs tridimensionais virtuais com base em texto ou imagens. Esta nova vertente de IA, conhecida como Inteligência Artificial Generativa (IAG), tem captado a atenção das organizações que agora procuram entender e explorar como podem utilizar estas novas tecnologias para ganhar competitividade.

Seguindo os passos de revoluções passadas, IAG promete transformar a nossa forma de trabalhar, e está já muito além de apenas automatização de tarefas repetitivas simples, com a sumarização de documentos e reuniões, criação de relatórios, suporte personalizado ao cliente, campanhas de marketing personalizadas, desenvolvimento acelerado de aplicações e websites, análises de mercado inteligentes, gestão de risco potenciada, identificação de riscos, criação de cenários hipotéticos, ações de mitigação e planos de contingência, cibersegurança reforçada,

entre muitas outras. Dadas as suas capacidades de analisar grandes quantidades de informação e gerar novo conteúdo, a criatividade dos utilizadores torna-se, agora, o *bottleneck* neste novo mundo de possibilidades.

Com a crescente complexidade ao nível terapêutico, o aumento da polimedicação e a necessidade de decisões clínicas cada vez mais fundamentadas e céleres, torna-se importante perceber como podem os serviços de saúde, e em particular os serviços hospitalares clínicos, integrar, de forma eficaz, modelos e sistemas baseados em IA para apoiar os seus processos, com especial ênfase numa segura, eficaz e racional tomada de decisão. Esta análise será realizada, especificamente, no contexto da FC, uma disciplina médica e científica essencial para garantir a qualidade terapêutica, a segurança dos doentes e a sustentabilidade dos sistemas de saúde.
Ressaltar que a presente dissertação não pretende aprofundar sobre a prática da FC, mas sim, perceber como tecnologias baseadas em modelos de linguagem natural podem contribuir para melhorar a eficiência e a qualidade numa atividade específica realizada por profissionais da Unidade de Farmacologia do Porto (UFC), nomeadamente, na elaboração de relatórios técnico-científicos, em resposta a pedidos realizados pela Comissão de Farmácia e Terapêutica (CFT), relativamente a questões de implementação de terapêuticas *off-label*, avaliação do custo-benefício de um fármaco em determinado contexto clínico, utilização de um medicamento em detrimento de outro, entre muitos outros.

1.2. Questão e objetivos de Investigação

Tendo em conta o enquadramento apresentado no capítulo anterior, a presente dissertação propõe-se responder à seguinte questão de investigação: de que formas podem as tecnologias de IAG, com ênfase (não exclusiva), em *Large Language Models*, contribuir para a agilização, qualidade e eficiência dos relatórios técnicos elaborados pelos profissionais de FC, promovendo, assim, uma tomada de decisão que resulte numa utilização mais segura, racional e eficaz dos medicamentos?
Com o objetivo geral de conceber e validar um modelo que auxilie as atividades de suporte à decisão na prática da FC, foram definidos os seguintes objetivos específicos:

1. Estudo do processo de elaboração do relatório técnico-científico (estrutura do relatório, tipologias dos pedidos, fontes utilizadas, métodos de trabalho e fluxo geral de informação);
2. Levantamento dos modelos existentes que melhor se adequam às necessidades identificadas, tendo em conta os requisitos da prática clínica e regulatória;
3. Definição dos requisitos técnicos necessários para a implementação responsável do modelo;
4. Elaboração de um PoC (Prova de Conceito) que demonstre a aplicabilidade da tecnologia no enquadramento em causa;
5. Testagem do modelo com base num conjunto simulado de pareceres técnicos e dados clínicos;

6. Avaliação dos resultados obtidos, com base em métricas de precisão, utilidade e aplicabilidade prática, e subsequente afinação do modelo;

Considerando os objetivos gerais e específicos apresentados anteriormente, optou-se por estruturar as opções metodológicas segundo os níveis de metodologia, método de investigação e técnicas e instrumentos de recolha e análise de dados, de acordo com a abordagem sugerida por Ramos (2011) e Caixeta & Fabricio (2018), que defendem a existência de uma relação hierárquica entre estes níveis, desde os mais amplos aos mais específicos.

1.3. Opções metodológicas

No que diz respeito à metodologia, e em consonância com as perspetivas de Ramos (2011) e Stockemer (2019), são habitualmente reconhecidas duas grandes abordagens: quantitativa e qualitativa. Segundo Ramos (2011), a abordagem quantitativa parte do pressuposto de que o problema de investigação tem uma solução objetiva, que pode ser analisada através da medição de variáveis. No presente caso, dado que se procura explorar um fenómeno emergente, a aplicação de modelos de linguagem natural na prática da Farmacologia Clínica, sem um corpo teórico amplamente estabelecido, a opção recai sobre uma abordagem qualitativa.

Quanto ao método de investigação, opta-se pela investigação-ação, uma vez que o estudo visa a resolução de problemas concretos e reais no seio da prática clínica, com envolvimento direto do investigador no processo de transformação. Trata-se de um processo iterativo e cíclico, que combina diagnóstico, planeamento, ação e avaliação (Ramos, 2011), sendo particularmente adequado a contextos de inovação tecnológica e melhoria de processos. Este método permite o desenvolvimento progressivo e participativo de soluções, nomeadamente: levantamento de necessidades junto de farmacologistas clínicos, desenvolvimento de um protótipo de modelo de linguagem, teste em cenários simulados, avaliação dos resultados, reformulação e reimplementação do modelo, entre outros.

A investigação será, assim, desenvolvida em ciclos sucessivos, em que se alternam momentos de desenvolvimento, experimentação, avaliação crítica e ajuste, articulando o conhecimento científico com a prática real. Esta abordagem permite não só validar empiricamente a aplicação dos LLMs na FC, como também promover uma solução adaptada às necessidades concretas dos profissionais de saúde, assegurando a sua relevância e aplicabilidade.

1.3.1. Metodologia CRISP-DM

O *Cross Industry Standard Process for Data Mining* (CRISP-DM) é um *framework* com o intuito de descrever e orientar os processos de análise e processamento de dados (Hotz, 2024), sendo, dentro de muitas outras, a metodologia mais comummente utilizada neste tipo projetos (Saltz, 2020).

Este modelo representa uma sequência idealizada de etapas. Na prática, muitas das tarefas podem ser realizadas noutra ordem e, frequentemente, torna-se necessário regressar a fases anteriores e repetir determinadas ações (Figura 1).

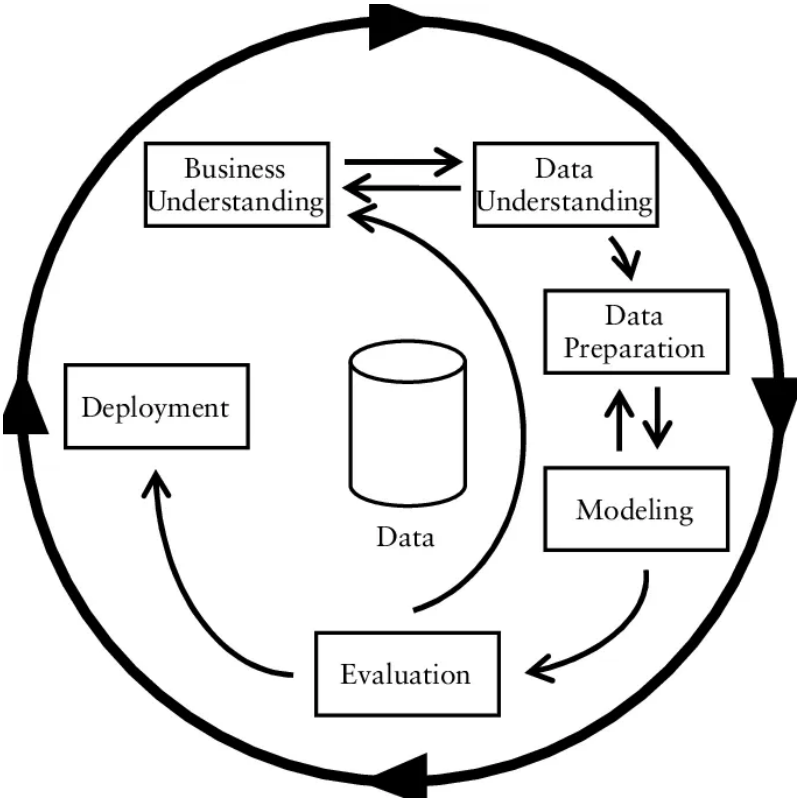


Figura 1 - Representação do modelo CRISP-DM

Fonte: (Löfström, 2009)

Tendo em conta a natureza exploratória, iterativa e aplicada do presente trabalho, optou-se pela adoção deste *framework* como linha orientadora do processo de desenvolvimento e avaliação do respetivo projeto. Desta forma, as etapas poderão ser descritas da seguinte forma (Casonatto et al., 2024; Hotz, 2024; IBM, 2021):

1. ***Business Understanding***: corresponde à compreensão do contexto clínico e farmacológico, bem como dos principais desafios enfrentados pelos profissionais na prática diária, mais concretamente, no que diz respeito ao processo de elaboração de pareceres técnicos, que envolvem o estudo da utilização de determinados medicamentos em casos clínicos específicos. Nesta fase, são definidos os objetivos gerais e específicos do projeto, identificados os requisitos prioritários e estabelecidos os critérios de sucesso da solução a desenvolver;

2. ***Data Understanding***: envolve o levantamento e análise dos dados disponíveis, incluindo literatura científica, bases de dados clínicas e histórico de pareceres realizados. Esta etapa permite mapear os fluxos de informação relevantes e compreender a estrutura, qualidade e heterogeneidade dos dados existentes.

3. ***Data Preparation***: abrange os processos de preparação dos dados para integração com os modelos de *embedding* e LLMs;

4. ***Modeling***: corresponde à seleção e *fine-tuning* dos modelos utilizados (*embeddings* e LLMs);

5. ***Evaluation***: refere-se à avaliação dos resultados obtidos nas diferentes fases do projeto, através da análise da utilidade, precisão, relevância e adequação dos outputs gerados pelo modelo, com base em métricas e *benchmarks* específicos;

6. ***Deployment***: embora não se preveja uma integração definitiva em ambiente clínico real, esta fase foi adaptada para simular a aplicação prática do modelo em contextos controlados, com recolha de *feedback* por parte dos profissionais.

1.3.2. Estrutura da Relatório

Este relatório encontra-se estruturado da seguinte forma: primeiramente, no Capítulo 2, é feita uma revisão da principal literatura existente no âmbito da IA, abordando a evolução da área e suas ramificações, dando especial ênfase às principais técnicas/ferramentas utilizadas no presente trabalho, como os LLMs, RAG e MCP *Servers*. Ainda neste capítulo, descreve-se a área da FC, a sua importância e impacto na prática médica, bem como alguns modelos e *benchmarks* desenvolvidos para serem utilizados neste domínio específico.
No Capítulo 3, inicia-se a análise à área central desta dissertação, onde é descrita, detalhadamente, não só a estrutura e conteúdo do relatório técnico-científico elaborado pela UFC, mas também o próprio processo da sua redação. Na secção 3.1 é apresentado uma proposta de automação desse mesmo processo, explorando-se as diversas técnicas de IA mencionadas posteriormente. Os capítulos seguintes detalham essas mesmas automações, com um eixo principal na recuperação de informação (IR). Na secção 4.1, descreve-se como modelos de *embedding*, indexação e pesquisa semântica são utilizados na redação da secção "*Fundamentação: Farmacologia*" do relatório técnico. A secção 4.2 aborda o uso do BioMCP, servidor MCP utilizado na redação da secção "*Fundamentação: Estudos Relevantes*". Já as restantes secções, por envolverem tarefas mais simples (essencialmente, sumarização), são tratadas no Capítulo 5, onde é descrito protótipo desenvolvido. Este capítulo apresenta em detalhe cada uma das etapas, desde a introdução dos *inputs* por parte do utilizador até à geração do relatório final. Mencionar que a apresentação das técnicas mencionadas acima é acompanhada pela respetiva validação e avaliação dos resultados com métricas específicas. Por fim, o Capítulo 6 encerra a dissertação, destacando as conclusões alcançadas e apontando possíveis trabalhos futuros.

2. Revisão da Literatura

Neste capítulo encontra-se um enquadramento do tema proposto, começando por abordar genericamente IAG, nomeadamente, em que consiste, como surgiu e respetivo impacto na indústria, seguido de usos atuais e potenciais futuros. Posteriormente, reduz-se o âmbito para os *Large Language Models* (LLMs) que servirão de base para o trabalho a desenvolver, onde será descrita a base para o seu funcionamento, o panorama de mercado com a análise das principais soluções existentes, finalizando-se com alguns desafios inerentes à sua utilização e abordagens e técnicas para os ultrapassar, com maior especificidade no domínio da FC.

2.1. Inteligência Artificial Generativa: Uma Introdução

Inteligência Artificial Generativa: Conceito

Ao redor do mundo, empresas, instituições e indivíduos procuram, incessantemente, aprender e aprofundar os seus conhecimentos em IAG, com o intuito final de agilizar o seu trabalho, nas suas mais variadas formas, e numa busca por uma crescente produtividade, eficiência e competitividade. De facto, de acordo com uma pesquisa realizada pela McKinsey, com cerca de 1500 participantes pertencentes às mais variadas indústrias e nacionalidades, a consultora concluiu que mais de 70% das organizações já utilizam esta tecnologia de forma regular em, pelo menos, um dos seus processos de negócio (Alex Singla et al., 2025), com preponderância em grandes empresas, que começam a redesenhar os seus fluxos de trabalho para esse fim. Exploremos as razões para esta nova tendência, que promete revolucionar várias indústrias em grande escala.

Quando falamos de IAG, falamos de uma ramificação da IA, com a capacidade de criar conteúdo original (em variados formatos), tendo, como ponto de partida, uma determinada pergunta ou pedido de um utilizador (Aamer Baig et al., 2024; Cole Stryker & Mark Scapicchio, 2024). Consiste em aplicar várias técnicas e algoritmos de IA, com o objetivo de gerar novos conteúdos sintéticos através de um conjunto de dados de treino, mais propriamente, dos seus padrões e distribuições (Jovanović & Campbell, 2022), criando outputs que representam esses padrões em novas formas (Foster & Friston, 2023).

Especificando, IAG é agora capaz de gerar textos, discursos, imagens, vídeos, código e até objetos tridimensionais em softwares CAD. Aliando estas capacidades aos dados específicos de indivíduos, sociedades ou organizações, é agora possível um grau de personalização com muito mais nuance e uma interação digital bastante mais fluída e simplificada. (Aamer Baig et al., 2024; Adam Zewe, 2023; Cole Stryker & Mark Scapicchio, 2024; NVIDIA, 2024).

Evolução da IA: Contexto Histórico e Teórico

Por forma a entender o que tudo isto significa concretamente e de que forma se processa, torna-se imperativo navegar (ainda que em alto nível) pela história da IA, realçando os marcos mais importantes que possibilitaram o seu desenvolvimento.

De facto, apesar de a IA só ter alcançado uma elevada notoriedade, aos olhos do grande público, no final de 2022 (com o lançamento da ferramenta ChatGPT (OpenAI, 2022)) e que possibilitou uma democratização da tecnologia nunca vista, na verdade, o seu percurso tecnológico começou muito antes , nos inícios do século XX, com um modelo bastante simples conhecido como a cadeia de Markov (*Markov Chain*), em 1906, com a capacidade de próxima palavra numa frase, baseando-se na(s) palavra(s) anterior(es), ainda de que de forma muito rudimentar (Adam Zewe, 2023). Seguido pelo Neurónio de McCulloch-Pitts, em 1943, com a simulação matemática do neurónio biológico, interligando, assim, a neurociência e a ciência da computação, base muito importante para as redes neurais que conhecemos hoje (Catarina Moreira, 2013; Chandra, 2018; Jeremy Norman, 2021; McCulloch & Pitts, 1990). Estas "simulações" foram ficando sucessivamente melhores, com o *Perceptron* de Frank Rosevelt (Rosenblatt, 1958), o *software* ELIZA (Weizenbaum, 1966) e o sistema DENDRAL (B. J. Copeland, 2025; Feigenbaum, 1969). Com um grande impacto ainda atualmente, é de grande importância o desenvolvimento da técnica de retropropagação (Rumelhart et al., 1986), onde os autores, no artigo, "*Learning representations by back-propagating errors*" estabelecem a base do algoritmo de retropropagação do erro ainda hoje utilizado no treino de redes neurais (Dave Bergmann & Cole Stryker, 2024). Mais recentemente, podemos mencionar o Google Autocomplete, realidade mais próxima do utilizador comum. Este foi um projeto desenvolvido por Kevin Gibbs, que oferece sugestões de pesquisa em tempo real enquanto o utilizador digita uma determinada *query*, com o objetivo de tornar a experiência de pesquisa mais rápida e conveniente, permitindo aos utilizadores realizar pesquisas com maior facilidade e explorar tópicos que talvez não tivessem considerado inicialmente (Kevin Gibbs, 2004; Liz Gannes, 2013). Vale, sobretudo, ressaltar os princípios dos Transformadores e o surgimento dos "GPTs (*Generative Pretrained Transformer)*". Fundamentais para os modelos de linguagem de ponta que vemos atualmente, a arquitetura dos transformadores surge em 2017, num trabalho realizado por uma equipa do Google Brain, juntamente com um grupo da Universidade de Toronto (Vaswani et al., 2017), e são agora empregados em uma ampla variedade de tarefas de *NLP*.

Embora não aprofundados neste trabalho, os marcos mencionados acima (e muitos outros) contribuíram significativamente para a consolidação da IA como área científica e para a sua sofisticação, tendo muitos deles servido de alicerce e/ou inspiração para os desenvolvimentos mais recentes que hoje assistimos.

Importa, ainda, sublinhar que o percurso da IA não foi, de todo, linear. Ao longo das últimas décadas, a área enfrentou períodos de grande entusiasmo seguidos por fases de desilusão e estagnação (os chamados AI *winters*) em que o financiamento e o interesse diminuíram significativamente (Edd Gent, 2024). Ainda assim, apesar dos altos e baixos, a evolução tecnológica e o crescimento exponencial de dados e de capacidade computacional permitiram avanços notáveis na indústria. Hoje, a IA é uma força motriz incontornável, com impacto

transversal em diversas áreas do conhecimento e da sociedade. O futuro da IA já não é apenas uma hipótese: é uma realidade em constante expansão.

Estatísticas, Aplicações e Vantagens Competitivas Consequentes

Não é difícil imaginar como estas novas tecnologias podem impactar várias indústrias, nas suas mais variadas vertentes. Um relatório publicado pela *Deloitte* denominado de "*The Generative AI Dossier*" identifica 60 casos de aplicação em 6 grandes vetores, nomeadamente: energia, recursos e indústria; serviços financeiros; governo e serviços públicos; saúde e ciências sociais; tecnologia, media e telecomunicações e, por fim, na vertente do consumidor (incluindo retalho, automóvel, estadias, restauração, viagens e transportes). Para cada caso de estudo, exploram as maneiras inovadoras de implementar as tecnologias de IAG com o intuito de ganhar eficiência nos seus processos, produtos e serviços, tanto para as organizações como para os indivíduos e sociedade, tendo, também, em consideração os riscos e respetivas ações de mitigação para ultrapassá-los (Beena Ammanath et al., 2023). De modo semelhante, um estudo realizado pela McKinsey identificou que, após análise de mais de 60 casos de potencial aplicação de IAG, 75% do valor total gerado pelos use cases pode ser categorizado em 4 grandes áreas: CRM, Marketing e Vendas, Engenharia de *Software* e, por fim, Pesquisa e Desenvolvimento, justificando-se, essencialmente, pelo facto de serem áreas que beneficiam da geração de conteúdo personalizado, excluindo-se áreas com aplicações predominantemente numéricas e de otimização, que acabam por beneficiar de outras técnicas de IA (Michael Chui et al., 2023).
De facto, o interesse em IAG ultrapassa agora o reconhecimento que antes era apenas detido pela comunidade científica, especialmente a partir de novembro/dezembro de 2022, que coincide com o lançamento da ferramenta ChatGPT (Figura 2) (OpenAI, 2022).

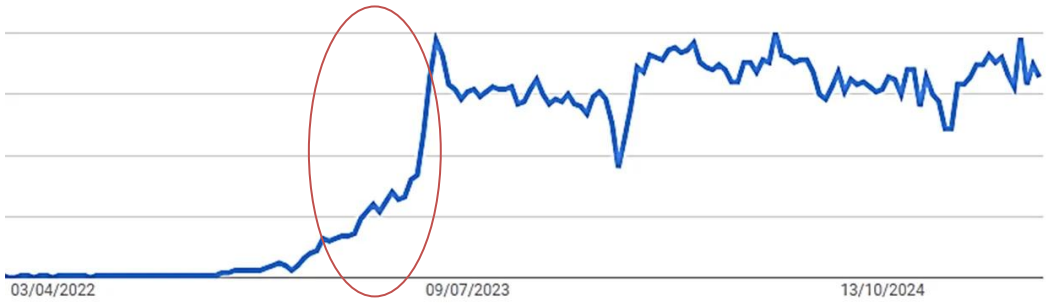


Figura 2 - Interesse global sobre o termo "*Generative AI*" nos últimos 5 anos (% relativamente ao pico máximo)

Fonte: https://trends.google.com/trends/explore?date=2021-01-01%202025-04-11&q=Generative%20AI&hl=pt-PT

Segundo o relatório da *McKinsey & Company*, a "Inteligência Artificial Generativa tem o potencial de mudar a anatomia do trabalho, aumentando as capacidades dos trabalhadores individuais ao automatizar algumas das suas atividades. As atuais tecnologias de IA generativa

têm o potencial de automatizar atividades de trabalho que consomem de 60 a 70 por cento do tempo dos funcionários atualmente." (Michael Chui et al., 2023).
Desde a agricultura (Pallottino et al., 2025), na saúde (D. Chen et al., 2024; Currie et al., 2024; Kanakala et al., 2024; Lyakhova & Lyakhov, 2024; Rodler et al., 2024; Rolls, 2024; Solaiman, 2024), na educação (Acosta-Enriquez et al., 2024; Collie & Martin, 2024; Foung et al., 2024; Jiang, 2024; Rasul et al., 2024; Tan et al., 2025; Ulla et al., 2024; H. Wang et al., 2024), até na robótica (Brunello et al., 2025), cibersegurança (Teo et al., 2024) e analítica financeira (B. Chen et al., 2023), são inúmeras as indústrias que procuram agora adotar a tecnologia, numa procura por uma maior competitividade e eficiência. Contudo, é bastante importante estabelecer expectativas ajustadas e realistas relativamente àquilo que é possível de ser realizado. Deve, então, olhar-se para IAG, não como uma solução omnipotente, mas sim uma ferramenta capaz de automatizar, auxiliar e potencializar algumas atividades. "Esta perspetiva facilitará a integração bem-sucedida da IA em diferentes domínios, sem esperar resultados utópicos, reduzindo a deceção que pode ocorrer quando a IA não se comporta conforme o esperado" (García-Peñalvo & Vázquez-Ingelmo, 2023).

2.2. Large Language Models

Large Language Models (LLMs) tornaram-se os principais protagonistas na forma como interagimos com sistemas computacionais, especialmente, no tratamento e compreensão de linguagem natural, o que possibilitou uma enorme democratização da tecnologia. Estes modelos, que aprendem padrões linguísticos a partir de grandes volumes de texto (como, no caso do GPT, toda a informação disponível na internet) são hoje aplicados numa vasta gama de tarefas, desde a tradução automática, a geração de respostas em tempo real, sumariação de documentos, entre muitas outras. Neste capítulo, propõe-se uma exploração estruturada dos LLMs, começando por uma definição geral e passando pelos principais componentes que os constituem, bem como pelos mecanismos fundamentais que lhes permitem "aprender" e produzir linguagem de forma coerente.
Será também discutido como estes modelos são utilizados na prática, com especial enfoque nas tarefas de sumariação e pesquisa semântica, tarefas com forte aplicabilidade na investigação científica (bastante relevantes no âmbito desta dissertação). Serão ainda abordadas métricas e *benchmarks* que permitem avaliar o desempenho destes sistemas nestas atividades, terminando com os principais fornecedores destes serviços.
Sendo o principal foco os LLMs, vale a pena ressaltar que estes constituem uma categoria daquilo a que chamamos genericamente de *Foundation Models* (FM), ou modelos de fundação. São designados desta forma por servirem de base a uma ampla gama de outros modelos, possuindo uma capacidade genérica para vários tipos de tarefas e aplicações (IBM Research, 2021a; Javier Canales Luna, 2024). As suas principais características são as seguintes:

- **Grande escala**: treinados com enormes volumes de dados e milhares de milhões de parâmetros (geralmente);
- **Elevada generalização**: capazes de realizar várias tarefas sem ser treinadas em um domínio específico;

- **Transferência de conhecimento**: o que aprendem em uma tarefa pode ser usado em outras, até de domínios diferentes;
- **Arquitetura multimodal** (em alguns casos): podem lidar com diferentes tipos de dados, como texto e imagem juntos;
- **Base para outras aplicações**: servem como plataforma para criar sistemas mais específicos, como *chatbots*, tradutores, etc.

Por esta razão, pode dizer-se que todos os LLMs são FM, enquanto o contrário nem sempre se verifica (Javier Canales Luna, 2024).

LLMs | Definição e conceitos-chave

Neural Networks (NN) com milhões, ou até, milhares de milhões de parâmetros (*weights*), e treinados com *datasets* textuais muito vastos (Vashisth, 2025), os LLMs têm um ótimo desempenho numa ampla variedade de tarefas e têm sido exaustivamente exploradas em variadas indústrias, maioritariamente, em atividades relacionadas a NLP, mais concretamente, em tradução, resumo, geração de texto e resposta a perguntas, entre outras (IBM Research, 2023; Stöffelbauer, 2023; Suvojit, 2023).
A arquitetura mais comummente utilizada nos modelos de linguagem atuais é, como mencionado anteriormente, o transformador (Figura 3), introduzido em 2017 com o artigo "*Attention is All You Need*" (Vaswani et al., 2017) e que revolucionou o campo de NLP. Diferentes modelos utilizam variações desta arquitetura. Contudo, pode dizer-se que um transformador é constituído, essencialmente, por 5 componentes principais:

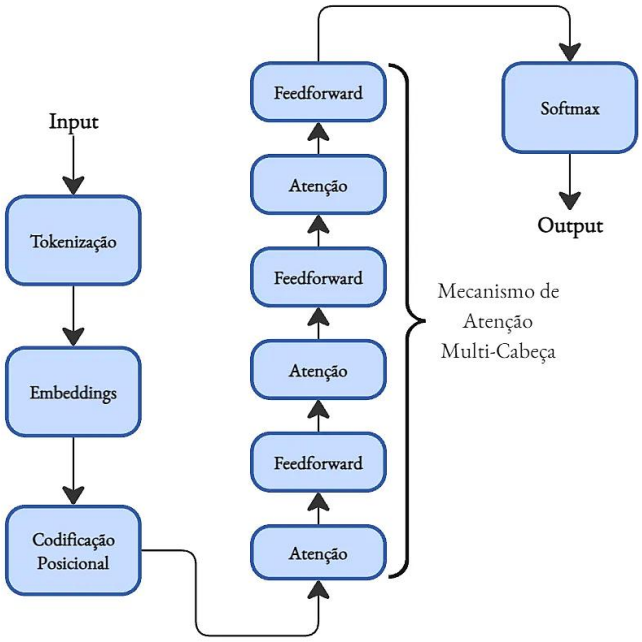


Figura 3 - Arquitetura genérica de um transformador

Fonte: Autor; Adaptado do *paper* "*Attention* is All You Need" (Vaswani et al., 2017)

1. **Tokenização**: A primeira etapa consiste em dividir uma sequência de texto em componentes menores, denominadas de *tokens* (
2. Figura 4). Estes *tokens* podem ser palavras, subpalavras, caracteres ou símbolos, dependendo do tipo de tokenização (Abid All Awan, 2024) .

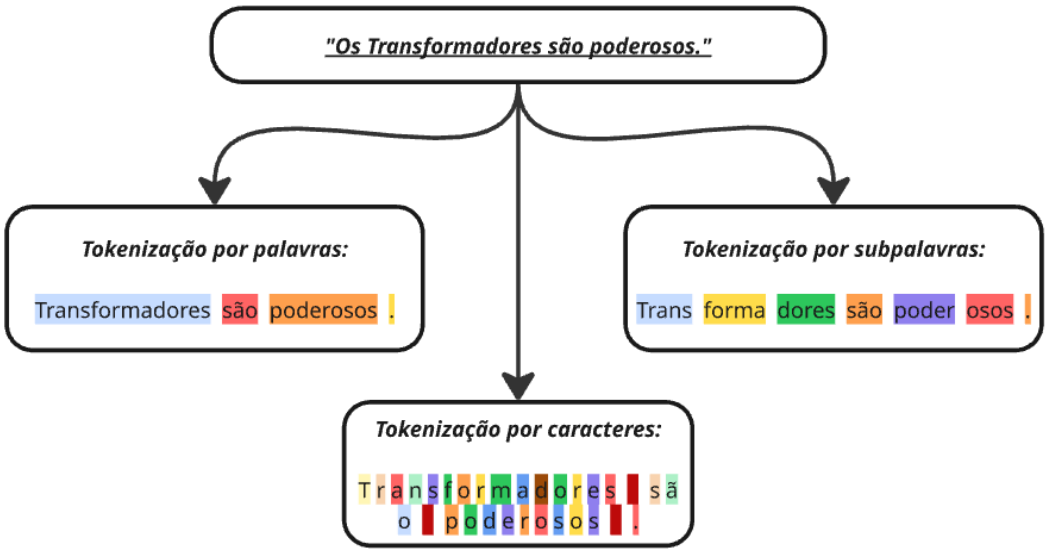


Figura 4 - Diferentes tipos de Tokenização

Este processo de transformação tem como principal objetivo a conversão do *input* do utilizador em componentes padronizados para que o modelo os consiga processar eficientemente, preservando o contexto, dividindo por sub-palavras quando necessário e facilitando o reconhecimento de padrões linguísticos (Menzli, 2022)

3. ***Embeddings***: Com o *input* convertido em *tokens*, a etapa seguinte consiste em "vetorizar" esses *tokens*. Esta vetorização transforma os *tokens* em representações numéricas contínuas que captam o seu significado e respetivas relações semânticas. Por outras palavras, fragmentos de texto com significados semelhantes tendem a gerar vetores mais próximos entre si no espaço vetorial / com números semelhantes (Amanatullah, 2023);

4. **Codificação Posicional**: Após a vetorização dos *tokens*, ocorre um processo de aglomeração em um único vetor para que a sequência possa ser processada como um todo. Visto que a arquitetura dos transformadores não é sequencial por natureza (o processamento dos *tokens* ocorre simultaneamente, em paralelo, sendo esta uma das principais vantagens face a arquiteturas anteriores), não existe uma noção intrínseca da ordem dos *tokens*. Desta forma, é adotado um esquema de codificação posicional, em matriz, onde cada linha representa um *token* vetorizado somado à sua informação de posição dentro da sequência (Amanatullah, 2023; Saeed, 2022);

5. **Atenção Multi-Cabeça e Camadas *Feedforward***: O mecanismo de atenção dos transformadores consiste, essencialmente, em atribuir pesos a diferentes partes de uma sequência de entrada, permitindo ao modelo entender as relações entre os elementos dessa sequência. O *self-attention* possibilita que o modelo observe e compare todos os elementos da sequência de uma vez, destacando aqueles mais relevantes para a tarefa em questão[4] (H20.ai, 2022; Raschka, 2023; Vanna, 2025). Pegando no exemplo "Transformadores têm grande impacto e são poderosos em NLP", o mecanismo de *self-attention* permite que o modelo entenda a relação entre "transformadores" e "poderosos", associando corretamente o contexto da frase, mesmo que as palavras não estejam próximas, como é o caso. Já as camadas de *feedforward* realizam transformações lineares e não-lineares sucessivamente, permitindo que o modelo aprenda representações mais complexas dos dados. Estas camadas são aplicadas a cada posição da sequência do *input* após o processo de *self-attention* mencionado acima, o que ajuda o modelo a capturar, processar e combinar informações de maneira mais abstrata, profunda e complexas.

6. ***Softmax***: Esta etapa captura o *output* do processo anterior e aplica uma função *softmax*. Esta função converte os valores atribuídos a cada *token* em probabilidade que variam entre 0 e 1, sendo que os *tokens* mais próximos de 1 terão maior probabilidade em se tornarem na próxima palavra da frase a ser construída (Amanatullah, 2023). Este processo é repetido até à conclusão do *output*.

Alguns exemplos de LLMs

Desde o já mencionado ELIZA (Weizenbaum, 1966), um dos primeiros programas com a capacidade de simulação de uma conversa com um ser humano, até aos modelos linguísticos avançados da atualidade, como os modelos de última geração GPT-4.5 (OpenAI, 2025), Claude 3.7 (Anthropic, 2025b), Llama 4 (Meta AI, 2025), ou até o novo modelo da *startup* chinesa DeepSeek V3 (Yang, 2025), a evolução tem sido notável (Figura 5). Estes LLMs representam o estado da arte no processamento de linguagem natural. Na verdade, aos dias de hoje, escolher o LLM mais adequado para um determinado tipo de atividade ou contexto pode revelar-se um desafio, dado o elevado número de opções disponíveis, cada uma com características e finalidades distintas.

[4] Em vez de processar os dados de forma sequencial, como acontece em redes neuronais recorrentes (RNNs), o mecanismo de *self-attention* possibilita que o modelo observe e compare todos os elementos de uma vez. Isso torna os transformadores extremamente eficientes, pois conseguem processar a informação paralelamente, sem as limitações de sequência das abordagens anteriores (Vanna, 2025).

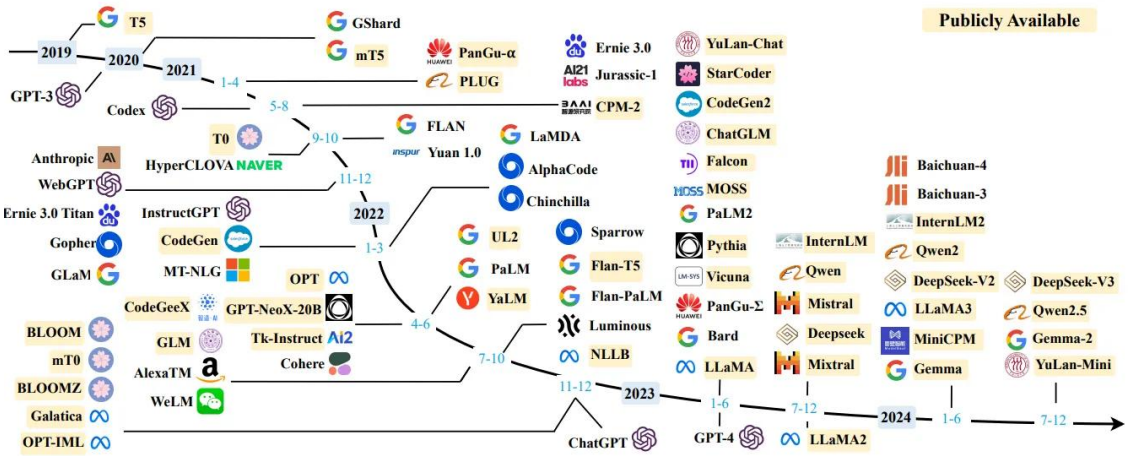


Figura 5 - LLMs existentes (com tamanho superior a 10 mil milhões de parâmetros), até final de 2024.

Fonte: (Kim, 2023)

Avaliação de LLMs - *Benchmarks*

Fatores como a versatilidade, a profundidade nas respostas, a adaptabilidade a tarefas específicas, a capacidade de raciocínio matemático ou de geração de código, a privacidade e segurança do modelo, bem como as capacidades multilinguísticas ou o grau de abertura do modelo (*open source* ou *open weights*)[5] tornam a escolha multifacetada e dependente dos objetivos do utilizador final. Assim, compreender os pontos fortes e limitações de cada modelo torna-se essencial para uma seleção informada e eficiente.
Algumas métricas a ter em consideração na escolha e implementação de LLMs podem ser (Basheer, 2025; Tripathi, 2023):

- **Precisão, completude e autenticidade**: quão precisas e completas são as respostas geradas pelo modelo;
- **Fluência e Coerência**: quão natural, fluído e lógico é o *output* gerado;
- **Relevância**: face ao *prompt* e ao contexto fornecido, quão relevante é a resposta;
- **Consistência**: quão consistente é a informação fornecida quando solicitada várias vezes, ou de forma distinta;
- **Diversidade**: variedade das respostas geradas, garantindo que o modelo não é enviesado para um determinado tipo de resposta;

[5] Modelos open-source disponibilizam não só os pesos treinados, mas também a arquitetura, o código-fonte e, por vezes, até os dados de treino, permitindo uma total transparência e modificação. Já os modelos com open weights partilham apenas os pesos já treinados, sem revelar necessariamente os detalhes da arquitetura ou permitir alterações profundas.
Enquanto os modelos open-source incentivam a colaboração da comunidade e a personalização através de retraining, os modelos com open weights são pensados para facilitar a utilização direta por developers que procuram aplicar o modelo tal como está, sem necessidade de compreender ou alterar a sua estrutura interna (Sunil Ramlochan, 2023).

- **Índice de Alucinação**: com que frequência o modelo alucina, gerando informação incorreta ou inventada;
- **Segurança e Responsabilidade**: mede a presença de linguagem ofensiva ou prejudicial nas respostas do modelo.
- **Versatilidade**: especificação ou generalização do modelo, dependendo do contexto em que será aplicado;
- **Custo**: custo de utilização, desenvolvimento, manutenção, etc.;
- ***Prompt Engineering***: o quão detalhado e específico deverá ser o *prompt* para que o modelo entenda aquilo que está a ser solicitado.

Dependendo do tipo de tarefa à qual o modelo vai ser utilizado, diferentes critérios poderão ser necessários. Daí ser bastante importante a utilização de frameworks de avaliação, ou *benchmarks*, completos e criteriosos que possibilitem uma escolha adequada do LLM. Os *benchmarks* consistem em testes padronizados que são desenvolvidos para medir e comparar as capacidades e a eficiência de diferentes modelos na realização de uma determinada tarefa (Evidently AI, 2025), destacando as suas principais forças e fraquezas (vellum, 2024). Estes *benchmarks* são compostos por um conjunto de dados de amostra, tarefas ou perguntas desenhadas, métricas de avaliação do desempenho e um sistema de pontuação associado (Rina Caballar & Cole Stryker, 2024). Daí ser importante escolher o mais adequado ao tipo de tarefa que pretendemos que o modelo empregue. Alguns dos *benchmarks* mais conhecidos encontram-se na Tabela 1.

Tabela 1 - Alguns exemplos de *benchmarks* para avaliação de LLMs

Benchmark	**Área de avaliação**	**O que avalia?**	**Como o faz?**
AI2 Reasoning Challenge *(Chollet, 2019)*	Raciocínio baseado em ciência e lógica aplicada.	Capacidade de resposta e de raciocínio numa série de 7787 perguntas na área das ciências naturais.	- 1 ponto por cada resposta correta. - 1/N pontos, caso acerte tendo dado N respostas.
HumanEval *(M. Chen et al., 2021)*	Geração de código.	Capacidade de geração de código funcionalmente correto.	Através de uma métrica denominada de pass@k, que mede a probabilidade, de pelo menos, uma entre k soluções geradas passa nos teste unitários associados ao problema.
MT-Bench *(Zheng et al., 2023)*	Capacidade de conversação e de resposta do modelo.	Qualidade das respostas obtidas a questões abertas relativas a código, extração de informação, conhecimento científico, matemática, capacidade de raciocínio e escrita.	Utiliza um outro LLM (geralmente mais poderoso, como o GPT-4) para avaliar as respostas geradas pelo modelo (*LLM-as-a-judge*).
MMLU *(Hendrycks et al., 2021)*	Conhecimento geral e capacidade de raciocínio.	Capacidade de raciocínio e conhecimento geral dos modelos de linguagem em áreas diversas, abrangendo um total de 57 disciplinas.	Apresentação ao modelo de mais de 15 mil perguntas de escolha múltipla, que vão desde o nível secundário até a um nível avançado. O desempenho do modelo é calculado com base na percentagem de respostas corretas por disciplina, sendo o resultado o valor médio obtido nas 57 áreas avaliadas.
TruthfulQA *(Lin et al., 2022)*	Capacidade de não alucinar respostas	Grau de veracidade e precisão factual das respostas geradas por um modelo de linguagem, especialmente em contextos onde há risco de desinformação.	Apresenta 817 perguntas distribuídas por 38 categorias, incluindo temas sensíveis como saúde, direito, finanças e política. O objetivo é verificar se o modelo evita responder com informações incorretas ou enganadoras, medindo a tendência para "alucinar" ou gerar factos falsos.
Berkeley Function-Calling Leaderboard (BFCL) *(Fanjia Yan et al., 2024)*	Capacidade dos LLMs em realizar chamadas de funções (*function calling*)	Capacidade dos LLMs em realizar chamadas de funções de forma precisa e eficiente, testando tanto a correção funcional como a adequação da função escolhida ao contexto da tarefa.	Com 2.000 pares pergunta-função-resposta em várias linguagens de programação e contextos de aplicação, como Java, JavaScript, REST API, SQL e Python, os cenários abrangem chamadas simples, múltiplas e paralelas de funções, bem como a deteção da relevância funcional, ou seja, se o modelo escolhe corretamente funções apropriadas para o que foi solicitado.

Desafios/Riscos/Limitações

Como já mencionado anteriormente, os modelos de linguagem apresentam vários riscos que devem ser mitigados aquando da sua utilização e desenvolvimento: **alucinações** - quando a geração de texto que não está de acordo com os *input*s fornecidos, com outputs falsos, mas convincentes, ou seja, alucinando sobre uma realidade inexistente (*OpenAI Research*, 2023). Este problema levanta preocupações de segurança e fiabilidade nas aplicações de modelos de linguagem no mundo real, não só porque pode gerar informação enganosa, como poderá violar a privacidade dos utilizadores (Ji et al., 2023). Torna-se, portanto, crucial abordar e mitigar este risco de maneira a garantir a utilidade e segurança de aplicação de LLMs em diferentes contextos de aplicação; **viés** (*Bias*): tanto na fase de desenvolvimento e treino destes modelos, como o próprio conjunto de dados utilizados para o treino podem enviesar as respostas oferecidas aos utilizadores, comprometendo a sua imparcialidade a vários níveis (Kaneko et al., 2024). Desta forma, deve procurar-se desenvolver estratégias que mitiguem estes riscos, promovendo o igual tratamento de todas as vertentes societárias e de trabalho (Bubeck et al., 2023; *OpenAI Research*, 2023); **eficiência** - o treino e lançamento destes modelos necessitam de elevado poder computacional, resultando numa pegada carbónica bastante elevada. Os modelos têm crescido constantemente, quer em tamanho, quer em capacidades, e muito se deve à infraestrutura por detrás (Jovanović & Campbell, 2022). E uma das formas mais fáceis e óbvias de escalar estes modelos é aumentar o poder computacional. É imperativo procurar soluções mais sustentáveis quando isso for possível, como por exemplo, a utilização de *Small Language Models* (SLM), *fine-tuned* para atividades específicas, ao invés de modelos grandes com elevada generalização; **explicabilidade** - os modelos de linguagem são capazes de gerar texto com base num grande conjunto de dados não estruturados. Contudo, é muitas vezes difícil de perceber as razões que originaram o *output* por eles produzidos, dificultando a sua validação. Este problema está diretamente associado às alucinações. A falta de explicabilidade não permite saber se o modelo está ou não a alucinar. Atualmente, as técnicas existentes não se traduzem num conhecimento suficientemente relevante para combater este problema (Samek et al., 2019); **ética** - estes modelos de linguagem, especialmente os de grande dimensão, são treinados num grande conjunto de dados, muitas vezes, sem quaisquer filtros. Sabendo que estes modelos geram novo conteúdo partindo dos dados que foram treinados e do *input* do utilizador, existe a probabilidade de o modelo gerar outputs indesejados, tendo por base preconceitos, discriminação, exclusão e toxicidade (Weidinger et al., 2021). São vários os exemplos de utilizadores que, propositadamente, levam modelos como o ChatGPT ao seu extremo, com respostas bastante alarmantes. Contudo, existe agora uma elevada consciencialização sobre esta temática, com muitos dos modelos a não responderem em situações que possam gerar esses tipos de respostas; ***accountability*** - é especialmente importante perceber as limitações e desafios proporcionados por estes modelos. Com toda esta revolução a acontecer, vê-se uma tendência para olhar para estes modelos de linguagem como uma solução para todos os males. É importante perceber as suas capacidades e desafios aquando da sua utilização e implementação. **privacidade e segurança** - muitos destes modelos de linguagem são treinados, não só por grandes quantidades de dados, mas também pelas interações com os seus utilizadores. Desta forma, a informação partilhada, passando a fazer

parte do treino dos modelos, pode ser gerada futuramente de forma indesejada (Michael Chui et al., 2023). Este risco torna-se particularmente importante de mitigar pelas organizações, de maneira a não correr o risco de partilhar informação sensível e vê-la futuramente exposta; por fim, **propriedade intelectual** - treinando em grandes porções de dados disponíveis na internet, existe um grande risco destes modelos de linguagem infringirem direitos autorais, marcas, patentes e outros legalmente protegidos, pelo que se torna essencial uma avaliação rigorosa na identificação destas infrações, garantindo compliance (Michael Chui et al., 2023).

Estratégias para melhoria do desempenho de um LLM

São já muitas as técnicas utilizadas para mitigar os riscos mencionados imediatamente acima, procurando uma melhoria nos resultados obtidos por LLMs nas mais variadas tarefas e contextos:

- **Técnicas de *Prompting***: *prompts* consistem nos *input*s dados aos modelos de linguagem pelo utilizador, e têm um grande impacto na qualidade das respostas geradas pelos modelos. Com o objetivo de ultrapassar algumas limitações dos LLMs, nomeadamente, a sua capacidade de raciocinar e agir sobre isso, bem como a segurança na interpretação e compreensão da *query* realizada pelo utilizador, foram propostas algumas técnicas, tais como:

    - *Few-Shot (FS) Learning*: de modo semelhante ao ser humano que é capaz de realizar de forma mais eficiente uma tarefa com base em exemplos, oferecendo exemplos ao modelo, espera-se que produza, também, melhores outputs (Brown et al., 2020);
    - *Chain of Thought Prompting (CoT)*: é providenciado ao modelo passos intermédios a seguir até atingir o objetivo final, possibilitando-o de seguir um raciocínio de modo semelhante ao ser humano (Wei et al., 2023)

    - *Chain of Thought Prompting com Self Consistency (CoT-SC)*: Para além de passos intermédios, é proposto ao modelo a criação de diferentes caminhos possíveis pertante um mesmo *input*, levando a uma resposta mais "informada" e precisa (Wei et al., 2023);

    - *Tree-of-Thoughts (ToT)*: Agrega as técnicas anteriores, permitindo adicionalmente que o modelo volte a passos anteriores à procura de novos e melhores caminhos para a resposta (Yao et al., 2024, p. 2), oferecendo uma maior flexibilidade.

- ***Retrieval Augmented Generation* (RAG)**: RAG pretendem ultrapassar as limitações de LLMs, no que toca a alucinações, explicabilidade, *accountability*, e pelo facto de LLMs terem os seu conhecimento limitado aos dados em que foram treinados. RAG consiste em introduzir um passo extra na interação utilizador-modelo. Em vez de os modelos se basearem nos seus dados (nos dados em que foram treinados) para responderem à *query*, o modelo procura a

informação numa base de dados externa (que poderá conter informação privada da organização, por exemplo), e que irá oferecer o contexto necessário para responder à pergunta de forma precisa e com disponibilização das fontes. Esta nuance permite que os modelos têm acesso à informação mais relevante e atualizada e possibilita, simultaneamente, a verificação dos outputs fornecidos (IBM Research, 2021b; Lewis et al., 2021). Dada a relevância da técnica de *Retrieval-Augmented Generation* (RAG) para a extensibilidade das capacidades dos modelos de linguagem em domínios específicos, especialmente aqueles que exigem elevado rigor técnico e sensibilidade quanto à privacidade dos dados, e considerando também a sua utilidade na promoção da explicabilidade e da factualidade das respostas geradas, o RAG será explorado com maior detalhe no subcapítulo seguinte.

- ***Fine-tuning***: Apesar de LLMs apresentarem um desempenho notável em uma variedade de tarefas, a sua generalização impacta a sua capacidade de especialização. Afinar estes modelos para realizarem atividades bastante específicas, em domínios específicos, pode ser uma solução bastante eficiente, pois o modelo passa a estar alinhado com as suas nuances e complexidades. Para além disso, afinar modelos pré-treinados é bastante menos dispendioso que treinar um modelo do zero. (Banjara, 2023)

- ***Small Language Models (SLMs)***: Embora possa existir um compromisso na capacidade de generalização dos modelos de linguagem com a redução de dezenas de milhares de milhões de parâmetros para modelos de menores dimensões, existem modelos capazes de executar determinados tipos de tarefas com uma grande eficiência, comparável, em certos casos, a LLMs. O lançamento do modelo Phi-2 pela Microsoft, com apenas 2,7 mil milhões, apresentou capacidades de raciocínio e compreensão de linguagem bastante notáveis, igualando-se e superando modelos até 25x maiores em determinadas tarefas complexas (Mojan Javaheripi, 2023);

2.3. *Retrieval Augmented Generation* (RAG)

Conceito e Pipeline

RAG é fundamental pelos vários motivos já mencionados acima. Pretende-se agora entender com maior rigor o funcionamento desta técnica, bem como explorar alguns *frameworks* já existentes neste domínio.

A procura da informação pelo modelo em bases de dados externas é possível através daquilo a que se chama *vector embeddings*, que consistem numa representação numérica que capta a semântica das palavras, suas relações e similaridades (em muito semelhante ao mencionado na camada de *embeddings* da arquitetura dos transformadores, mencionada no subcapítulo anterior).

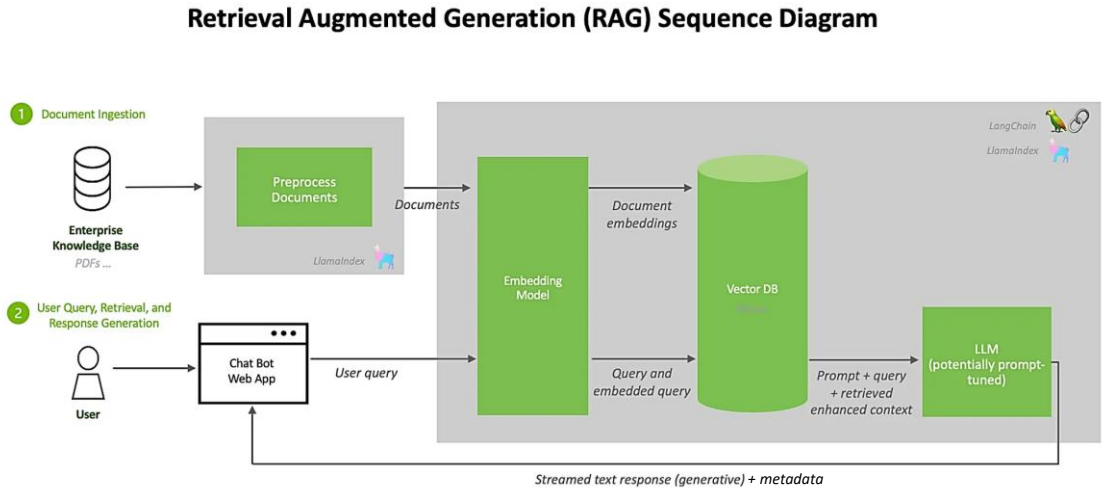


Figura 6 - Pipeline de *Retrieval Augmented Generation* (RAG)

Fonte: https://developer.nvidia.com/blog/rag-101-demystifying-retrieval-augmented-generation-pipelines/

Através destas representações numéricas, é possível fazer comparações matemáticas que possibilitarão estabelecer semelhanças entre os dados. Na sua essência, estes vetores representam dados (textuais ou outros) matematicamente. Estes vetores são agrupados em bases de dados (*vector databases*), o que permite, tal como outras bases de dados, realizar *query*s e obter informações. Desta forma, documentos de grandes dimensões são divididos em *chunks*, agrupados nestas bases de dados e, de seguida, "chamados" para estes modelos, respeitando aquilo que são as limitações em termos de tamanho do contexto máximo admissível (*context length*) e, simultaneamente, oferecendo a informação mais relevante ao modelo, sem este ter de analisar os documentos na sua totalidade. Isto reduz, drasticamente, o poder computacional necessário.
Genericamente, um processo de RAG (Figura 6) é composto pelas seguintes fases (Databricks, 2023; Hayden Wolff, 2023; IBM Research, 2021b; Lewis et al., 2021):

1. **Ingestão de Documentos**: tudo começa com a ingestão de dados brutos provenientes de diversas fontes: bases de dados, documentos, websites, emails ou até dados através de APIs;

2. **Pré-processamento dos Documentos**: depois de carregados, os documentos passam por um processo de transformação. Um dos passos cruciais aqui é o fragmentação de texto (*text splitting*), que divide conteúdos longos em segmentos menores, chamados *chunks*. Isto é necessário pois os modelos de *embedding* têm um limite de *tokens* que conseguem processar de uma só vez (o limite do modelo de embedding da OpenAI "*text-embeddings-3-large*", por exemplo, é de 8191 *tokens* (Open AI Platform, 2025)). A escolha do tamanho dos *chunks* é crítica: se forem demasiado grandes, podem conter informação pouco específica; se forem demasiado pequenos, podem perdem coerência semântica, ou seja, informação relevante pode ser "cortada ao meio".

3. **Geração de *Embeddings***: com os textos já divididos, o próximo passo é transformá-los em vetores numéricos, ou seja, em *embeddings*. Como mencionado anteriormente, esta etapa consiste em aplicar um modelo de *embedding* que representa semanticamente o texto num espaço vetorial multidimensional. Assim, textos semelhantes ocupam posições próximas nesse espaço vetorial, o que facilita as buscas por similaridade semântica mais à frente;

4. **Armazenamento em Base de Dados Vetorial**: os *embeddings* resultantes são armazenados numa base de dados vetorial que permite procurar rapidamente informação relevante com base em semelhança semântica. Estas bases de dados são projetadas para fazer buscas eficientes em dados vetorizados e devem ser atualizadas regularmente para garantir que o conhecimento disponível está sempre atualizado;

5. ***Query***: quando o utilizador submete uma pergunta ou *prompt*, o sistema converte essa entrada num vetor e realiza uma busca semântica na base de dados vetorial. O objetivo passa, então, por identificar *chunks* de texto relevantes, aqueles mais próximos ao vetor da consulta, para construir uma base de contexto para fornecer ao modelo (juntamente com a *query* do utilizador, como mencionado de seguida);

6. **Engenharia do *Prompt* (*Prompt Augmentation*)**: com base nos dados recuperados da base dados vetorial, o sistema constrói um *prompt* aumentado, que é passado ao modelo de linguagem. Este *prompt* combina a pergunta original com os conteúdos mais relevantes retirados da base de dados.

7. **Geração da Resposta**: Por fim, o *prompt* aumentado é fornecido ao modelo de linguagem que gera uma resposta. Desta forma, ainda que a resposta tenha sido criada por um modelo generativo, é agora fundamentada no conteúdo recuperado pela base de dados externa, tornando-se assim mais factual, específica e explicável.

São várias as frameworks já existentes que realizam o *heavylifting* de todo este processo. Langchain é uma das mais conhecidas, *open source*, possui um vasto conjunto de funcionalidades que facilitam a integração de bases de dados vetoriais externas com modelos de linguagem, tendo sido um dos pioneiros a colmatar este *gap*. Llama Index é também bastante popular entre os desenvolvedores de sistemas RAG, também *open source* e com forte capacidade na execução das etapas mencionadas acima, facilitando a transformação de dados estruturados e não estruturados, provenientes de inúmeras fontes e formatos em dados prontos para serem alimentados no LLM. Outros como Haystack, desenvolvido pela Deepset ou REALM, ferramenta da Google (Guu et al., 2020) são só algumas das inúmeras ferramentas que foram desenvolvidas para ajudar desenvolvedores a construírem pipelines de recuperação de informação de forma bastante facilitada e eficiente.

Também são inúmeros os fornecedores de bases de dados vetoriais, cada uma com as suas forças e limitações, variando em termos de desempenho, escalabilidade, facilidade de uso e integração com outros sistemas (Figura 7).

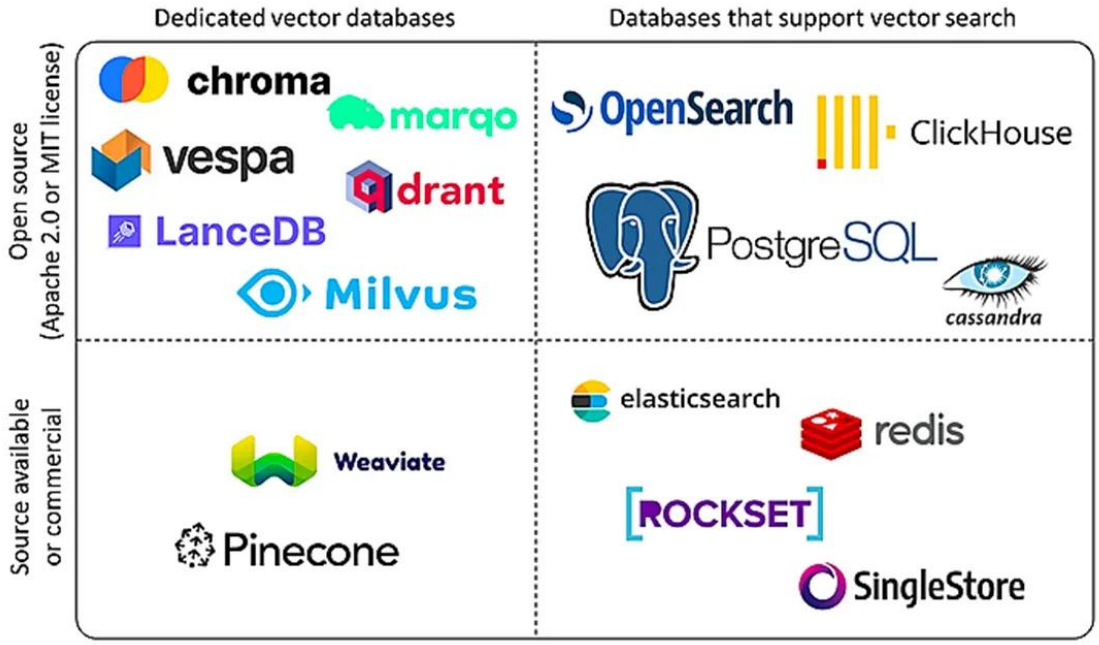


Figura 7 - Mercado das Bases de Dados Vetoriais.

Fonte: (Wu, 2023)

O FAISS (Hervé Jegou et al., 2017), por exemplo, oferece desempenho elevado, sendo ideal para tarefas onde alta eficiência é essencial, mas exige um nível técnico considerável para sua configuração e manutenção. Já o Milvus (J. Wang et al., 2021) se destaca pela sua escalabilidade, tornando-se uma excelente escolha para grandes volumes de dados distribuídos, e oferece uma versão completamente gerenciada em *cloud*, fornecida pela Zilliz, facilitando sua adoção em ambientes empresariais. Pinecone, por outro lado, é reconhecido pela sua facilidade de uso e por ser otimizado para aplicações em tempo real, com baixa latência, além de contar com inúmeras integrações com diferentes fornecedores e plataformas. Weaviate, também voltada para tarefas de pesquisa semântica, oferece uma API simples e intuitiva, tornando-se uma opção atrativa para projetos que buscam praticidade sem comprometer a funcionalidade. Chroma DB, uma base de dados vetorial *open source*, é especialmente otimizada para o armazenamento e recuperação de *embeddings*, com foco em pesquisa semântica, proporcionando uma solução eficiente na execução dessa tarefa. Por fim, o MongoDB, embora não seja uma base de dados vetorial nativa, é um banco de dados NoSQL amplamente utilizado para armazenar grandes volumes de dados não estruturados. Com a adição de extensões, o MongoDB pode ser adaptado para armazenar *embeddings* e realizar buscas de similaridade. Ainda assim, em termos de desempenho e otimização, não se compara diretamente a soluções específicas para dados vetoriais, como as mencionadas anteriormente.

A escolha da base de dados ideal, portanto, depende das necessidades específicas do projeto, incluindo o tipo de dados a serem processados, a escala da aplicação e a complexidade da implementação (Jim Holdsworth & Matthew Kosinski, 2024).
Tal como acontece na escolha de LLMs, também aqui é importante avaliar a qualidade do pipeline desenvolvido e de que maneira a escolha do framework, a interação com o LLM, a base de dados vetorial escolhida e a qualidade da fase de *embeddings* e recuperação impactam os resultados obtidos (IBM Research, 2021b).
De facto, pode parecer aliciante a utilização de RAG face às promessas de aproveitar o melhor dos dois mundos: as potencialidades dos LLMs e consequente ganho de produtividade e competitividade, com a segurança, privacidade, conhecimento de domínio externo e consequente explicabilidade dos *outputs*. Ainda assim, desenvolver um sistema RAG com qualidade nem sempre é fácil. Algumas das principais dificuldades que surgem são as seguintes (Ismail Eruysaliz, 2024; Johnson et al., 2023; Willsmore, 2024):

- **Falta de conteúdo necessário nas bases de dados**: A ausência de informações essenciais às *queries* dos utilizadores pode fazer com que o LLM forneça respostas incorretas;

- **Dificuldade na extração da informação relevante**: o modelo pode falhar ao extrair a resposta correta, especialmente em contextos com informações conflitantes, ruído, métodos de *ranking* desadequados, *chunking* desadequado, perda de contexto, etc. → resultando em *outputs* erróneos ou incompletos;

- **Escalabilidade na ingestão de dados**: grandes volumes de dados podem sobrecarregar o pipeline de ingestão, prejudicando a capacidade do sistema de gerenciar e processar os dados de maneira eficiente. Daí a importância na escolha de um framework e BD adequada;

- **Documentos com estruturas complexas**: a utilização de documentos em PDF é bastante comum em vários tipos de trabalho. A extração de dados de PDFs complexos, com tabelas e gráficos incorporados, exige uma lógica de análise sofisticada devido à inconsistência nos *layouts* e formatos;

- **Execução de consultas de múltiplas etapas, com necessidade de tarefas e ferramentas distinta**s: consultas complexas podem exigir múltiplas etapas de consulta e a orquestração de *prompts* diferentes. Neste caso, entramos no território dos agentes, com diferentes chamadas aos LLMs, a executarem tarefas diferentes, a partilharem contexto entre eles, etc.;

- **Contextualização e estrutura de dados**: agrupar os dados em hierarquias de documentos ou grafos de conhecimento pode melhorar a precisão da recuperação (Edge et al., 2025), mas esses métodos podem ser difíceis de implementar e exigem recursos intensivos.

2.4. Model Context Protocol (MCP)

Model Context Protocol, ou MCP, desenvolvido pela Anthropic em novembro de 2024 (Anthropic, 2024), é um protocolo de comunicação aberto e modular que estabelece uma linguagem comum entre modelos de IA, servidores especializados e aplicações. A sua conceção foi inspirada em princípios semelhantes aos protocolos de rede e APIs, mas adaptados para o ecossistema da IA generativa, que consiste, essencialmente, em padronizar uma forma de conexão entre aplicações IA e serviços externos.

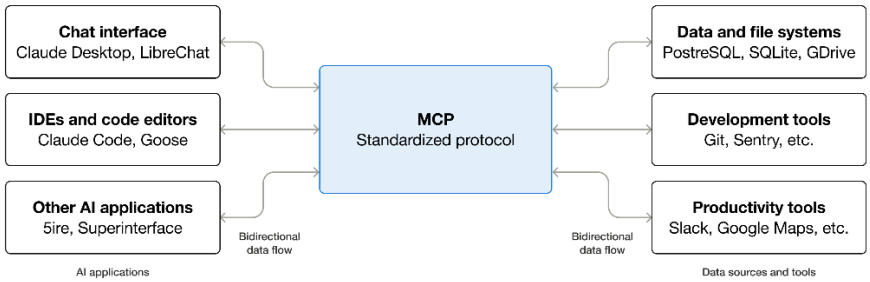


Figura 8 - *Model Context Protocol* (MCP)

Fonte: https://modelcontextprotocol.io/docs/getting-started/intro

Em termos funcionais, o MCP define três camadas principais (Nebius, 2025):

1. **Camada de transporte** – garante a comunicação segura entre o cliente (p. ex., um *frontend* de texto) e o servidor (como repositórios ou bases de dados biomédicas). Esta camada abstrai os detalhes do canal de comunicação, suportando protocolos como HTTP, stdio ou WebSockets, com base em JSON-RPC 2.0;
2. **Camada de protocolo** – define a estrutura das mensagens trocadas (requisições, respostas, eventos) e padroniza os formatos de entrada e saída (*inputs/outputs*). Serve como uma interface formal entre cliente e servidor, assegurando interoperabilidade;
3. **Camada de aplicação** – contém os recursos e ferramentas (*tools*) acessíveis ao modelo. É neste nível que se encontram os utilitários expostos pelo servidor, como motores de busca, bases de dados, calculadoras ou outras APIs, e que o cliente pode invocar conforme necessário.

Existem inúmeras vantagens na utilização de servidores MCP (Anthropic, 2025). Para os desenvolvedores, reduzem significativamente o tempo e a complexidade associados ao desenvolvimento ou integração de aplicações e agentes de IA. No que diz respeito a aplicações e agentes de IA, os MCP disponibilizam acesso a um ecossistema diversificado de fontes de dados, ferramentas e aplicações, ampliando as capacidades e melhorando a experiência do

utilizador. Já para os utilizadores finais**,** possibilitam aplicações e agentes de IA mais robustos, capazes de aceder aos seus dados e executar ações em nome do utilizador, sempre que necessário.

No domínio da FC, a aplicabilidade do MCP revela-se particularmente vantajosa. A elaboração de relatórios técnicos exige acesso a múltiplas fontes de evidência, nomeadamente, literatura biomédica indexada (PubMed) e diretrizes de sociedades médicas e científicas. A integração destas fontes, tradicionalmente realizada de forma manual pelos farmacologistas clínicos, pode ser significativamente acelerada por um sistema baseado em LLMs com suporte MCP.

2.5. Inteligência Artificial na Farmacologia Clínica

Farmacologia Clínica: Definição e Objetivos

A FC é uma especialidade médica que investiga e avalia os efeitos dos fármacos no organismo humano, tanto em indivíduos saudáveis quanto doentes (Beigel et al., 2019; Rawlins, 2013). Esta disciplina abrange todos os aspetos da interação entre os medicamentos e os seres humanos, sendo uma ciência multidisciplinar que envolve médicos, farmacologistas, farmacêuticos, cientistas biomédicos e enfermeiros (Dr. Luís Almeida, 2020). O principal objetivo da FC é melhorar os cuidados prestados aos pacientes por meio de um uso mais seguro e eficaz dos medicamentos. Isso inclui, entre outros aspetos, a personalização da medicação de acordo com as características individuais do paciente, a aplicação de testes farmacogenéticos que ajudam a prever como o corpo reage a certos fármacos, o desenvolvimento de novos tratamentos farmacológicos (Choi et al., 2021; Merry & Flexner, 2008), tudo isto com o intuito de aumentar a eficácia dos tratamentos (Merry & Flexner, 2008).

Importância da FC na prática médica e no desenvolvimento de novos tratamentos

A importância da FC na prática médica é imensurável, pois permite que os tratamentos sejam ajustados de forma mais precisa e personalizada, o que resulta em melhores *outcomes* para os pacientes. Esta área torna-se particularmente crucial quando se considera o crescente número de medicamentos utilizados em polifarmácia e a necessidade de garantir que esses tratamentos sejam eficazes e seguros. No contexto de doenças complexas, como VIH ou até cancros, o trabalho do farmacologista clínico é essencial para a escolha da terapia mais adequada e para o ajuste das dosagens dos medicamentos, por forma a evitar falhas no tratamento ou o desenvolvimento de resistência medicamentosa (Gulmez et al., 2020).
Além disso, a FC desempenha um papel fundamental no desenvolvimento de novos medicamentos, envolvendo o estudo de biomarcadores, farmacocinética e farmacodinâmica, e também a realização de ensaios clínicos para testar a segurança e eficácia de novos compostos. Já em termos de gestão de custos de saúde, farmacologistas clínicos têm demonstrado que sua atuação pode resultar em economias significativas para os sistemas de saúde, promovendo o

uso racional de medicamentos e evitando prescrições inadequadas ou ineficazes, o que reduz os custos associados a reações adversas, hospitalizações e falhas no tratamento (Falconer et al., 2021).

Aplicações de IA existentes específicas para o domínio

Atualmente, a literatura sobre aplicações de IA especificamente direcionadas para a FC ainda é limitada, não tendo sido possível identificar um corpo de trabalhos vasto ou consolidado nesta área. Entre os poucos exemplos já existentes, destaca-se o MediAlbertina (Nunes et al., 2024), um *encoder* treinado em português com enfoque biomédico, que representa um avanço relevante para o processamento e análise de informação clínica no contexto nacional, modelo este que será explorado em capítulos posteriores. Outro exemplo é o *OpenEvidence* (Intelligence Artificielle & la Néphrologie, 2025), uma iniciativa que procura sistematizar e organizar evidência científica de forma automatizada, facilitando a consulta de dados biomédicos para suporte à decisão clínica. Apesar de ser um projeto interessante a explorar mencionado pela própria equipa da UFC, é importante para a mesma uma independência desta plataforma externas, por questões de confidencialidade dos dados sensíveis de pacientes que utilizam no processo de elaboração destes relatórios.

Processo de tomada de decisão na área da FC

O processo de tomada de decisão terapêutica representa um eixo fundamental da atividade do farmacologista clínico, exigindo uma análise minuciosa da informação disponível e uma integração criteriosa de múltiplas dimensões clínicas, farmacológicas, regulatórias e éticas (Figura 9).

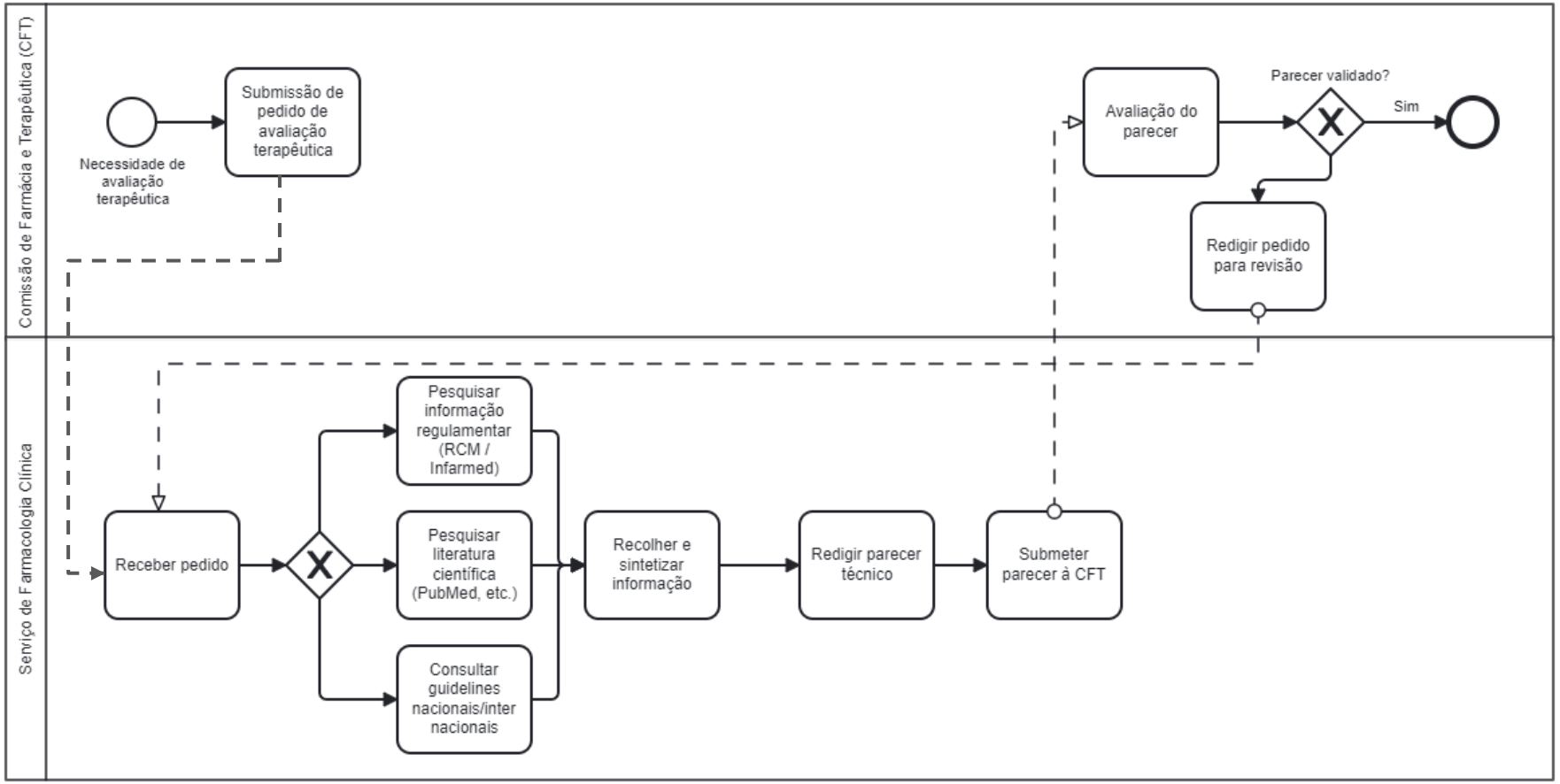


Figura 9 - Diagrama BPMN do Processo de Avaliação em Farmacologia Clínica

O potencial da IA no processo de tomada de decisão

LLMs e sistemas RAG oferecem um potencial significativo na otimização e aceleração dos processos de apoio à decisão clínica, com a determinante nuance de não comprometer a qualidade ou o rigor da análise. Os sistemas RAG funcionariam, essencialmente, através da integração entre dois mecanismos complementares:

1. Recuperação (*Retrieval*): acesso automático e seletivo a documentos relevantes provenientes de bases de dados estruturadas (e.g., literatura científica, *guidelines*, documentos regulatórios);

2. Geração (*Generation*): síntese e apresentação da informação recuperada sob a forma de um texto coerente, estruturado e adaptado ao contexto clínico ou decisório específico.

Prevê-se que a adoção destes sistemas neste âmbito (desde que devidamente treinados com dados contextuais e validados cientificamente), possa resultar em:

- Redução significativa do tempo de elaboração de pareceres, permitindo respostas mais céleres a necessidades clínicas urgentes;
- Padronização e uniformização da estrutura dos relatórios técnicos, com ganhos imensuráveis de eficiência e produtividade;
- Melhoria da qualidade da decisão, ao garantir o acesso sistemático à evidência mais recente e relevante;
- Potenciação do raciocínio clínico, ao libertar os profissionais das tarefas mais mecânicas de triagem e resumo de informação, focando nas atividades em que geram mais apurado, como na análise crítica e interpretação contextual do caso clínico;
- Preservação da segurança e confidencialidade da informação clínica, desde que os modelos sejam operados em ambientes informáticos fechados e validados, respeitando os princípios da ética biomédica e da proteção de dados.

Importa, contudo, sublinhar que o recurso a estes sistemas não tem como objetivo a substituição dos profissionais, mas antes funcionar como uma ferramenta de amplificação cognitiva, ao serviço da tomada de decisão baseada na evidência.

Principais aliados da sinergia IA – FC: Pesquisa Semântica e Sumarização

Tendo em conta o âmbito da presente dissertação, existe especial interesse na aplicação prática de modelos de linguagem em contextos de recuperação e síntese de informação, tornando-se, por isso, particularmente relevante olhar com atenção para as tarefas de pesquisa semântica e sumarização. Estas duas áreas refletem necessidades reais em diversos domínios. Assim, a escolha dos *benchmarks* corretos para avaliar o desempenho dos modelos nessas tarefas é essencial para garantir a adequação e eficácia das soluções propostas. Consequentemente, esta escolha influencia também a seleção dos LLMs mais apropriados, privilegiando-se aqueles que demonstram melhor desempenho em métricas específicas de relevância semântica, coerência e fidelidade factual.

Dada a natureza altamente sensível e técnica da área da FC, a escolha dos *benchmarks* mais adequados para avaliação de modelos de linguagem assume uma importância crítica. Neste domínio, onde erros factuais podem ter implicações clínicas sérias, é essencial garantir que os modelos utilizados não apenas demonstrem competência linguística, mas também rigor científico, precisão factual e capacidade de sintetizar conhecimento especializado a partir de fontes fiáveis.

Neste contexto, quatro *benchmarks* destacam-se pela sua relevância: MMLU (Massive Multitask Language Understanding), TruthfulQA, PubMedQA (Jin et al., 2019) e BioASQ (Nentidis et al., 2023). O MMLU permite avaliar a profundidade do conhecimento geral e técnico de um modelo, cobrindo disciplinas como medicina, biologia e química, que são altamente pertinentes à farmacologia. Já o TruthfulQA avalia diretamente a veracidade das respostas, ajudando a detetar "alucinações".

Por outro lado, *benchmarks* como o PubMedQA e o BioASQ são ainda mais especializados, uma vez que se baseiam diretamente em literatura científica biomédica, nomeadamente, em artigos indexados no PubMed. O PubMedQA centra-se na capacidade do modelo de responder a perguntas clínicas com base em abstracts científicos, promovendo o raciocínio baseado em evidência. O BioASQ, por sua vez, oferece um conjunto mais abrangente de tarefas, que inclui respostas a perguntas abertas, sumarização e recuperação de informação, sendo particularmente relevante para avaliar competências de pesquisa semântica e síntese de conhecimento biomédico.

Assim, tendo em conta os objetivos desta dissertação, que envolvem a análise de modelos de linguagem no contexto da FC, será dada especial atenção aos *benchmarks* PubMedQA e BioASQ, tanto pela sua orientação biomédica, como pela capacidade de testar diretamente as competências de sumarização e compreensão semântica de documentos técnicos. O MMLU e o TruthfulQA serão também considerados como instrumentos complementares, oferecendo uma visão mais ampla da robustez e fiabilidade geral dos modelos avaliados. Os leaderboards atualizados em tempo real dos respetivos *benchmarks* encontram-se em:

- **PubMedQA**: https://pubmedqa.github.io/
- **BioASQ**: https://paperswithcode.com/sota/question-answering-on-bioasq
- **MMLU**: https://huggingface.co/spaces/StarscreamDeceptions/Multilingual-MMLU-*Benchmark*-Leaderboard
- **TruthfulQA**: https://paperswithcode.com/sota/question-answering-on-truthfulqa

Tabela 2 - *Benchmarks* úteis na avaliação de modelos para a área da Farmacologia Clínica

Benchmark	Área de avaliação	Formato de Tarefa	Domínio Específico	Relevância para FC	Pontos Fortes	Limitações
MMLU	Conhecimento geral e técnico em múltiplas disciplinas	Escolha múltipla	Amplo (inclui biologia, medicina, química)	Alta – cobre tópicos relevantes de ciências da vida e raciocínio clínico	Avalia profundidade de conhecimento em áreas complexas; boa generalização	Não é exclusivamente biomédico; perguntas não são extraídas de literatura científica
TruthfulQA	Veracidade e precisão factual	Perguntas abertas	Amplo (inclui saúde, política, ciência)	Alta – deteta "alucinações" factuais, muito relevante em contexto clínico	Excelente para testar se o modelo gera informações incorretas	Foco mais genérico, com menor profundidade técnica em tópicos médicos
PubMedQA	Capacidade de raciocínio clínico baseado em abstracts	Sim/Não/Indeterminado	Biomédico (PubMed)	Muito alta – usa dados reais da literatura médica	Baseado em literatura científica real; muito focado na linguagem clínica	Formato limitado a perguntas com resposta Sim/Não; não cobre sumarização
BioASQ	Resposta a perguntas factuais e sumarização biomédica	QA aberta + Sumariação + Informação estruturada	Biomédico	Muito alta – ideal para sumarização de artigos e recuperação de informação	Abrange sumarização, pesquisa semântica e QA com base em dados biomédicos	Complexidade elevada; requer modelos treinados em terminologia biomédica

3. Relatório Técnicos UFC

O sistema proposto tem o intuito de automatizar grande parte do processo de elaboração dos pareceres técnicos redigidos pela UFC. Por forma a entendermos como isso se processa, é importante, numa primeira instância, perceber qual é a sua estrutura final e qual o conteúdo que compõe cada uma das suas secções.
Analisando o histórico de pareceres realizados pelos profissionais da UFC nos últimos 3 anos, verifica-se que estes seguem uma estrutura relativamente padronizada, garantindo consistência entre relatórios, rigor e clareza. Essa estrutura é composta pelos seguintes capítulos:

1. **Enquadramento**: Esta secção contém a contextualização do pedido submetido pela CFT. É definido o âmbito do parecer, explicitando a questão clínica ou terapêutica que lhe dá origem (por exemplo, avaliação da utilização de um fármaco num determinado contexto/caso clínico). Fornece-se o enquadramento normativo, científico e institucional, salientando a relevância da decisão a tomar.

2. **Caso Clínico:** Aqui descreve-se a situação clínica concreta do doente que motivou o pedido. Incluem-se dados demográficos essenciais (idade, sexo, antecedentes pessoais relevantes), registam-se os antecedentes patológicos e terapêuticos, bem como a evolução clínica recente e apresentam-se exames complementares de diagnóstico pertinentes, diagnósticos confirmados e critérios de elegibilidade para a terapêutica em análise. Esta secção garante que a avaliação do medicamento ou da estratégia terapêutica se ancora em necessidades clínicas reais e individualizadas ao doente em questão.

   Nota: nem sempre esta secção está presente. Poderá haver casos em que o pedido em questão não envolve um doente em específico, envolvendo, uma pesquisa mais generalizável.

3. **Fundamentação:** núcleo científico do parecer, este capítulo é organizado em subsecções que estruturam a análise crítica da evidência disponível:

   - ***Farmacologia***: esta secção contém a descrição dos principais aspetos farmacológicos do medicamento em avaliação: mecanismo de ação, farmacocinética, farmacodinâmica, vias de administração, esquema posológico e perfil de segurança, integrando a informação disponível nos Resumos das Características do Medicamento (RCM) e de outras fontes regulamentares (como, por exemplo, as informações de prescrição do fármaco com as indicações terapêuticas para as quais este foi aprovado pela *Foods and Drugs Administration*). Menciona, também, potenciais interações medicamentosas e considerações práticas de utilização em contexto clínico real, etc.

   - ***Estudos Relevantes***: nesta secção, é feita uma análise crítica da evidência científica publicada relativamente à eficácia e segurança do fármaco, sempre com a aplicabilidade ao caso clínico em análise. Faz-se uma síntese dos principais ensaios clínicos randomizados, revisões sistemáticas e meta-análises, com destaque para indicadores clínicos de relevância (como, por exemplo, sobrevivência global, resposta clínica, qualidade de vida, entre outros).

   - ***Pareceres de sociedades médicas e científicas***: nesta secção é feita uma revisão das recomendações emitidas por sociedades científicas nacionais e internacionais, bem como por entidades regulatórias (ex.: EMA, FDA, INFARMED). Integram-se as orientações clínicas mais recentes publicadas por sociedades médicas de referência, tais como a *European Society for Medical Oncology* (ESMO), o *National Institute for Health and Care Excellence* (NICE), a *American Society of Clinical Oncology* (ASCO), a *European Crohn's and Colitis Organisation* (ECCO), entre outras.

     Nota: As entidades/sociedades consultadas variam substancialmente consoante a área terapêutica em causa. Desta forma, em cada parecer, são selecionadas as diretrizes mais pertinentes, garantindo que a fundamentação segue as melhores práticas internacionais e se adapta ao contexto clínico específico.

4. **Considerações Finais:** nesta secção procede-se a uma integração sintética dos elementos discutidos nos capítulos anteriores. É feita uma avaliação global da relação custo-benefício do medicamento no contexto clínico específico e/ou discussão de alternativas terapêuticas viáveis e/ou considerações práticas para a monitorização da resposta ao tratamento e gestão de efeitos adversos (depende bastante do pedido realizado).

5. **Conclusão:** corresponde ao momento decisório do parecer, onde se emite uma recomendação clara e fundamentada por parte da equipa da UFC, com base na evidência recolhida até ao momento relativamente à aprovação ou não da utilização do medicamento na situação clínica em questão, bem como da definição de eventuais condições associadas à decisão (ex.: critérios de monitorização, critérios de suspensão, necessidade de reavaliação periódica, …).

6. **Referências Bibliográficas:** Listagem (em estilo APA) de todas as fontes de informação utilizadas na redação do parecer.

Processamento do Histórico de Pareceres

Como mencionado acima, todos os pareceres possuem a mesma estrutura, com as diversas secções bem definidas. É possível verificar um exemplo de um parecer redigido pela UFC no Anexo A. Importante mencionar que foram removidos todos os dados sensíveis (informações pessoais do paciente) para que não sejam alimentados aos modelos de linguagem em nenhum momento, garantindo, assim, a confidencialidade tão importante para os profissionais e doentes.
Este histórico de pareceres foi ainda categorizado de maneira a poder ser utilizado como referência na elaboração de novos relatórios. Segundo os profissionais da UFC, os pedidos da CFT podem agrupar-se em 6 grandes grupos:

1. Eficácia e segurança de um medicamento vs comparadores ou outras terapêuticas disponíveis.
2. Evidência científica do medicamento em determinado contexto clínico/doença;
3. Opções terapêuticas no tratamento de determinada doença;
4. Evidência de eficácia e/ou segurança no tratamento de determinada doença vs *standard of care;*
5. Critérios de monitorização de um fármaco;
6. Critérios de iniciação/suspensão de um fármaco.

Esta categorização poderá ser consultada no Anexo B.

3.1. Automação do Processo de Elaboração dos Relatórios Técnicos (UFC)

Tradicionalmente, o processo de elaboração de pareceres técnicos por parte da UFC baseia-se na análise manual de evidência científica, revisão das bulas de medicamentos (RCM) e integração crítica, clara e coesa da informação para responder a questões clínicas e terapêuticas específicas colocadas pela CFT. Com a introdução das inúmeras tecnologias já mencionadas acima (*Large Language Models*, sistemas de *Retrieval-Augmented Generation* e *Model Context Protocol Servers*, procura-se avaliar a possibilidade de automatizar e acelerar grande parte das etapas, mantendo o rigor científico exigido. Assim, o diagrama apresentado de seguida (Figura 10 e Figura 11) ilustra, precisamente, a forma como estes componentes se poderão integrar com o fluxo de trabalho entre a CFT e a UFC.

Comecemos, então, por descrever em que consistem cada uma das etapas para passarmos, seguidamente, à sua implementação (a descrição com terminologia mais técnica será feita em capítulos posteriores).

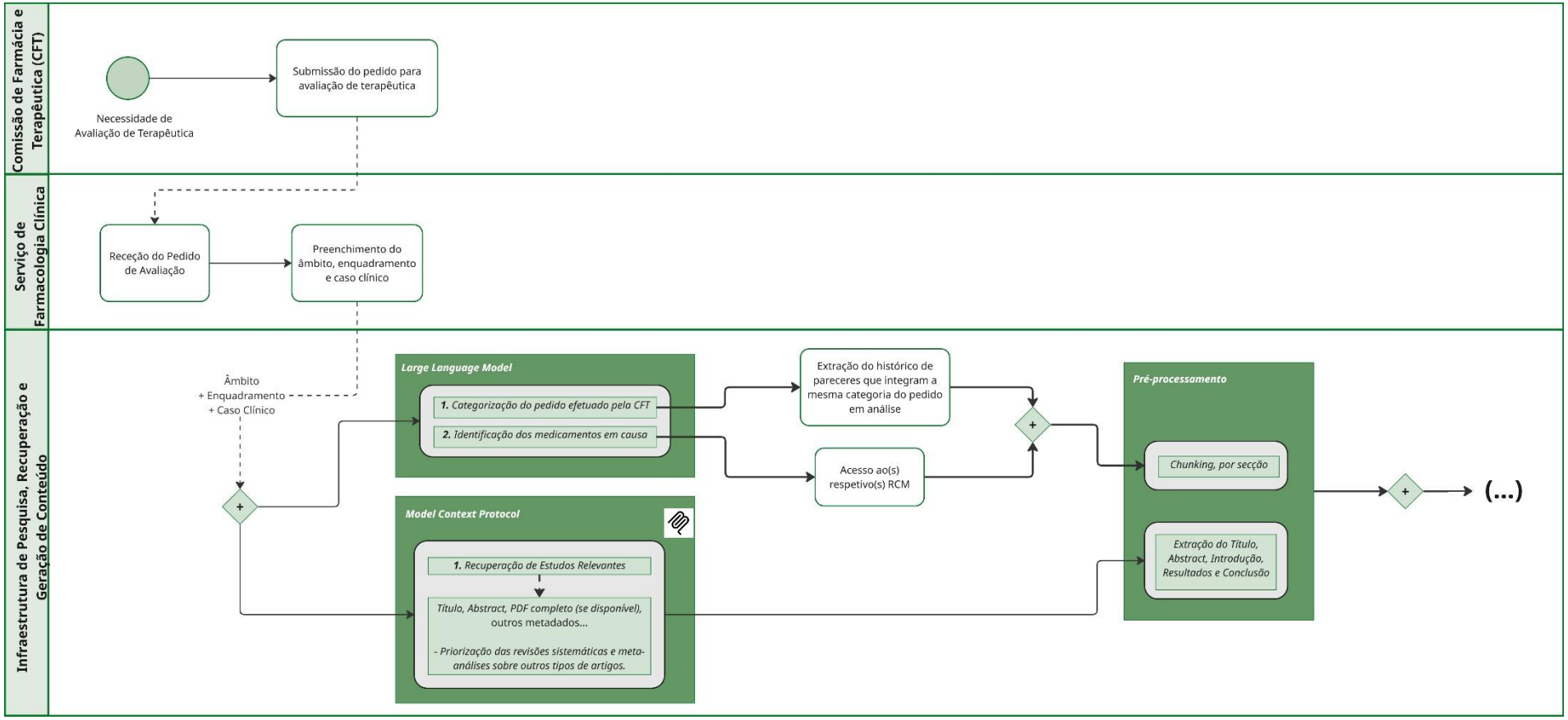


Figura 10 – Automação do Processo de Elaboração de Relatórios com Infraestrutura de Recuperação e Geração de Conteúdo (1)

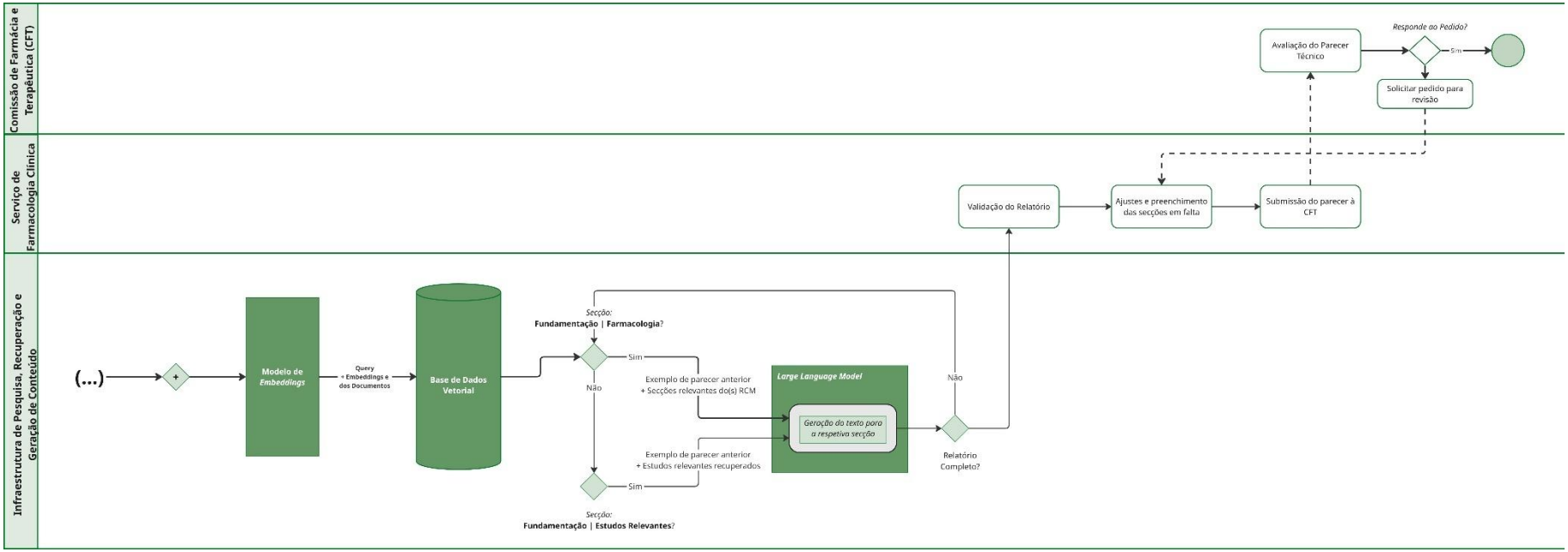


Figura 11 - Automação do Processo de Elaboração de Relatórios com Infraestrutura de Recuperação e Geração de Conteúdo (2)

O processo inicia-se com a necessidade de avaliação terapêutica, formalizada pela CFT através da submissão de um pedido. Este pedido é constituído pelos seguintes elementos fundamentais (como já mencionado no capítulo 3:

- **Âmbito**: especifica o tipo de decisão terapêutica a suportar (e que dará título ao relatório final);
- **Enquadramento**: descreve o contexto clínico e institucional;
- **Caso clínico**: fornece os detalhes individuais do doente em questão (se aplicável. Por vezes, os pedidos são generalizados e não relativos a um doente específico. Nestes casos, o pedido não contém o caso clínico).

Estas 3 variáveis serão a base para todo o trabalho realizado a seguir. Comummente, são a base para os profissionais de farmacologia realizarem a sua pesquisa e elaborarem o relatório. Para o caso da automação aqui descrita, serão utilizados como os principais *inputs* a serem introduzidas no sistema, que os utilizará para orientar a pesquisa e geração dos respetivos conteúdos subsequentes.

Introduzidos os *inputs*, chegamos ao primeiro ponto de intervenção dos modelos de linguagem, que são responsáveis por categorizar do pedido, identificando a área terapêutica e a natureza da questão colocada. Uma das fases do processamento dos relatórios técnicos (descrita no subcapítulo 4.1) consiste precisamente nesta categorização, que agrupa os pedidos da CFT em classes homogéneas. Esta etapa tem como objetivo permitir que, em relatórios subsequentes, seja possível fornecer exemplos representativos aos modelos de linguagem, garantindo que estes reproduzem não apenas a terminologia adequada, mas também uma estrutura semelhante. Assim, promove-se uma padronização progressiva dos relatórios elaborados (em termos técnicos, falamos de de *few-shot prompting*). Para além disso, o LLM é solicitado a identificar o(s) medicamento(s) em análise**,** mais concretamente, o nome comercial do medicamento e/ou substância ativa. Isto permite uma pesquisa focada no fármaco em questão, com a recuperação dos documentos que reúnem as suas características, nomeadamente, o ”Resumo das Características do Medicamento” (documento este que consta na base de dados da EMA (*European Medicines Agency*) e INFARMED (Instituto Nacional da Farmácia e do Medicamento)), bem como o documento de informação de prescrição (*HIGHLIGHTS OF PRESCRIBING INFORMATION*) que consta na base de dados da FDA (*Foods and Drugs Administration*).

Este processamento inicial possibilita que o sistema direcione a pesquisa de forma mais eficiente, assegurando a recuperação de informação historicamente relevante.

No que respeita à categorização do pedido, a seleção do parecer histórico utilizado como exemplo é realizada com base na identificação daquele cuja secção de “Enquadramento” apresenta maior proximidade semântica ao novo pedido. Essa proximidade é determinada através do cálculo da similaridade do cosseno, metodologia descrita em detalhe mais à frente, no Subcapítulo 4.1.1, aquando do processo de geração de *embeddings*.

Seguidamente, passamos à fase de pré-processamento do RCM e parecer de referência (subcapítulo 4.1). O “Resumo das Características do Medicamento” constitui a principal fonte

para a construção da secção *"Fundamentação: Farmacologia"*. Este documento é segmentado em *chunks*, de modo a possibilitar a sua posterior indexação numa base de dados vetorial e facilitar a recuperação semântica de passagens relevantes. Já o parecer de referência, isto é, o relatório histórico selecionado a servir como exemplo, é igualmente dividido nas suas secções constituintes. Este seccionamento permite, como já mencionado acima, que o sistema recorra a trechos específicos durante a fase de geração, assegurando consistência no estilo e na estrutura dos relatórios. (Ex.: para a geração da secção "Considerações Finais", utiliza essa mesma secção do relatório exemplo como inspiração de estrutura, linguagem e tom).

Segmentados os documentos nos respetivos *chunks*, ocorre a indexação dos documentos processados numa base de dados vetorial e pesquisa semântica (subcapítulo 4.1.1). Nesta etapa, é utilizado um modelo de *embedding* para converter os documentos previamente segmentados em representações vetoriais capazes de capturar a sua semântica. Assim, sempre que o sistema necessita de gerar texto para uma determinada secção do relatório, é formulada uma *query* específica que é comparada com os vetores armazenados, através de métricas de similaridade. Este procedimento permite, então, identificar os segmentos mais relevantes.

Paralelamente a este processo de indexação, é feita uma pesquisa no PubMed por parte de um *MCP Server* (Subcapítulo 4.2). Nesta etapa, o *MCP Server* escolhido atua como intermediário entre o sistema desenvolvido e as bases de dados bibliográficas externas, em particular o PubMed. Fá-lo da seguinte forma:

- Traduz a necessidade de informação numa *query* estruturada, adequada aos operadores de pesquisa do PubMed;
- Submete a pesquisa em tempo real, assegurando a recuperação dos artigos mais relevantes;
- Filtra e organiza os resultados, extraindo metadados essenciais (título, autores, ano, *abstract*, PMID) e priorizando publicações de maior nível de evidência (ex.: revisões sistemáticas, meta-análises, ensaios clínicos randomizados).

Com todos os documentos processados e a pesquisa pela evidência científica mais relevante feita, chegamos à fase de geração de conteúdo (Subcapítulo 5.2):

1. **Secção de "Fundamentação | Farmacológica"**: redigida com base na secção homóloga do relatório exemplo e os RCM em causa, previamente indexados;
2. **Secção de "Fundamentação | Estudos Relevantes"**: é construída a partir da secção homóloga do relatório exemplo e da evidência científica recuperada pelo *MCP Server.*
3. **Secções "Considerações Finais" e "Conclusão":** construídas com base nas secções homólogas do relatório exemplo e em todo relatório redigido até ao momento.

Após a construção inicial do relatório pelo sistema, este é submetido a uma fase de validação interna pela UFC, onde, iterativamente, farão ajustes manuais, preenchimento de secções em falta e revisão crítica da fundamentação.
Concluída esta etapa, o parecer técnico é formalmente submetido à CFT, que procede à avaliação do parecer técnico. Caso o relatório responda de forma adequada ao pedido, o

processo encerra. Caso contrário, é solicitado à UFC um pedido de revisão, reiniciando o ciclo de elaboração.

4. Recuperação de Informação

O processo de Recuperação de Informação (RI) constitui um dos pilares do sistema proposto. O objetivo central consiste em permitir a recuperação eficiente e contextualizada de informação relevante, a partir de diferentes fontes documentais, para posterior integração nos relatórios técnicos elaborados pela UFC.
Aqui descreve-se com maior detalhe o primeiro passo mencionado no capítulo anterior, em particular, o processamento do Resumo das Características do Medicamento (RCM), documento de referência regulamentar e científica que serve de base à secção *"Fundamentação: Farmacologia"* dos pareceres mencionado previamente.

4.1. Documento "Resumo das Características do Medicamento" (RCM)

O RCM é uma fonte estruturada que contém informação farmacológica, clínica e regulamentar sobre cada medicamento autorizado. É com base neste documento que os profissionais recolhem várias informações relativas ao fármaco que acabam por incluir na secção da Farmacologia. Estas informações poderão ser, entre muitas outras:

- **Identificação do medicamento:** nome do medicamento comercial e o respetivo titular da AIM (Autorização de Introdução no Mercado);
- **Indicações terapêuticas:** terapêuticas para as quais o medicamento é aprovado pela EMA;
- **Mecanismo de ação/Propriedades farmacodinâmicas**: modo como a substância ativa atua a nível molecular, celular e/ou sistémico;
- **Perfil farmacocinético:** dados relativos à absorção, distribuição, metabolismo e excreção da substância ativa (bem como outras propriedades relevantes, mediante o fármaco em análise);
- **Posologia**: doses a administrar, frequência, duração e via de administração (poderá apresentar variações mediante os grupos etários, peso, coexistência de outras comorbilidades, etc.);
- **Eficácia e Segurança:** perfil de segurança do medicamento, incluindo contraindicações, advertências especiais, precauções de utilização e potenciais interações, bem como evidência científica relevante que comprovam a sua eficácia no tratamento das terapêuticas mencionadas.

Este documento é, por natureza, extenso e altamente detalhado, frequentemente composto por múltiplas secções técnicas. No entanto, a relevância das diferentes secções para a elaboração de um parecer técnico depende intrinsecamente do contexto em análise, isto é, do tipo de pedido submetido pela CFT, das características clínicas do doente e do medicamento

em avaliação. Logo, nem todas as secções do RCM serão igualmente úteis em todas as situações. Por exemplo, enquanto num pedido de avaliação inicial de um fármaco oncológico pode ser essencial explorar o mecanismo de ação, farmacocinética e eficácia clínica, já num pedido de ajuste posológico num doente com insuficiência renal, a secção relativa a alterações farmacocinéticas em populações especiais adquire maior relevância.

Consequentemente, torna-se fundamental que o relatório final inclua apenas a informação estritamente necessária para responder de forma objetiva e fundamentada ao pedido, evitando a sobrecarga documental e assegurando clareza e foco na tomada de decisão. Para alcançar esta flexibilidade, é necessário que o sistema seja capaz de aceder dinamicamente às diferentes secções do RCM, discernindo de forma autónoma quais os conteúdos mais relevantes para o caso em questão. É neste ponto que se justifica a adoção de metodologias baseadas em *embeddings*, indexação vetorial e pesquisa semântica.

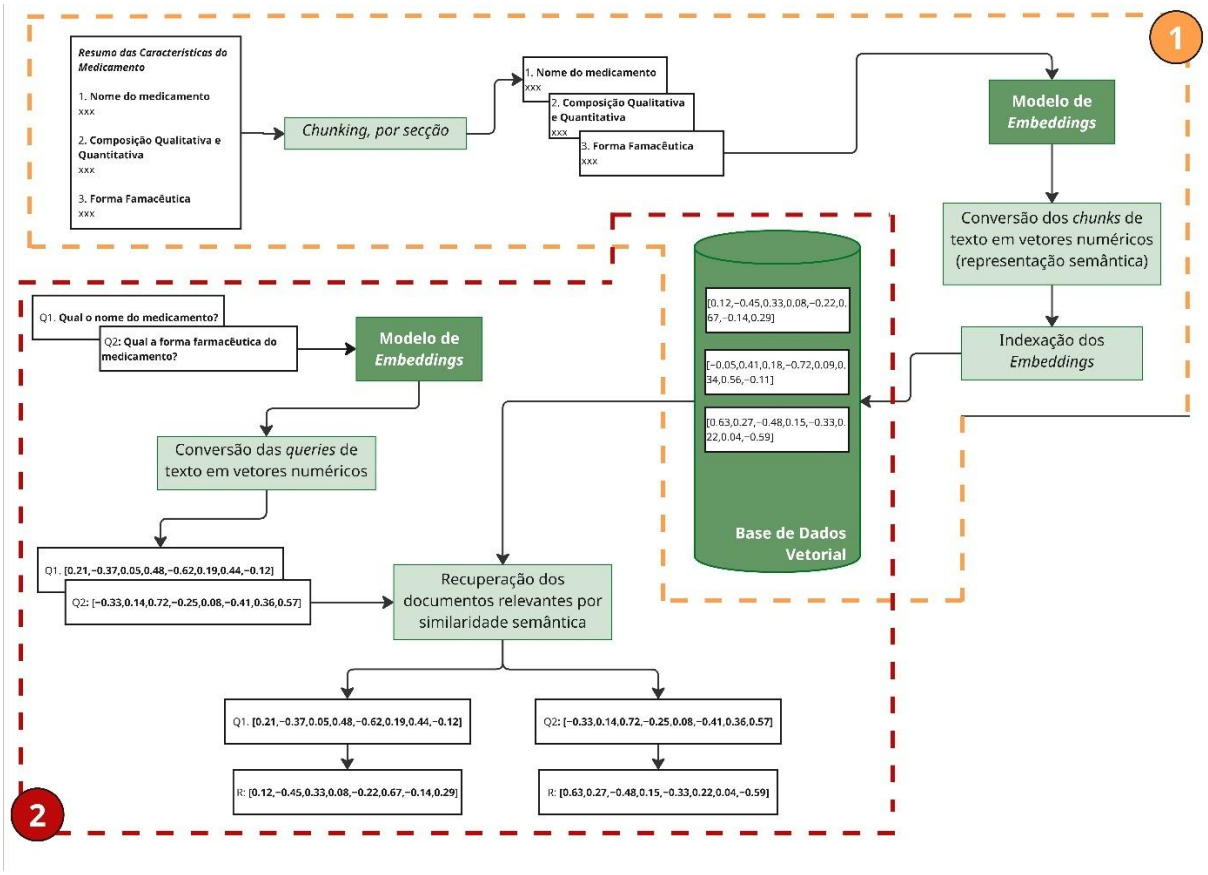


Figura 12 - Processamento, Indexação e Pesquisa Semântica no Documento "Resumo das Características do Medicamento"

Na Figura 12, está representado todo o processo de seccionamento, indexação e pesquisa semântica do RCM, no contexto em questão:

Fase 1 | *Chunking* e indexação em base de dados vetorial: na primeira etapa (delimitada a laranja), o objetivo é preparar e armazenar a informação do RCM numa base de dados vetorial, para que possa ser consultada posteriormente. Primeiramente, o RCM é segmentado em *chunks* correspondentes a secções bem definidas (capítulos e subcapítulos, como, ”Nome *do medicamento”*, “*Composição Qualitativa*" *e Quantitativa”* e “*Forma Farmacêutica”,* entre muitos outros. A opção de realizar o *chunking* segundo as secções do RCM justifica-se pelo facto de cada segmento corresponder a um tópico semântico distinto. Procurando, garantir que cada *chunk* preserva um significado autónomo. De seguida, é feita a conversão dos *chunks* em *embeddings*, onde cada *chunk* é processado por um modelo de *embedding*, que converte o texto em vetores numéricos multidimensionais. Estes vetores capturam a semântica de cada secção, permitindo a comparação baseada em significado e não apenas em palavras exatas. Por fim, é feita a indexação destes vetores numa base de dados vetorial, preparada para suportar consultas semânticas. Aqui, cada vetor fica associado ao texto original, possibilitando a sua recuperação quando necessário.

Relativamente ao processo de *chunking* do RCM, importa salientar que, em alguns casos, não foi possível preservar a totalidade de um capítulo/subcapítulo como um único *chunk*, uma vez que a sua extensão excedia a dimensão máxima de *tokens* suportada pelos modelos de *embedding*. Este é um dos principais desafios inerentes à utilização desta tecnologia: os modelos operam sobre janelas de contexto limitadas, o que obriga a dividir textos mais longos em segmentos menores. Para ultrapassar esta limitação, optou-se por fragmentar as secções extensas em múltiplos *chunks*, aplicando um *overlap* de 2 a 3 frases entre segmentos consecutivos. Este procedimento visa assegurar que a coesão semântica e a continuidade do raciocínio não se perdem no processo de segmentação, permitindo que conceitos ou informações que se encontram na transição entre parágrafos permaneçam acessíveis em ambos os *chunks* e minimizando o risco de descontextualização.
No que diz respeito à base de dados vetorial desempenha, igualmente, um papel central neste processo, funcionando como repositório para armazenamento e consulta das representações numéricas geradas pelos modelos de *embedding*. Ao contrário das bases de dados tradicionais, que indexam documentos a partir de chaves e campos estruturados, as bases vetoriais organizam a informação em função da proximidade semântica entre vetores. Isto permite que consultas em linguagem natural sejam traduzidas em *queries* vetoriais e comparadas diretamente com os *chunks* armazenados, recorrendo a métricas de similaridade (como a similaridade do cosseno aqui utilizada). Anteriormente, no subcapítulo 2.3, fez-se uma análise comparativa das diversas bases de dados vetoriais existentes atualmente no mercado e que poderiam ser utilizadas neste projeto. Optou-se pela utilização da ChromaDB, sobretudo por ser *open source* e de fácil utilização em projetos de prototipagem, sendo, simultaneamente, otimizada para *embeddings* e pesquisa semântica.

Fase 2 | Pesquisa Semântica: na segunda etapa (delimitada a vermelho na Figura 12), o sistema “responde” a *queries* específicas, utilizando o *chunks* previamente indexados na base de dados vetorial. Esta fase corresponde ao momento em que a informação é efetivamente consultada e disponibilizada para integrar o processo de elaboração do relatório técnico. Primeiramente,

o LLM discerne, com base no pedido da CFT, quais as informações mais relevantes que deve incluir na secção "Farmacologia" e, com base nisso, formula múltiplas *queries* em linguagem natural. Ex.: *"Qual é a forma farmacêutica do medicamento?"* ou *"Quais são as propriedades farmacodinâmicas descritas no RCM?"*. De seguida, ocorre a conversão destas *queries* em *embeddings,* isto é, tal como acontece nos documentos mencionados na Fase 1, cada *query* é convertida num vetor numérico (através do mesmo modelo de *embedding* usado na fase de indexação), assegurando que tanto os *chunks* como as *queries* partilham o mesmo espaço vetorial, o que permite a comparação direta em termos de proximidade semântica. Feito isso, o vetor da *query* é, então, comparado com os vetores previamente armazenados na base de dados vetorial. Esta comparação é realizada através de métricas de similaridade. Neste projeto, optou-se pela similaridade do cosseno[6], que quantifica o grau de proximidade semântica entre a *query* e cada *chunk* do RCM. Os resultados são ordenados em função dessa proximidade, permitindo, assim, recuperar o conjunto dos *chunks* com maior similaridade em relação à *query* (neste projeto, optou-se por recuperar o top-3) e, com base neles, o LLM redige a secção.

Relativamente à formulação das *queries*, é importante sublinhar que, ao recorrer a *embeddings*, a pesquisa não depende de correspondências lexicais exatas. Isto significa que termos semanticamente relacionados (ex.: *"forma de apresentação"* em vez de *"forma farmacêutica"*) conduzem (à partida) ao mesmo resultado, tornando o sistema mais robusto a variações terminológicas e sinónimos.
Já no que respeita à comparação e recuperação, a escolha da similaridade do cosseno prendeu-se sobretudo pelo facto de ser comummente utilizada em tarefas de recuperação de informação com *embeddings*.

4.1.1. Modelos de Geração de Embeddings

A conversão dos *chunks* em *embeddings* (tanto do RCM como, posteriormente, das *queries*), é das etapas mais críticas de todos este processo, visto que condiciona, diretamente, a relevância dos documentos recuperados e, consequentemente, a qualidade final do relatório gerado. Vale a pena ressaltar que diferentes modelos de *embedding* produzem representações distintas, uma vez que variam nos dados de treino, na dimensão dos vetores e na arquitetura subjacente. Um modelo treinado em textos biomédicos, por exemplo, poderá capturar melhor as relações semânticas entre termos clínicos e farmacológicos do que um modelo generalista, maximizando assim a pertinência dos resultados recuperados. Por este motivo, torna-se essencial testar diferentes modelos de *embedding* e avaliar o seu desempenho em cenários reais de utilização.

[6] A similaridade do cosseno mede a proximidade entre dois vetores calculando o cosseno do ângulo formado entre eles. É definida como:

$$sim(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$$

onde $A \cdot B$ é o produto escalar dos vetores e $\|A\|$ e $\|B\|$ são as suas normas (comprimento). O valor varia, então, entre -1 e 1, sendo que valores mais próximos de 1 indicam maior similaridade.

Desta forma, foram testados alguns dos principais modelos de *embedding* existentes atualmente (com base no MTEB - *Massive Text Embedding Benchmark Leaderboard*), modelos estes dos 3 principais *providers* (OpenAI, Google e Anthropic). Para além disso, optou-se por testar, também, o Albertina (*encoder* em português de Portugal) e o MediAlbertina (treinado para ser utilizado no domínio biomédico, igualmente, em pt-pt). A escolha destes dois modelos para além dos principais do mercado, prende-se com o facto de os relatórios redigidos pela UFC serem escritos em português. Assim, presumiu-se que poderiam, eventualmente, apresentar uma melhor adaptação linguística e terminologia mais adequada.

Na Tabela 3, constam os diferentes modelos de *embedding* utilizados, bem como algumas das suas características:

Tabela 3 - Modelos de *Embedding* utilizados no processo de IR

Modelo / Provider	Idioma principal	Domínio de treino	Dimensão do vetor	Características distintivas
text-embedding-ada-002 (OpenAI)	Multilingue	Geral, ampla cobertura linguística e temática	1536	Padrão de mercado, eficiente, excelente equilíbrio entre custo e desempenho; amplamente validado em *benchmarks* de IR.
gemini-embedding-001 (Google)	Multilingue	Geral + otimizações contextuais	768	Modelo de *embedding* mais recente da Google, com foco em flexibilidade semântica e integração nativa com outros modelos Gemini.
voyage-3.5 (Voyage AI) - *recomendação Anthropic*	Multilingue	Geral, adaptado a tarefas de IR	1024	Alta performance em tarefas de recuperação, otimizado para *retrieval-augmented generation*.
albertina-pt-pt-1.5b (Hugging Face)	Português (Portugal)	Geral (*corpus* pt-pt)	768	*Encoder* específico para português europeu; melhor captação de nuances linguísticas e semânticas em pt-pt.
medialbertina-pt-pt-1.5b (Hugging Face)	Português (Portugal)	Biomédico (literatura médica e farmacológica pt-pt)	768	*Encoder* especializado em terminologia biomédica; captura mais fiel das relações semânticas em textos clínicos e farmacológicos.

Para avaliar, sistematicamente, o desempenho de cada modelo, foi criado um *dataset* próprio (Tabela 4), composto por pares de perguntas e respostas referentes a cinco medicamentos distintos (*Tremfya, Eylea, Lemtrada, Orencia e Piqray*). Para cada medicamento, foram formuladas entre 3 a 5 *queries* (por cada secção do RCM), cobrindo tópicos como forma farmacêutica, indicações terapêuticas, mecanismo de ação, posologia e segurança. Este

procedimento permitiu obter uma amostra significativa e representativa das consultas esperadas no contexto real de elaboração de relatórios. Optou-se por construir um *dataset* próprio, apesar da existência de *benchmarks* específicos para o domínio biomédico, por forma a garantir que as *queries* e respostas refletissem de forma fidedigna o contexto real de utilização dos relatórios técnicos da UFC, assegurando maior relevância e aplicabilidade prática.

Tabela 4 - *Dataset* de pares "*queries-answers"* para cada *chunk* do RCM (medicamento: Eylea)

Query	**Resposta**	***Chunk* de Referência**
Qual é o nome do medicamento?	O nome do medicamento é Eylea.	1. NOME DO MEDICAMENTO Eylea 40 mg/ml solução injetável em seringa pré-cheia
Qual é a dosagem do medicamento?	A dosagem do medicamento é de 40 mg/ml.	1. NOME DO MEDICAMENTO Eylea 40 mg/ml solução injetável em seringa pré-cheia
Em que forma é apresentado o medicamento?	O medicamento é apresentado como uma solução injetável em seringa pré-cheia.	1. NOME DO MEDICAMENTO Eylea 40 mg/ml solução injetável em seringa pré-cheia
Qual é a via de administração do medicamento?	O medicamento é administrado por injeção.	1. NOME DO MEDICAMENTO Eylea 40 mg/ml solução injetável em seringa pré-cheia
O medicamento já vem preparado para uso?	Sim, o medicamento já vem preparado para uso, em uma seringa pré-cheia.	1. NOME DO MEDICAMENTO Eylea 40 mg/ml solução injetável em seringa pré-cheia
Qual é a composição qualitativa e quantitativa do medicamento?	1 ml de solução injetável contém 40 mg de aflibercept. Uma seringa pré-cheia contém um volume extraível de pelo menos 0,09 ml, equivalente a pelo menos 3,6 mg de aflibercept.	2. COMPOSIÇÃO QUALITATIVA E QUANTITATIVA 1 ml de solução injetável contém 40 mg de aflibercept*. Uma seringa pré-cheia contém um volume extraível de pelo menos 0,09 ml, equivalente a pelo menos 3,6 mg de aflibercept. Isto proporciona uma quantidade utilizável para administração de uma dose única de 0,05 ml contendo 2 mg de aflibercept em doentes adultos ou uma dose única de 0,01 ml contendo 0,4 mg de aflibercept em recém-nascidos prematuros. (…)
(…)	(…)	(…)

Uma vez definido o *dataset*, para cada uma das *queries*, procedeu-se à recuperação dos documentos mais semelhantes a partir da BD vetorial (Tabela 5), utilizando todos os diferentes modelos mencionados acima. No caso, foram recuperadas as 3 secções que apresentavam menor distância semântica com a *query*.

Tabela 5 - Recuperação das secções do RCM (Eylea) mais próximas semanticamente de cada uma das *queries* (Modelo: *text-embedding-ada-002*)

Query	1ª Secção Recuperada	2ª Secção Recuperada	3ª Secção Recuperada	Distância 1	Distância 2	Distância 3
Qual é o nome do medicamento?	1. NOME DO MEDICAMENTO /	3. FORMA FARMACÊUTICA /	2. COMPOSIÇÃO QUALITATIVA E QUANTITATIVA /	0,316	0,365	0,395
Qual é a dosagem do medicamento?	4. INFORMAÇÕES CLÍNICAS / 4.9 Sobredosagem	2. COMPOSIÇÃO QUALITATIVA E QUANTITATIVA /	5. PROPRIEDADES FARMACOLÓGICAS / 5.2 Propriedades farmacocinéticas / Outras Propriedades	0,320	0,342	0,345
Em que forma é apresentado o medicamento?	3. FORMA FARMACÊUTICA /	1. NOME DO MEDICAMENTO /	5. PROPRIEDADES FARMACOLÓGICAS / 5.2 Propriedades farmacocinéticas / Outras Propriedades	0,306	0,349	0,360
Qual é a via de administração do medicamento?	5. PROPRIEDADES FARMACOLÓGICAS / 5.2 Propriedades farmacocinéticas / Outras Propriedades	4. INFORMAÇÕES CLÍNICAS / 4.2 Posologia e modo de administração / Modo de administração	3. FORMA FARMACÊUTICA /	0,337	0,362	0,371
O medicamento já vem preparado para uso?	1. NOME DO MEDICAMENTO /	3. FORMA FARMACÊUTICA /	2. COMPOSIÇÃO QUALITATIVA E QUANTITATIVA /	0,347	0,368	0,389
Qual é a composição qualitativa e quantitativa do medicamento?	2. COMPOSIÇÃO QUALITATIVA E QUANTITATIVA /	3. FORMA FARMACÊUTICA /	5. PROPRIEDADES FARMACOLÓGICAS / 5.2 Propriedades farmacocinéticas / Outras Propriedades	0,300	0,364	0,373
(…)	*(…)*	*(…)*	*(…)*	*(…)*	*(…)*	*(…)*

A geração de *embeddings* e a consequente recuperação de documentos foram realizadas através da *framework* LangChain, que simplifica bastante o teste de vários modelos distintos, bases de dados vetoriais e fluxos de *retrieval*, permitindo uma integração eficiente e modular de todo o processo.

Ainda assim, ao contrário dos modelos disponibilizados diretamente pela OpenAI, Google ou Voyage, os modelos Albertina e o MediAlbertina são disponibilizados apenas como *encoders* de texto no *Hugging Face*, sem *endpoints* prontos para geração de *embeddings*. Logo, para os integrar no pipeline, foi necessário criar *endpoints* personalizados. Desta forma, criou-se um *endpoint* dedicado em *Hugging Face*, através do qual era possível enviar texto em português de Portugal e receber o vetor correspondente. O processo consistiu em passar o texto pelo *encoder* do modelo, extraindo os embeddings de cada token, e em seguida aplicar *mean pooling*, isto é, a média desses embeddings, de forma a gerar um único vetor representativo da sequência/frase (optou-se por *mean pooling* em detrimento de outras estratégias, como [CLS], pela evidência empírica existente que demonstra que o *mean pooling* fornece, geralmente, representações mais consistentes e eficazes em tarefas de similaridade semântica e recuperação de informação (Reimers & Gurevych, 2019)). Após o cálculo dos vetores, procedeu-se à normalização L2, um processo em que cada vetor é dividido pela sua norma (comprimento), garantindo que a sua magnitude total seja sempre igual a 1. Esta etapa assegura que, ao aplicar medidas de similaridade como o cosseno, a comparação entre *embeddings* dependa apenas da direção do vetor (isto é, da informação semântica) e não da sua escala, resultando numa comparação mais justa, estável e consistente entre representações distintas.

Uma vez operacionalizados os *endpoints* de ambos os modelos, tanto o Albertina e MediAlbertina puderam ser utilizados da mesma forma que os restantes, através da *framework* LangChain.

4.1.2. Avaliação do Processo de Recuperação (RCM)

Para avaliar a qualidade do processo de recuperação de todos os modelos, foi construído um *dataset ground truth* para cada um dos medicamentos. Este *dataset* é em tudo semelhante ao mencionado anteriormente, com a única diferença a ser que, para cada *query*, é indicado se os *chunks* recuperados são (1) ou não (0) objetivamente relevantes para responder à *query* em questão. Tal irá permitir comparar o desempenho do sistema de recuperação com uma referência validada manualmente.

Para quantificar a eficácia da recuperação, foram aplicadas algumas métricas clássicas de recuperação de informação, nomeadamente:

- **MAP (*Mean Average Precision*)**: avalia a precisão ao longo de todos os documentos recuperados, dando maior peso à ordenação correta dos documentos relevantes:

$$AP = \frac{1}{R}\sum_{k=1}^{N} P(k) \cdot rel(k)$$

$$MAP = \frac{1}{Q}\sum_{q=1}^{Q} AP_q$$

onde R é o nº total de documentos relevantes, N é o número de documentos recuperados, $P(k)$ é a precisão no ponto k, $rel(k)$ indica a relevância (1) ou não relevância (0) do documento recuperado e Q representa o nº total de *queries.*

- **MRR (*Mean Reciprocal Rank*)**: mede a posição do primeiro documento relevante recuperado, refletindo a rapidez com que o sistema apresenta informação útil:

$$MRR = \frac{1}{Q}\sum_{q=1}^{Q} \frac{1}{rank_q}$$

onde $rank_q$ representa a posição do primeiro documento relevante para a *query* q.

- **Precision@3**: proporção de documentos relevantes que surgem nos 3 documentos recuperados:

$$P@3 = \frac{nº\ documentos\ relevantes\ nos\ 3\ recuperados}{3}$$

- **Recall@3**: proporção de documentos relevantes que surgem nos três primeiros resultados, em relação ao total de relevantes existentes no *ground truth*.

$$Recall@3 = \frac{nº\ documentos\ relevantes\ nos\ 3\ recuperados}{nº\ total\ de\ documentos\ relevantes}$$

- **NDCG@3 (*Normalized Discounted Cumulative Gain*)**: avalia a qualidade da ordenação dos documentos nos três primeiros lugares, atribuindo maior peso aos mais bem classificados.

$$DCG@3 = \sum_{k=1}^{3} \frac{2^{rel(k)} - 1}{\log_2(k+1)}$$

$$NDCG@3 = \frac{DCG@3}{IDCG@3}$$

onde $rel(k)$ indica a relevância do documento na posição k e $IDCG@3$ representa o valor ideal do DCG, ou seja, quando todos os relevantes estão ordenados no topo.

Para prosseguir com o processo de avaliação, foram calculadas as métricas acima para cada um dos modelos utilizado na geração de *embeddings* (relativas ao *dataset* específico construído para o medicamento *Eylea*) (Gráfico 1).

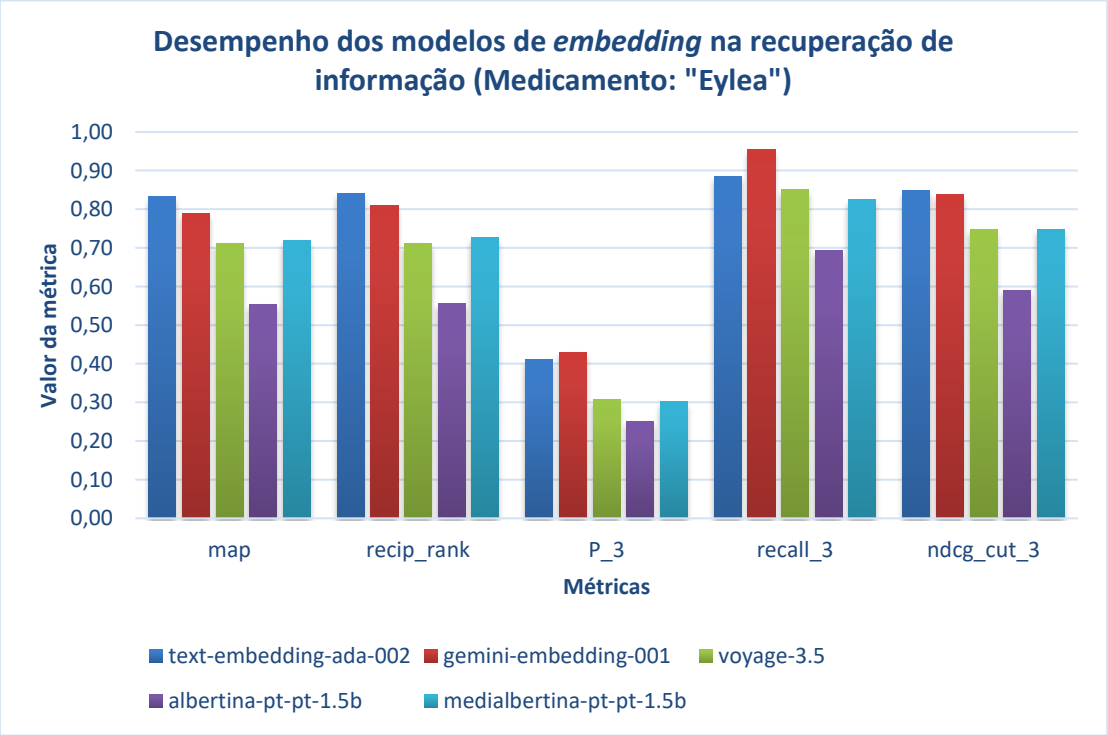


Gráfico 1 - Comparação dos modelos de *embedding* no processo de Recuperação de Informação (Medicamento: "Eylea")

Analisando o gráfico comparativo das métricas para cada um dos modelos testados, podemos observar uma quase supremacia dos modelos OpenAI e Google. Efetivamente, o *text-embedding-ada-002* foi o melhor na ordenação (MAP = 0.832; MRR = 0.841 e NDCG = 0.849), com o *gemini-embedding-001* a vencer no que diz respeito à cobertura e precisão (*Recall* = 0.955 e *Precision* = 0.428). Já o *voyage-3.5* e o medialbertina-pt-pt-1.5b obtiverem valores intermédios, com o *encoder* albertina-pt-pt-1.5b a ser, como já esperado, o mais fraco em todas as métricas.

Comparando diretamente os dois melhores modelos, podemos, então, assumir que o *gemini* encontra mais frequentemente documentos relevantes no top-3 (alta cobertura), enquanto o modelo da OpenAI tende a colocar os documentos mais relevantes nas posições acima, ou seja, em média, devolve os documentos mais relevantes em primeiro lugar, reduzindo a probabilidade de *chunks* irrelevantes no topo (redução de ruído). Ênfase na inexistência de ruído para o primeiro, e a recuperação dos "mais" relevantes no topo para o segundo.

No fundo, tanto um como outro apresentaram resultados bastante satisfatórios, com diferenças marginais. Ainda assim, optou-se por priorizar as métricas MAP, MRR e NDCG@3, uma vez que avaliam, não apenas a presença de documentos relevantes entre os resultados recuperados, mas também a sua correta ordenação nas primeiras posições, fator determinante em sistemas RAG com espaço de contexto limitado. Este critério reveste-se de particular importância no domínio da FC, dado que os documentos são extensos e é fundamental que os LLMs sejam alimentados preferencialmente com evidência de elevada relevância, minimizando a incorporação de ruído. Assim, optou-se por priorizar o modelo *text-embedding-ada-002*.

Os resultados acima foram calculados apenas com os resultados obtidos para o medicamento Eylea. Para tentar validar que o modelo apresenta um bom desempenho de forma generalizável, o mesmo processo foi realizado para mais 4 medicamentos, tendo sido obtidos os resultados seguintes:

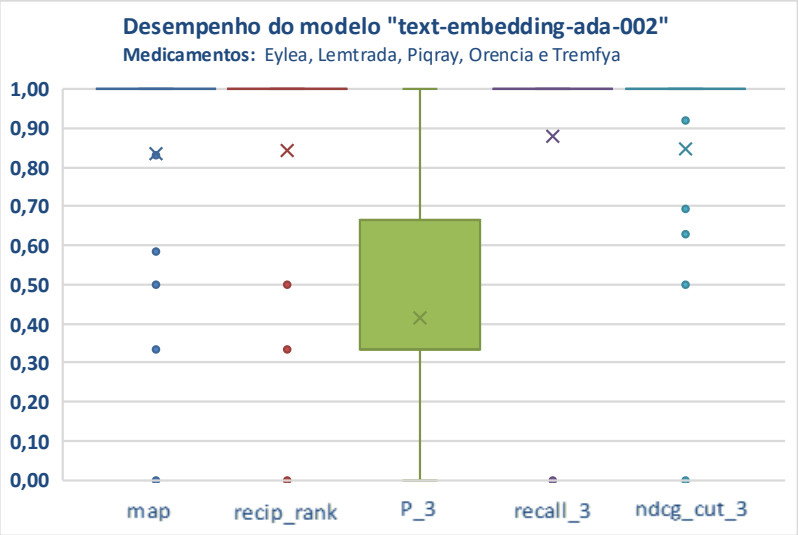


Gráfico 2 - Distribuição das métricas na avaliação do processo de IR

Interpretando os resultados obtidos, podemos concluir que o MAP e MRR estão muito concentrados, perto de 1, ainda que com alguns *outliers* mais baixos. Isto mostra que, na maioria das *queries*, o primeiro relevante surge logo no topo e a ordenação é quase perfeita. Quanto ao *recall,* este está também muito perto do máximo (1) para quase todas as *queries,* ou seja, o sistema recupera praticamente todos os documentos relevantes já nos três primeiros resultados. No que diz respeito à precisão, verifica-se uma maior dispersão, indo de 0 até, aproximadamente, 0,66. Isto significa que, apesar de os documentos relevantes estarem geralmente presentes entre os primeiros resultados, nem sempre os três documentos no topo são todos relevantes, existindo algum ruído. Tal decorre do facto de, para responder a determinadas *queries*, apenas ser necessário 1 ou 2 *chunks* relevantes, sendo os restantes recuperados, inevitavelmente, irrelevantes. Exemplificando, num caso em que apenas 1 *chunk* seja relevante, caso o sistema RAG o recupere, o *recall* será 1 (todos os documentos relevantes foram recuperados) e a *precision* será apenas 0,33 (1 de 3). Assim, o valor relativamente mais

baixo desta métrica, quando comparado com as restantes, não compromete, necessariamente, o desempenho global do modelo. Por fim, a métrica NDCG, com valores altos e estáveis, tal confirma boa ordenação nos primeiros lugares.

4.1.3. Finetuning do encoder Albertina

Apesar dos bons resultados obtidos com o modelo *text-embedding-ada-002*, colocou-se a hipótese de que um *finetuning* direcionado, recorrendo aos *datasets* de pares já construídos, poderia melhorar o desempenho do *encoder* Albertina. Esta estratégia, para além de potencialmente melhorar os resultados das métricas de recuperação, teria a vantagem estratégica de reduzir a dependência de *providers* externos (neste caso, da OpenAI), garantindo maior controlo sobre o modelo em cenários futuros.

O *finetuning* do *encoder* Albertina foi conduzido em ambiente Google Colab Pro+, utilizando uma GPU NVIDIA A100 (40GB de RAM). Este recurso permitiu suportar cargas de treino consideráveis, embora com algumas limitações práticas em termos de memória quando se exploravam *batch sizes* e sequências mais extensas (como será evidente de seguida).

Para o treino, utilizou-se *contrastive learning*, na qual pares de textos relevantes (ex.: *query–chunk* e *query–answer*) são aproximados no espaço vetorial, enquanto exemplos não relevantes são afastados. Para tal, foram definidos tanto positivos (pares semanticamente relevantes) como negativos (pares construídos aleatoriamente ou selecionados como *hard negatives*).

Foram conduzidas várias iterações no treino do *encoder* (Tabela 6), com o objetivo de avaliar a sensibilidade do modelo às diferentes condições e configurações. Os resultados obtidos mediante cada experiência de *finetuning* realizada encontram-se na Tabela 7.

Tabela 6 - Experiências de *finetuning* realizadas no *encoder* Albertina

Experiência	Descrição	Condições principais/Principais alterações
1. Treino base	*Finetuning* com parâmetros padrão	*Epochs* = 2 *L*earning rate = 3e-5 *Batch size* = 2 Sequência truncada a 256 tokens Positivos apenas compostos pelos *chunks.*
2. Ajuste dos parâmetros	Aumento da duração (mais *epochs*) e redução da taxa de aprendizagem.	*Epochs* = 5 *L*earning rate = 2e-5
3. Inclusão de *answers*	Integração das respostas finais como pares positivos	Pares positivos passam a incluir também as *answers* do *dataset* de pares originalmente construído (e não só os *chunks* relevantes).
4. *Hard negatives*	Introdução de negativos semanticamente próximos	Junção de exemplos incorretos (negativos) mas semelhantes (*hard*), forçando o modelo a distinguir termos sobrepostos.
5. Modelo reduzido (900M parâmetros)	Substituição do modelo completo pelo menor	Sequência truncada a 512 tokens e *batch size* = 8 Compromisso: maior contexto e mais robustez vs. menor capacidade do modelo

Tabela 7 - Resultados obtidos por cada experiência de *finetuning*

Modelo Embeddings	ID Experiência	Experiência	map	recip_rank	P_3	recall_3	ndcg_cut_3
albertina-pt-pt-1.5b	-	*Encoder* Base	0,553	0,557	0,251	0,693	0,590
albertina-pt-pt-1.5b	1	Treino Base	0,724	0,724	0,263	0,781	0,739
albertina-pt-pt-1.5b	2	Ajuste de parâmetros	0,719	0,727	0,301	0,825	0,748
albertina-pt-pt-1.5b	3	Inclusão de *answers*	0,740	0,749	0,310	0,851	0,771
albertina-pt-pt-1.5b	4	*Hard Negatives*	0,745	0,753	0,313	0,860	0,777
albertina-pt-pt-900m	1	Treino Base	0,719	0,727	0,301	0,825	0,748
albertina-pt-pt-900m	2	Ajuste de parâmetros	0,727	0,735	0,307	0,842	0,759
albertina-pt-pt-900m	3	*Hard Negatives*	0,758	0,766	0,316	0,868	0,789

Os resultados obtidos com o *finetuning* do Albertina evidenciaram diferenças relevantes entre as várias configurações testadas, com melhorias significativas em todas elas quando comparadas com o *encoder* base. No modelo *albertina-pt-pt-1.5b*, o treino inicial apresentou valores razoáveis, mas com fraca precisão nos três primeiros documentos recuperados, com bastante ruído no top3. O aumento do número de *epochs*, em conjunto com a redução da *learning rate*, não resultou em melhorias nas métricas de ordenação, mas permitiu ganhos nas métricas de topo, sugerindo que esta estratégia privilegia a relevância imediata dos primeiros resultados devolvidos (NDCG = 0,739), ainda que à custa de alguma instabilidade no *ranking* completo (MAP = 0,719).

A introdução das respostas finais (*answers*) como pares positivos demonstrou um impacto claro e positivo em todas as métricas, em particular no aumento da relevância dos documentos mais próximos do topo da lista (MRR = 0,749 e NDCG = 0,771).

De igual modo, a inclusão de *hard negatives* reforçou a capacidade discriminativa do modelo perante exemplos semanticamente próximos, com uma melhoria em todas as métricas. Este aspeto é particularmente relevante em contexto clínico, onde a terminologia utilizada é bastante precisa, e ligeiras mudanças poderão fazer com que certos documentos deixem de apresentar utilidade para responder à questão em análise.

No caso do modelo *albertina-pt-pt-900m*, apesar da sua menor dimensão, o que poderia, à partida, indicar que o desempenho seria menor, a verdade é que a possibilidade de aumentar o limite de *tokens* para 512 e o *batch size* para 8 revelou-se determinante. Com o ajuste deste parâmetros, o modelo apresentou melhorias face ao treino inicial em todas as métricas, sendo que, com a adição de *hard negatives*, atingiu o melhor desempenho global de todas as experiências (MAP = 0,758; MRR = 0,766; P@3 = 0,316; R@3 = 0,868; NDCG@3 = 0,789). Efetivamente, verificou-se que o facto de o modelo truncar as sequências em 256 *tokens* estava a impedir que as secções dos *chunks* fossem indexadas na sua totalidade, o que poderia levar a que conteúdo importante não fosse considerado.

A comparação entre modelos confirma que, neste contexto, o fator determinante não foi o número de parâmetros, mas antes a capacidade de processar mais contexto e treinar com *batch sizes* mais elevados. O modelo *albertina-pt-pt-900m* com *hard negatives* superou consistentemente o modelo maior em todas as métricas, com destaque para o aumento de cerca de 20% em precisão e 11% em *recall* face aos resultados do modelo sem *finetuning*.

Apesar das melhorias alcançadas, estas configurações não foram suficientes para superar o desempenho do *text-embedding-ada-002*, ainda que os resultados obtidos se revelem promissores e demonstrem o potencial do *finetuning* direcionado em português europeu. Futuramente, de forma a explorar e maximizar esse potencial, sugere-se:

1. aumentar a diversidade e dimensão do *dataset* de pares de treino;

2. explorar técnicas de *in-batch negatives* e mineração automática de exemplos negativos a partir de métodos tradicionais de recuperação (com *negatives* de várias dificuldades (não só *hards*);
3. melhorar infraestrutura de treino, com consequente ajuste dos respetivos parâmetros de treino.

4.2. Pesquisa de Literatura Científica (PubMed)

Na secção "Fundamentação | Estudos Relevantes", a equipa da UFC sintetiza a evidência científica mais atual e pertinente que responda diretamente ao pedido da CFT, selecionando estudos que abordam explicitamente o fármaco em análise, o enquadramento e o caso clínico. Dado a área científica em questão, grande parte desta pesquisa é realizada na base de dados do PubMed.

Na tentativa de agilizar este processo de pesquisa, optou-se por tirar partido das potencialidades dos MCP *Servers* que abstraem uma boa parte da orquestração e integração que seriam necessárias, de outra forma, desenvolver. Assim, através de *queries* em linguagem natural, o MCP (alimentado por um LLM) realiza a pesquisa no PubMed, utilizando diversas *tools* que tem à sua disposição, retornando, por fim, os respetivos artigos e um resumo da evidência científica encontrada.

Existem inúmeros servidores MCP, cada um com as suas forças e fraquezas. Procurou-se, desta forma, encontrar o melhor MCP no momento que permitisse cumprir o objetivo proposto acima. Dentro dos principais repositórios de MCP, foi possível encontrar, até à data, 3 servidores principais que permitem realizar pesquisa no PubMed:

1. **Simple PubMed** (https://github.com/andybrandt/mcp-simple-pubmed): servidor MCP minimalista que permite pesquisa na base de dados PubMed. Facilita a procura por palavras-chave, a recuperação de resumos (*abstracts*) e, quando disponíveis, o download dos artigos completos, em formato XML;
2. **PubMed MCP Server** (@JackKuo666/pubmed-mcp-server): disponibiliza uma interface MCP robusta e abrangente, permitindo, não só a busca por artigos e acesso a metadados detalhados, como também a tentativas de obtenção de artigos completos e uso de *prompts* especializados para guiar a respetiva análise ao pedido em questão;
3. **BioMCP** (https://biomcp.org/): desenvolvido pela GenomOncology, o BioMCP é um servidor de domínio biomédico mais abrangente que os restantes mencionados acima, principalmente por integrar múltiplas fontes autorizadas de dados (não apenas PubMed, mas também ClinicalTrials.gov, PubTator (literatura com anotações), bioRxiv/medRxiv (preprints), MyVariant.info (variantes genéticas), MyGene.info, OpenFDA, entre outras), fornecendo uma interface unificada e estruturada para pesquisas em linguagem natural sobre literatura biomédica, ensaios clínicos e variantes

genéticas (Oncology, 2025). Certificado pelo MCPHub, disponibiliza um conjunto extensivo de ferramentas (cerca de 24 *tools*) que ampliam o suporte aos fluxos de trabalho de investigação biomédica.

Face às características descritas, a opção mais promissora apontava naturalmente para o BioMCP, pela sua integração num ecossistema biomédico alargado e pelo vasto conjunto de ferramentas que disponibiliza, permitindo elevar a pesquisa de evidência científica a um nível mais avançado, requisito fundamental no contexto em análise. Ainda assim, procedeu-se à avaliação prática dos três servidores, uma vez que não se podia excluir a possibilidade de uma solução mais simples, mas potencialmente mais estável, se revelar a escolha mais adequada nesta fase.

Análise dos MCP *Servers*:

Importa sublinhar que este protocolo constitui uma tecnologia ainda recente, pelo que era expectável que nem todos os servidores apresentassem níveis de robustez equivalentes. Tal hipótese confirmou-se no caso do Simple PubMed, que demonstrou significativa instabilidade, com falhas frequentes na recuperação de resultados e limitações evidentes nos parâmetros configuráveis. Além disso, a qualidade e a abrangência da evidência científica obtida mostraram-se bastante aquém do necessário. Com efeito, poucos meses após os primeiros testes, este MCP acabaria mesmo por ser removido da plataforma *Smithery*.

As duas opções restantes mostraram-se consideravelmente mais estáveis durante a sua execução. O *PubMed MCP Server*, com as suas quatro *tools* disponíveis (*search_pubmed_key_words, search_pubmed_advanced, get_pubmed_article_metadata* e *download_pubmed_pdf*), revelou um desempenho rápido e eficaz, permitindo recuperar artigos, metadados detalhados e, em alguns casos, os textos completos. Apesar destes pontos fortes, verificou-se uma limitação relevante: o número de artigos devolvidos por pesquisa era reduzido (tipicamente entre dois e três), mesmo quando se especificava explicitamente a necessidade de um maior volume de resultados. No contexto das *queries* realizadas, tal restrição revelou-se insuficiente para garantir a abrangência e profundidade de evidência científica exigidas.

O BioMCP destacou-se de forma evidente entre as alternativas avaliadas. O seu extenso conjunto de *tools* permite, não apenas realizar pesquisas no PubMed, mas também aceder a informação complementar proveniente de ensaios clínicos (particularmente relevante, face aos temas utilizados nos testes), variantes genéticas e outras fontes biomédicas. Acresce a possibilidade de acompanhar o raciocínio subjacente à geração do *output*, o que evidenciou uma notável proximidade com a metodologia habitualmente seguida por profissionais de farmacologia na análise crítica da literatura e na pesquisa, recolha e síntese da mesma. Esta combinação de abrangência, transparência e rigor analítico posicionou o BioMCP como a solução mais completa e adequada para responder às exigências do contexto em estudo.

4.2.1. BioMCP – MCP Server

As principais características do BioMCP podem ser consultadas no Anexo C. Escolhido o MCP *Server*, passou-se a uma fase mais intensiva de testes. Utilizou-se:

1. **GitHub Copilot** (VS Code) como *front-end* conversacional;
2. **BioMCP:** chamadas padronizadas por STDIO (local);
3. **Modelos LLM testados:** para controlar o efeito do LLM na utilização do MCP, nas *tools* utilizadas, nº de chamadas e *output final*, as mesmas *prompts* foram enviadas a três modelos de referência distintos, nomeadamente, GPT-5, Claude Sonnet 4 e Gemini 2.5 Pro.

Para testes, foram escolhidos 3 pareceres históricos realizados por profissionais da UFC. Enquadramento e caso clínico trabalhado em cada um dos relatórios podem ser analisados na Tabela 8 e no Anexo D.

- Relatório 1: "*Osimertinib adjuvante no tratamento do adenocarcinoma do pulmão localmente avançado*"
- Relatório 2: *"Ustecinumab em regime posológico off-label na Doença Inflamatória Intestinal"*
- Relatório 3: *"Tratamento com Burosumab no adulto com raquitismo"*

Tradicionalmente, a pesquisa em bases de dados científicas, como o PubMed, é realizada através de *queries* estruturadas, que combinam *keywords*, operadores booleanos, filtros de data e tipos de artigo, etc. Com a utilização de um MCP, torna-se possível recorrer às capacidades dos LLMs para gerar automaticamente essas *queries* a partir de uma descrição em linguagem natural, introduzida pelo utilizador. Esta abordagem oferece maior flexibilidade, uma vez que o próprio servidor pode ajustar iterativamente a *query* em função do pedido inicial e dos resultados que vai obtendo. Assim, as *queries* aqui formuladas em linguagem natural serão posteriormente convertidas em queries estruturadas equivalentes para execução no PubMed. Este processo é suportado pelo estudo de avaliação do (Luo et al., 2025), onde os autores demonstram que a substituição da formulação direta de *queries* em SQL por uma interface declarativa em linguagem natural, traduzida automaticamente em SQL por um modelo *text-to-SQL*, resulta em ganhos de desempenho significativos (até +22 pontos de acurácia).

Desta forma, recorreu-se ao título, enquadramento e caso clínico específicos de cada um dos relatórios para gerar a respetiva *query*[7,8] a fazer ao BioMCP. A Tabela 8 contém a *query* relativa ao relatório 1. Para os dois restantes relatório, verificar o Anexo D.

[7] As *queries* a serem enviadas para o servidor MCP serão em inglês, visto ser este o idioma de grande parte da literatura científica no PubMed.
[8] Como será possível verificar no subcapítulo 5.1, a geração deste resumo em forma de *query* será um dos passos a delegar ao LLM.

Tabela 8 - Formulação da *Query* para o BioMCP relativa ao relatório 1, com base no Título, Enquadramento e Caso Clínico

Relatório	Enquadramento	Caso Clínico	*Query* formulada para o BioMCP
Osimertinib adjuvante no tratamento do adenocarcinoma do pulmão localmente avançado	A respeito de um caso clínico, solicita a Comissão de Farmácia e Terapêutica (CFT) da Unidade Local de Saúde S. João (ULS São João) parecer científico à Unidade de Farmacologia Clínica (UFC) relativo à eficácia e segurança do tratamento de adenocarcinoma do pulmão localmente avançado com osimertinib, em adjuvância.	**Paciente**: XXXXX, sexo feminino, 75 anos, ECOG 1 **Antecedentes Pessoais**: Hipertensão arterial; dislipidemia; perturbação depressiva; histerectomia, sem anexotomia; exérese de nódulos mamários, benignos; apendicectomia; gonartroses bilaterais; prótese da anca à esquerda; doença hemorroidária. **Medicação Habitual**: Rosuvastatina + ezetimiba, 20 mg + 10 mg, 1cp/dia; valsartan + hidroclorotiazida, 80 mg + 12,5 mg, 1cp/dia; bisoprolol 5 mg, 1cp/dia; duloxetina 60 mg, 1cp/dia; ácido fólico 5 mg, 1cp/dia. (…)	*"Search for high-quality studies (randomized controlled trials, meta-analyses, clinical guidelines, and cohort studies) evaluating osimertinib as consolidation therapy after definitive chemoradiotherapy in unresectable, locally advanced non-small cell lung cancer (NSCLC), specifically adenocarcinoma, with EGFR exon 19 deletion mutations (and other EGFR mutations if evidence is limited). Focus on patients with stage III disease (e.g., IIIB, T3N2M0) who did not progress following platinum-based chemoradiotherapy with curative intent, and consider data on efficacy (progression-free survival, overall survival, response rates) and safety/tolerability in older patients (around 70–80 years, ECOG 1)."*

4.2.2. Resultados e Avaliação do BioMCP

De forma semelhante ao processo de avaliação do "Resumo das Características do Medicamento" (subcapítulo 4.1), também aqui foi construído um *gold standard dataset* (validado por profissionais de farmacologia), com uma listagem de artigos científicos relevantes e atuais que permitem responder aos 3 pedidos em questão. Vale a pena ressaltar que este *dataset* poderá não se manter atualizado por muito tempo, com a publicação de novos estudos relevantes ao tópico em análise. O mesmo se pode dizer dos artigos mencionados no histórico de pareceres que, à luz do *corpus* científico atual, poderão já não estar mais atualizados. Por este motivo, para construção destes *datasets,* não se limitou aos artigos já identificados no parecer, mas nova pesquisa foi feita para garantir que a informação mais atualizada até ao momento era corretamente identificada. Ainda assim, o efeito de atualização do *corpus* deverá ser tido em consideração na análise dos resultados obtidos pelas métricas de avaliação do servidor MCP. Os *datasets* encontram-se no Anexo E para consulta.

Com as *queries* definidas e o conjunto de referência estabelecido, avançou-se para a fase de avaliação. Tendo em conta o carácter não determinístico dos LLMs e o objetivo de verificar a estabilidade, robustez e variabilidade da recuperação, cada *query* foi executada cinco vezes por modelo sob as mesmas condições, tendo sido recuperados e registados os 5 artigos mais relevantes (top 5).

Um dos principais benefícios da utilização do BioMCP integrado no GitHub Copilot é a possibilidade de acompanhar, em detalhe, todo o trilho de raciocínio seguido pelo modelo até à obtenção do *output* final. De facto, constatou-se que a interação com o BioMCP varia consideravelmente entre os diferentes LLMs:

- **GPT-5** recorre extensivamente às *tools* disponíveis, procurando iterativamente refinar os resultados e colmatar lacunas identificadas ao longo da pesquisa;
- **Gemini 2.5 Pro**, em contraste, realiza um número consideravelmente menor de chamadas às ferramentas, formulando *queries* mais amplas que resultam na recuperação de um grande volume de artigos de uma só vez;
- **Claude Sonnet 4** adota uma abordagem intermédia, caracterizada por múltiplos comentários e reflexões à medida que o processo avança, o que acrescenta valor em termos de explicabilidade e transparência.

A título de exemplo, a tabela presente no Anexo F mostra os resultados obtidos para o relatório nº 1, utilizando o GPT-5 como LLM orquestrador.

Numa primeira instância, observa-se uma estabilidade bastante considerável: em todas as 5 execuções, os dois estudos nucleares (LAURA e o protocolo NCT03521154) surgiram, sistematicamente, nas primeiras posições. Os restantes lugares oscilaram moderadamente entre revisões/meta-análises recentes (p.ex., Luo 2025; Chen 2024; Li 2024; Dai 2025), refletindo a variabilidade esperada dos LLMs, mas sem perda da cobertura do "*core evidence*", o que é fundamental. Ocasionalmente entraram estudos menos elegíveis (SPIRAL-0,

LOGIK0902/OLCSG0905, subanálise PACIFIC) que acabam por não ser relevantes para o caso, daí terem sido anotados como irrelevantes.

Para avaliar os resultados obtidos de uma forma mais objetiva, optou-se por calcular dois tipos de métricas, nomeadamente, métricas de estabilidade (para avaliar a variabilidade e robustez do MCP na recuperação de artigos) e métricas funcionais (para avaliar a qualidade e relevância dos artigos recuperados, contrastando com o *gold standard*):

1. **Métricas de estabilidade (top-5)**: para cada *query,* recolher o top-5, em cada repetição:
    a. **Similaridade de Jaccard@5** (entre listas): $|A \cap B|/|A \cup B|$ (razão entre o número de elementos em comum e o número total de elementos distintos presentes em ambos, neste caso, entre os artigos recuperados e o *gold standard*);
    b. **RBO@5** (p=0,9) (*Rank Biased Overlap*): *overlap* ponderado das primeiras posições (a explicação do cálculo desta métrica encontra-se no Anexo G);
    c. **Stability@5**: proporção de itens que aparecem em todas as repetições;
    d. **Unique@5**: nº médio de PMIDs distintos no conjunto de repetições;
    e. **Volatility@5**: $1 - RBO$;

2. **Métricas funcionais**:
    a. **Precision@5:** proporção de documentos relevantes dentro dos 5 resultados (medida de eficiência):

$$P@5 = \frac{nº\ documentos\ relevantes\ nos\ top-5}{5}$$

    b. **Recall@5:** proporção de documentos relevantes que foram recuperados nos top-5 em relação a todos os relevantes conhecidos (*gold standard*) (medida de cobertura):

$$Recall@5 = \frac{nº\ documentos\ relevantes\ nos\ top-5}{nº\ total\ de\ documentos\ relevantes\ no\ gold\ standard}$$

    c. **NDCG@5:** *Normalized Discounted Cumulative Gain*, mede a qualidade do *ranking*, atribuindo mais valor aos artigos relevantes que aparecem nos primeiros lugares (varia entre 0 e 1. Medida de *ranking*):

        i. Atribuição da relevância a cada documento (1 → relevante; 0 → irrelevante);
        ii. Cálculo do DCG@5 (*Discounted Cumulative Gain*):

$$DCG@5 = \sum_{i=1}^{5} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

iii. Normalização pelo IDCG@5 (trata-se do DCG se estivesse ordenado perfeitamente, i.e., o valor máximo possível):

$$NDCG@5 = \frac{DCG@5}{IDCG@5}$$

d. **Recência**: % em <24 meses;

$$R@5 = \frac{nº\ de\ artigos\ publicados\ há\ menos\ de\ 2\ anos}{5}$$

e. **Integridade dos metadados** (títulos, autores, PMIDs, data e *abstract*)**:**

$$I@5 = \frac{nº\ de\ artigos\ recuperados\ com\ metadados\ corretos}{5}$$

f. **Latência**: mediana e p95 (tempo de resposta típico e piores casos)

Tabela 9 - Resultados (valores médios dos 3 relatórios) das Métricas de Avaliação dos Artigos Recuperados pelo BioMCP

Métrica	Claude Sonnet 4 (Anthropic)	Gemini 2.5 Pro (Google)	GPT-5 (OpenAI)
Jaccard@5	0,39	0,39	0,57
RBO@5	0,56	0,52	0,76
Volatility@5	0,44	0,48	0,24
Stability@5	0,12	0,13	0,21
Unique@5	11,00	10,67	8,33

Os resultados obtidos para as métricas de estabilidade (Tabela 9) evidenciam diferenças claras entre os três modelos avaliados, com o GPT-5 a destacar-se de forma consistente, apresentando melhores resultados na similaridade de Jaccard (0,57) e maior RBO@5 (0,76), o que reflete listas de resultados mais homogéneas e estáveis entre execuções, com preservação da ordem dos artigos recuperados e menor volatilidade (0,24). Por contraste, tanto o Claude Sonnet 4 como o Gemini 2.5 Pro registaram valores mais baixos (similaridade de Jaccard: 0,39 e RBO@5: 0,56 e 0,52, respetivamente), associados a uma maior instabilidade na ordenação (volatilidade de

0,44 e 0,48, respetivamente). A métrica *Stability* revelou valores baixos em todos os casos, sugerindo dificuldade generalizada em manter artigos idênticos em todas as repetições, embora o GPT-5 volte a evidenciar melhor desempenho relativo (0,21). Quanto à exclusividade dos resultados (métrica *Unique*), Claude (11,00) e Gemini (10,67) recuperaram um número mais alargado de PMIDs distintos, contrastando com o GPT-5 (8,33), que apresentou menor diversidade. Este resultado pode, eventualmente, sugerir que tanto o Claude e Gemini exploram de forma mais ampla o espaço documental. Contudo, analisando mais minuciosamente o processo de pesquisa e raciocínio de ambos os modelos, é possível verificar que a pesquisa realizada pelo GPT-5 é, em grande parte das situações, bem mais extensa e ampla comparativamente aos restantes. O menor valor desta métrica é, na verdade, positivo, na medida em que o modelo privilegia consistência e replicabilidade.

Tabela 10 - Resultados das Métricas de Avaliação dos Artigos Recuperados pelo BioMCP para o Parecer 1

	Jaccard@5			RBO@5			Volatility@5		
Modelo	**Média**	**IC (95%)**	**p95**	**Média**	**IC (95%)**	**p95**	**Média**	**IC (95%)**	**p95**
Claude Sonnet 4 (Anthropic)	0,48	(0,31; 0,65)	0,85	0,70	(0,60; 0,81)	0,92	0,30	(0,19; 0,40)	0,46
GPT-5 (OpenAI)	0,57	(0,48; 0,66)	0,67	0,77	(0,71; 0,83)	0,86	0,23	(0,17; 0,29)	0,33
Gemini 2.5 Pro (Google)	0,33	(0,22; 0,44)	0,56	0,50	(0,39; 0,61)	0,71	0,50	(0,39; 0,61)	0,72

Tabela 11 - Resultados das Métricas de Avaliação dos Artigos Recuperados pelo BioMCP para o Parecer 2

	Jaccard@5			RBO@5			Volatility@5		
Modelo	**Média**	**IC (95%)**	**p95**	**Média**	**IC (95%)**	**p95**	**Média**	**IC (95%)**	**p95**
Claude Sonnet 4 (Anthropic)	0,34	(0,26; 0,43)	0,43	0,53	(0,44; 0,61)	0,65	0,47	(0,39; 0,56)	0,66
GPT-5 (OpenAI)	0,56	(0,41; 0,72)	0,85	0,78	(0,69; 0,87)	0,94	0,22	(0,13; 0,31)	0,39
Gemini 2.5 Pro (Google)	0,40	(0,31; 0,49)	0,56	0,48	(0,39; 0,58)	0,67	0,52	(0,42; 0,62)	0,71

Tabela 12 - Resultados das Métricas de Avaliação dos Artigos Recuperados pelo BioMCP para o Parecer 3

	Jaccard@5			RBO@5			Volatility@5		
Modelo	**Média**	**IC (95%)**	**p95**	**Média**	**IC (95%)**	**p95**	**Média**	**IC (95%)**	**p95**
Claude Sonnet 4 (Anthropic)	0,35	(0,17; 0,52)	0,74	0,44	(0,28; 0,60)	0,80	0,56	(0,41; 0,72)	0,72
GPT-5 (OpenAI)	0,59	(0,37; 0,84)	1,00	0,73	(0,53; 0,92)	1,00	0,27	(0,08; 0,47)	0,62
Gemini 2.5 Pro (Google)	0,43	(0,28; 0,59)	0,74	0,57	(0,40; 0,75)	0,87	0,43	(0,25; 0,60)	0,72

Analisando os resultados separadamente (por relatório) (Tabela 10, Tabela 11 e Tabela 12), verificamos que, no 1º relatório, o GPT-5 evidencia clara superioridade, com os valores médios superiores já mencionados acima a serem acompanhados de intervalos de confiança estreitos, refletindo, assim, maior consistência que os restantes. O p95 mostra que o Claude (0,85 em Jaccard e 0,92 em RBO) apresentou o melhor resultado de entre os 3, ainda que os seus valores médios e IC revelem maior variabilidade. O Gemini apresenta resultados menos robustos, com médias significativamente inferiores (Jaccard de 0,33; RBO de 0,50) e volatilidade mais elevada (0,50), confirmando um comportamento algo instável.

No Parecer 2, a mesma tendência se mantém: o GPT-5 regista novamente os melhores valores (Jaccard de 0,56; RBO de 0,78; volatilidade de 0,22), mantendo margens de incerteza relativamente reduzidas. Neste caso, o Claude, não conseguiu manter a fasquia obtida no 1º parecer, com os melhores resultados (p95) a apresentarem um desempenho consideravelmente inferior (Jaccard de 0,43 e RBO de 0,65). O mesmo se passa no último relatório, que acaba por reforçar a sumpremacia do GPT-5 para a tarefa em questão, com o Claude e Gemini a apresentarem médias inferiores e maior volatilidade, com intervalos relativamente alargados.

Em síntese, a análise por parecer confirma que o GPT-5 não só apresenta melhores desempenhos médios, como o faz de forma mais consistente e previsível, com intervalos de confiança mais estreitos e valores p95 elevados, consolidando a sua robustez em diferentes cenários. No contexto da FC, onde a previsibilidade e a fiabilidade dos resultados assumem particular relevância para apoiar a decisão, o perfil do mais recente modelo da OpenAI revela-se, claramente, o mais adequado.

Ainda assim, esta análise não se revela suficiente para tomar uma decisão final. Isto porque o modelo pode ser bastante estável e previsível, mas, ainda assim, não recuperar artigos suficientemente relevantes para responder aos pedidos da CFT. Por isso, torna-se necessário o cálculo de uma tipologia de métricas que meça isso mesmo: a qualidade e relevância dos artigos recuperados (Gráfico 3).

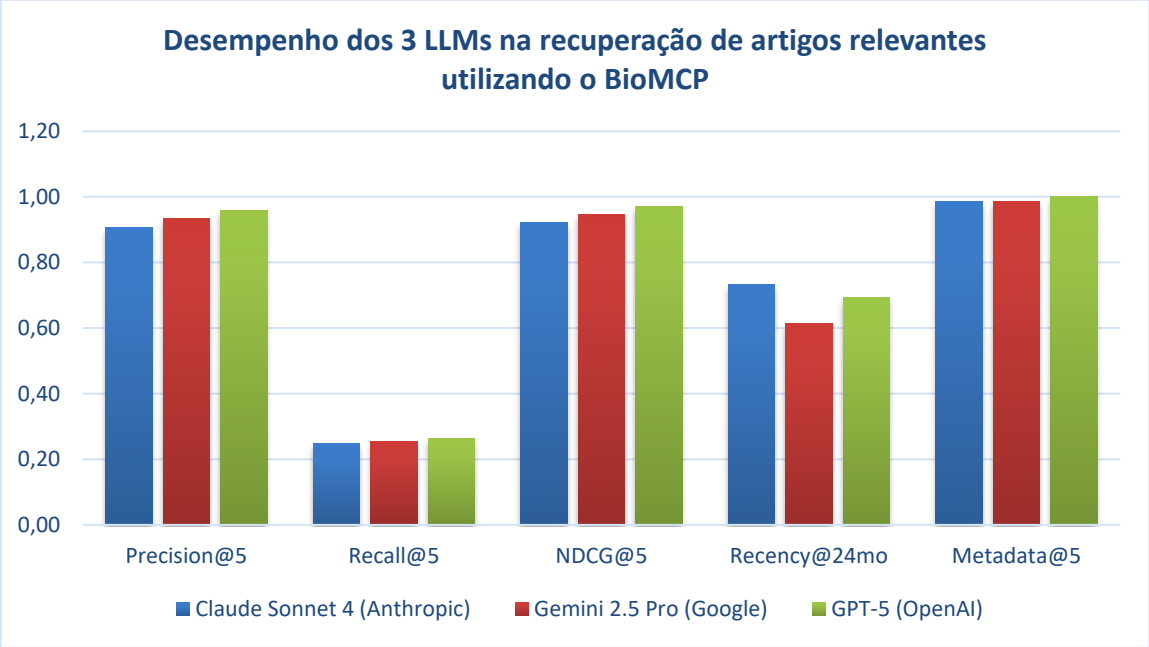


Gráfico 3 - Resultados das Métricas Funcionais no Processo de IR utilizando o BioMCP

Os resultados mostram níveis muito elevados de precisão nos três modelos (≥0,91), com o GPT-5 novamente a liderar (0,96), o que indica que a maioria dos artigos recuperados é, efetivamente, relevante. A ordenação das listas também foi adequada em todos os casos (NDCG@5 ≥0,92), com vantagem clara para o GPT-5 (0,97), confirmando que os artigos mais pertinentes tendem a surgir nas primeiras posições. Em contrapartida, o *recall* revelou-se baixo e semelhante em todos os modelos (0,25–0,26), evidenciando limitações na cobertura dos artigos relevantes disponíveis. Contudo, estes resultados justificáveis tendo em conta que a quantidade de artigos no *gold standard* é significativa, relativamente ao total de artigos recuperados (limitado a 5). Ou seja, os valores baixos nesta métrica devem-se, sobretudo, a opção metodológica de limitar o nº total de artigos recuperados pelo MCP.

Quanto à recência da literatura recuperada, o Claude destacou-se (0,73), seguido do GPT-5 (0,69) e do Gemini (0,61). Relativamente a esta métrica, vale a pena ressaltar que 2 dos 3 relatórios foram redigidos em 2023, com cerca de 40% dos artigos presentes no *gold standard* a serem publicados antes de 2023. Ainda assim, optou-se por considerar esta métrica por ser imperativo que o MCP recupere a evidência científica mais relevante e recente para o caso em questão. Por fim, quanto à completude e integridade dos metadados, mencionar que, apesar de estes terem sido identificados corretamente em praticamente todas as situações, os modelos da Anthropic e Google identificaram erradamente o PMID errado de um artigo numa das execuções. Além disso, o Gemini, na 1ª execução para o relatório 1, alucinou, "recuperando" um artigo que não existia de todo.

Relativamente aos tempos de execução (desde o início da pesquisa até ao *output* final dado pelo LLM), houve também algumas discrepâncias significativas entre modelos, tal como observado abaixo (Gráfico 4).

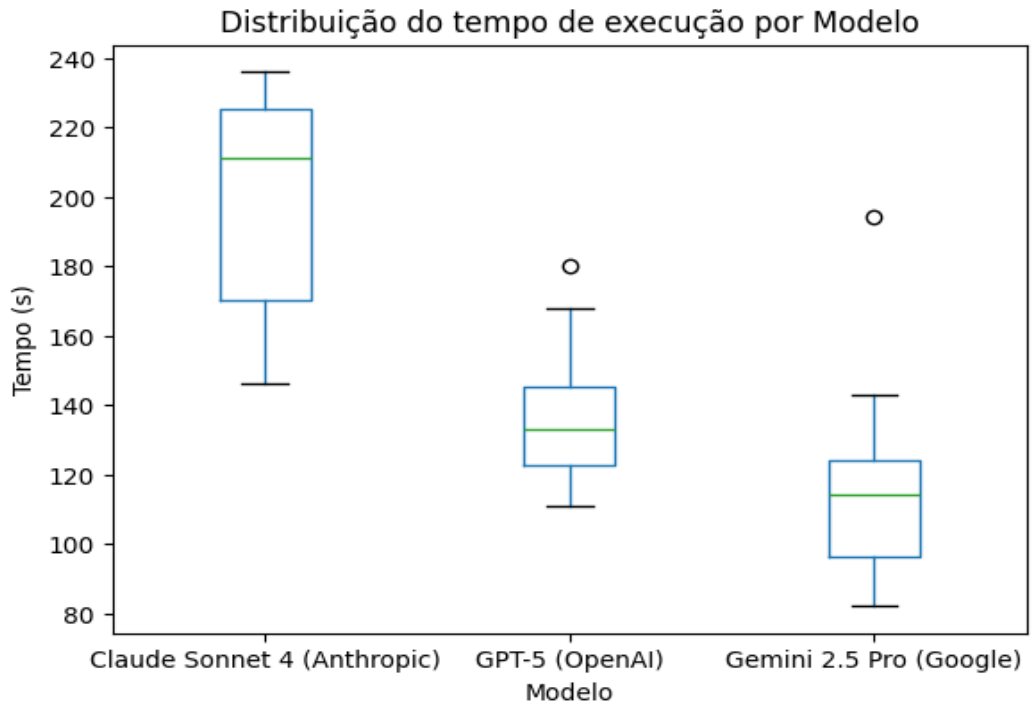


Gráfico 4 - Distribuição do Tempo de Execução Médio da Pesquisa, por Modelo

O Claude Sonnet 4 revelou-se o mais lento, com um tempo médio global de 200,9 segundos, em contraste com o Gemini 2.5 Pro, que foi o mais rápido (115,7 segundos). O GPT-5 posicionou-se num patamar intermédio (137,3 segundos). Considerando o número de chamadas às diferentes *tools* do BioMCP e a qualidade do *output* final, estas diferenças entre modelos podem ser consideradas como pouco significativas, comparativamente ao ganho efetivo proporcionado. De facto, a etapa de pesquisa e síntese da evidência científica constitui, tradicionalmente, a fase mais morosa da redação destes relatórios, podendo prolongar-se por vários dias, ou até mesmo uma semana, de acordo com os profissionais da UFC. Neste sentido, os tempos registados pelos modelos representam uma melhoria substancial na eficiência do processo, pelo menos, na elaboração de um primeiro *draft*.

Em síntese, os resultados demonstram que o GPT-5 acaba por ser o modelo mais consistente e preciso, oferecendo maior fiabilidade para aplicação no caso em questão. O Claude Sonnet 4 revelou vantagens na atualização dos artigos e maior diversidade, podendo ser útil em fases exploratórias, enquanto o Gemini 2.5 Pro mostrou menor robustez e episódios pontuais de erros, reduzindo a sua adequação para o caso.

Uma das limitações mais evidentes desta análise prende-se com a ausência de validação por parte dos profissionais de saúde. Além disso, estando a recuperação restrita a apenas cinco artigos, ainda que todos sejam individualmente relevantes para o pedido da CFT, a sua combinação pode não ser suficiente para assegurar uma resposta completa. Assim, torna-se indispensável considerar, em fases futuras, uma avaliação qualitativa que envolva diretamente os peritos, de forma a garantir que a evidência recuperada corresponde de facto às necessidades.

5. Prova de Conceito

5.1. Descrição do Protótipo

O protótipo, desenvolvido em Streamlit, foi concebido para operacionalizar todas as etapas do processo de elaboração de um parecer técnico, a partir dos *inputs* fornecidos pela CFT (título, enquadramento e caso clínico) e sequenciando as diversas etapas mostradas nos capítulos anteriores. A plataforma apresenta-se de forma modular e estruturada em diferentes secções (Figura 13), correspondendo à lógica do relatório final, ou seja: na secção 1 encontra-se a fundamentação farmacológica, que deriva, essencialmente, da informação contida no RCM e devidamente normalizada; a secção 2 integra os estudos relevantes identificados através do BioMCP, assegurando uma visão sintetizada da evidência disponível; a secção 3, referente a pareceres de sociedades médicas e científicas, não foi alvo de desenvolvimento nesta fase, dada a necessidade de técnicas adicionais de pesquisa automatizada em bases específicas. Isto é, dependendo do tema, as entidades relevantes vão variar bastante. Cada entidade tem a sua base de dados, a sua forma de aceder aos documentos, etc.. Tal exigiria explorar o conceito de agentes que fazem pesquisa na web, ou até MCP que realizam esta pesquisa. Com o intuito de reduzir o âmbito e focalizar naquilo que seria possível, dados os limites temporais para desenvolvimento deste trabalho, optou-se por explorar esta secção em fases futuras; a secção 4 corresponde às considerações finais, nas quais se procede a uma síntese crítica e integrativa da informação reunida nas secções anteriores, destacando consistências, incoerências, lacunas e incertezas; por fim, a secção 5 apresenta a conclusão, materializando-se no momento decisório do parecer: aprovar ou não a utilização do medicamento na situação clínica em causa, bem como definir eventuais condições associadas, tais como critérios de monitorização, suspensão ou necessidade de reavaliação periódica.

Por motivos de simplificação no desenvolvimento do protótipo, o fluxo de dados segue uma lógica sequencial (cada secção é trabalhada separadamente e sequencialmente, da primeira secção até à secção final).

Figura 13 - Página inicial do protótipo desenvolvido em Streamlit

Todo este processo é descrito na própria UI do protótipo, na secção "*Status"* (Figura 14).

Resumidamente:

Secção "***Fundamentação | Farmacologia*"**:

1. Após a inserção dos *inputs* por parte do utilizador, o LLM selecionado começa por identificar o(s) fármaco(s) em análise, bem como a categorização do pedido numa das 6 categorias mencionadas no capítulo 3;
2. São extraídos o(s) RCM do(s) fármaco(s) identificado(s) da base de dados da EMA, bem como a informação de prescrição do(s) mesmo(s) da base de dados da FDA;
3. Os RCM são seccionados e indexados em base de dados vetorial;
4. Com base no enquadramento do novo pedido, é calculado, dentro de todos os pareceres já redigidos no passado, qual o parecer mais semelhante (da mesma categoria);
5. O LLM, com base no parecer exemplo e no pedido em questão, realiza inúmeras *queries* à base de dados que contém os *embeddings* do RCM previamente indexados, recuperando, assim, as secções mais relevantes;
6. O LLM, com base nas secções recuperadas, redige a secção "Fundamentação | Farmacologia";

Secção "***Fundamentação | Estudos Relevantes*"**:

7. O LLM, com base no parecer exemplo e no pedido em questão, elabora uma *query* onde descreve, detalhadamente, a evidência científica que pretende que o BioMCP recupere;
8. O BioMCP, executa uma multiplicidade de chamadas às *tools* para recuperar artigos que respondam ao pedido efetuado pelo LLM, terminando com uma síntese dessa mesma pesquisa, os artigos identificados e respetivos metadados (Secção "Raciocínio IA" permite que utilizador siga esse o processo de raciocínio do LLM à medida que vai interagindo com o BioMCP (Figura 15);
9. O LLM, com base na síntese redigida pelo BioMCP e no parecer exemplo, redige a secção "Fundamentação | Estudos Relevantes";

Secção "***Considerações Finais":***

10. O LLM, com base no parecer exemplo, no pedido em questão e em todo o relatório redigido até ao momento, redige a secção "Considerações Finais";

Secção "***Conclusão":***

11. O LLM, com base no parecer, exemplo, no pedido em questão e em todo o relatório redigido até ao momento, redige a secção "**Conclusão**";

Secção "***Referências Bibliográficas":***

12. São incluídas as referências bibliográficas, com *hyperlinks,* do(s) RCM da EMA e prescrição(ões) de informação da FDA, bem como artigos recuperados pelo BioMCP.

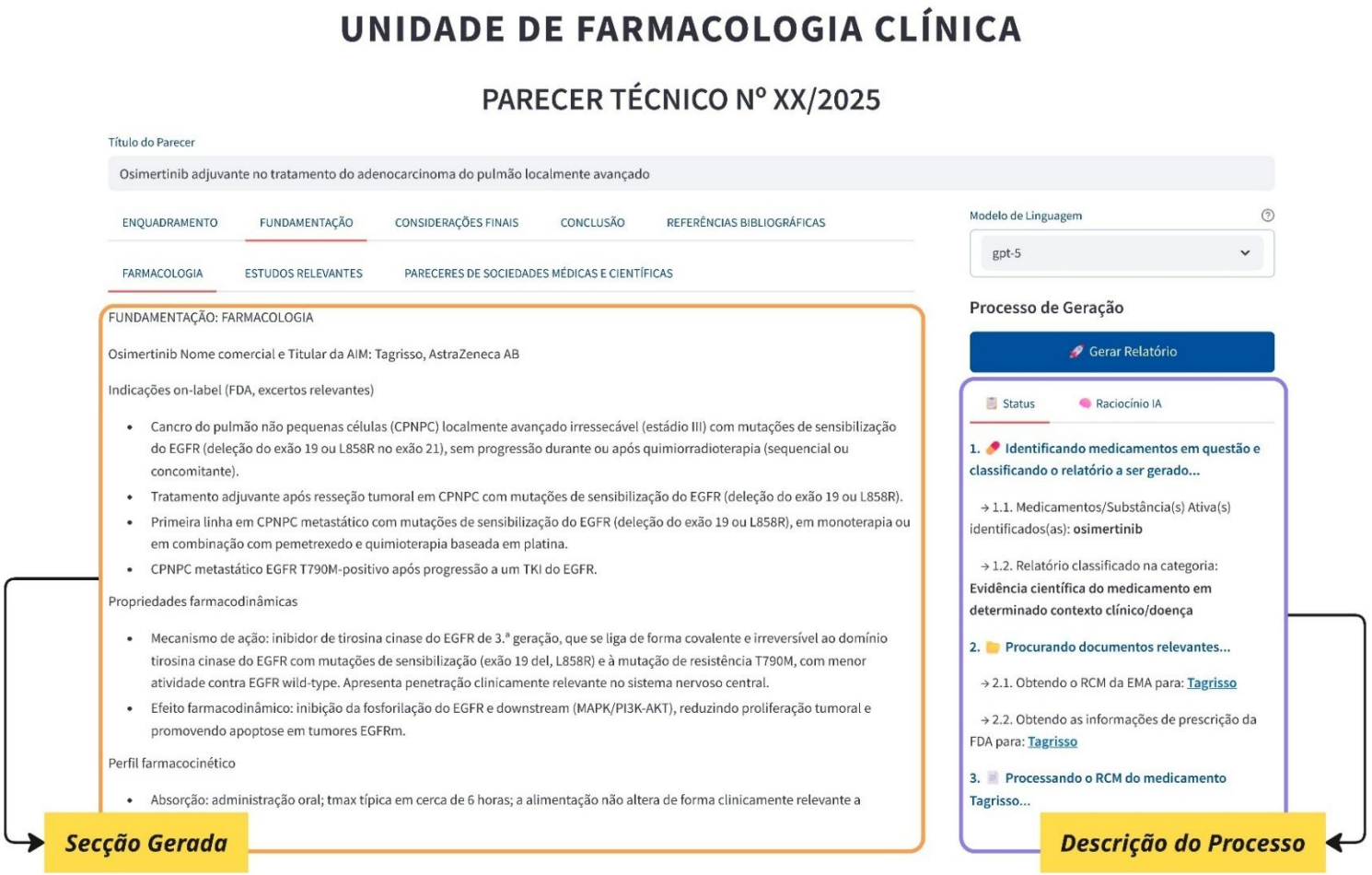


Figura 14 - Geração de um Relatório Exemplo utilizando o Protótipo Desenvolvido

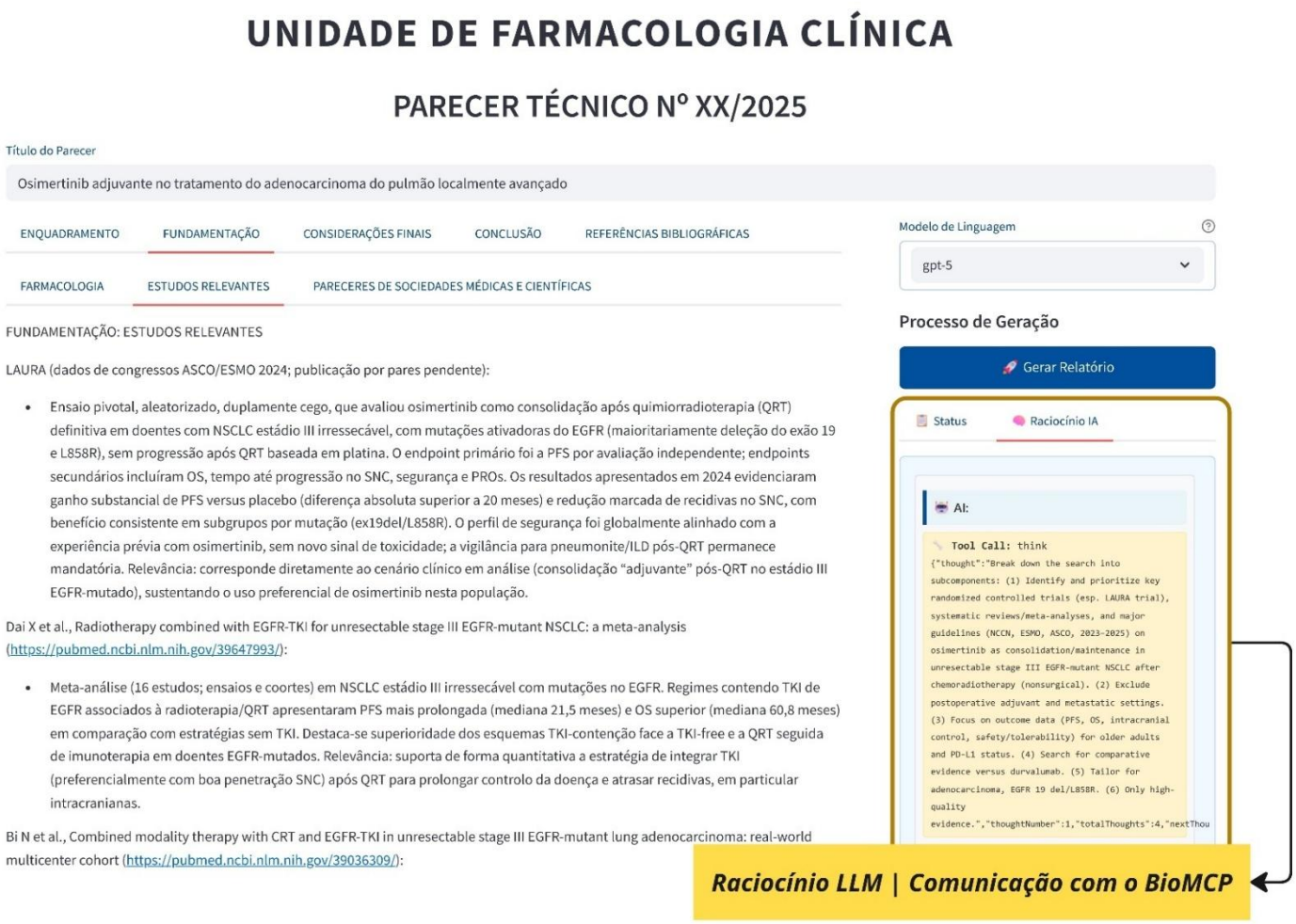


Figura 15 - Demonstração da Interação LLM - BioMCP na Protótipo Desenvolvido

Nos passos acima, nas situações em que são mencionadas as etapas realizadas pelo LLM, para todas as elas foi elaborado um *system prompt* específico para que o modelo saiba exatamente aquilo que necessita de fazer a cada momento, que informação utilizar e qual deve ser o formato dos *outputs.* Estes *prompts* encontram-se no Anexo H.

5.2. Geração e Avaliação dos Relatórios Gerados

Seguindo a mesma metodologia aplicada nos capítulos anteriores, o protótipo desenvolvido foi utilizado para gerar os três relatórios previamente mencionados, recorrendo aos três LLMs em análise. A avaliação rigorosa do seu conteúdo em termos de qualidade científica, precisão e completude só poderá ser devidamente assegurada com o contributo direto dos profissionais de farmacologia. No entanto, é possível, numa primeira instância, recorrer a algumas estratégias que permitem aferir se os relatórios produzidos correspondem, minimamente, àquilo que seria expectável. Nesse sentido, optou-se pela aplicação dos seguintes métodos:

1. **Avaliação baseada em características linguísticas (métricas quantitativas de texto):** quantificar a aderência lexical e terminológica dos relatórios em relação ao domínio clínico e ao enunciado do parecer. Nesta avaliação calcular-se-ão duas métricas, nomeadamente, a frequência de termos relevantes, isto é, total de ocorrências de palavras-chave obrigatórias que surgem no relatório, e a cobertura lexical, ou seja, a percentagem de palavras-chave obrigatórias que surgem no relatório (ambas as métricas calculadas mediante uma lista pré-definida de *keywords*, que poderá ser consultada no Anexo I);
2. **Avaliação por LLM (*LLM-as-a-Judge*):** explorar a utilização de modelos de linguagem para avaliar os resultados gerados por um outro modelo. O procedimento inicia-se no fornecimento ao modelo um relatório de referência (produzido por profissionais, possível visto que havia acesso ao histórico de pareceres). De seguida, solicita-se a avaliação comparativa do relatório gerado vs. relatório de referência em dimensões como *clareza*, *precisão científica*, *adequação ao pedido clínico*, *completude*. O LLM classifica, então, cada uma das secções utilizando uma escala ordinal (1–5) por dimensão, complementada com justificação textual.
3. ***Checklist* estruturada (com base em critérios clínicos):** verificar de forma objetiva se o relatório cobre todos os pontos essenciais para responder ao pedido da CFT. Este é um passo a ser realizado por profissionais de farmacologia. Nesta avaliação, recorre-se a *checklists* que contêm os pontos essenciais que devem ser abordados em cada secção. De seguida, é feita uma simples verificação binária (*sim/não*) ou gradual (0–2 pontos por item) conforme o relatório gerado cumpre com o esperado ou não. Como resultado, obtém-se uma matriz de cobertura que mostra se o relatório responde ao pedido em todas as suas dimensões essenciais.

Nota: Este 3º ponto não será realizado no âmbito deste projeto na sua completude, visto que a validação por parte dos especialistas dos *outputs* gerados pelo sistema será feito numa fase posterior. Ainda assim, foi desenvolvida uma *checklist* para o 1º relatório em análise que poderá servir como inspiração e ponto de partida para realizar essa respetiva avaliação.

5.2.1. Métricas Linguísticas

Para o cálculo das métricas linguísticas definidas, foi necessário desenvolver uma lista abrangente de termos considerados determinantes para a temática de cada relatório. Para além da identificação das *keywords* principais, tornou-se igualmente essencial incluir as suas variações linguísticas, de forma a evitar que conceitos relevantes fossem excluídos por diferenças de denominação. Por exemplo, no relatório nº 1, "adjuvante" constitui um termo chave, mas que pode surgir na forma "adjuvância", dependendo da construção frásica. De modo semelhante, "concomitante" pode aparecer como "em concomitância" e, mais crítico ainda, a própria patologia pode ser descrita de múltiplas formas, como "cancro do pulmão de não pequenas células", "carcinoma do pulmão de não pequenas células", "carcinoma pulmonar de não pequenas células", "CPNPC" ou "NSCLC".

O primeiro passo consistiu, portanto, em criar uma lista extensiva de *keywords* e respetivas variações para cada relatório (Anexo I). Seguidamente, para simplificar a análise e torná-la mais informativa, optou-se por agrupar os termos em categorias, em vez de avaliá-los individualmente. Assim, "osimertinib" e "Tagrisso" foram incluídos na categoria Fármaco, todas as variações da doença mencionadas integraram a categoria CPNPC, e termos como "EGFR", "exão 19" e "del19" foram reunidos na categoria Biomarcadores e mutações.

Com esta lista estruturada e organizada por categorias, tornou-se então possível proceder ao cálculo das métricas linguísticas definidas, cujos resultados se encontram apresentados nas Tabela 13, Tabela 14 e Tabela 15.

Tabela 13 - *Term Frequency*, por categoria, no relatório 1 (gerado)

Categorias | Relatório 1	OpenAI	Anthropic	Gemini
Fármaco	46	30	29
Contexto	26	15	16
Doença	60	45	35
Biomarcadores e mutações	40	34	35
Ensaio clínico	9	11	14
Endpoints e estatística	22	16	12
Eventos adversos	7	5	5
Tratamento prévio	4	4	4
Posologia e Farmacocinética	0	0	0
Diretrizes / agências	10	10	10

Tabela 14 – *Term Frequency*, por categoria, no relatório 2 (gerado)

Relatório 2	OpenAI	Anthropic	Gemini
Fármaco	27	33	29
Doença	21	11	12
Regimes on-label (RCM) e posologia	18	10	2
Estratégias off-label / intensificação	31	38	22
Ensaios pivotais	5	0	0
Marcadores clínicos e laboratoriais	16	4	0
Tratamentos prévios	6	1	0
Outcomes	0	0	0
Segurança e contraindicações	3	2	4
Abreviaturas	37	18	6

Tabela 15 - *Term Frequency*, por categoria, no relatório 3 (gerado)

Relatório 3	OpenAI	Anthropic	Gemini
Fármaco	30	24	22
Doença	26	21	14
Caso clínico	41	20	6
Posologia	14	2	0
Monitorização	40	10	4
Endpoints	1	0	0
Marcadores	0	0	0
Reações adversas	1	0	0
Contraindicações e precauções	11	0	0

Os resultados obtidos no cálculo do TF evidenciam diferenças assinaláveis entre os três modelos analisados, refletindo não apenas o volume de informação recuperada, mas também a abrangência temática de cada relatório. No conjunto dos três relatórios, o GPT-5 destacou-se de forma consistente, apresentando maior completude e diversidade de termos nas categorias mais relevantes, como fármacos, contexto, doença, biomarcadores, posologia, caso clínico e monitorização. O Claude Sonnet 4 revelou um desempenho intermédio, com melhor cobertura em algumas dimensões específicas, como estratégias *off-label*, mas falhas em áreas essenciais como marcadores e contraindicações. Já o Gemini demonstrou resultados globalmente mais fracos e inconsistentes, com omissões críticas em variadas categorias (nomeadamente, *endpoints*, marcadores, reacções adversas e contraindicações).

Tabela 16 - Cobertura lexical em cada relatório gerado

Cobertura global (%)	OpenAI	Anthropic	Gemini
Relatório 1	68,98	54,9	70,78
Relatório 2	63,73	48,51	43,07
Relatório 3	61,15	44,59	41,48

Quando analisada a cobertura lexical global (percentagem de *keywords* que surge, pelo menos, uma vez) (Tabela 16), observa-se um padrão menos linear do que o identificado nas métricas anteriores. No Relatório 1, o Gemini apresenta maior cobertura (70,78%), seguido do GPT-5 (68,98%) e do Claude (54,9%), sugerindo maior diversidade na utilização de termos da lista de *keywords*. Já no Relatório 2, o GPT-5 assume clara vantagem (63,73%) face ao Claude (48,51%) e, sobretudo, ao Gemini (43,07%), enquanto no Relatório 3 repete-se esta tendência, com o GPT-5 a liderar (61,15%), superando novamente Claude (44,59%) e Gemini (41,48%). Estes resultados demonstram que, embora o Gemini consiga abranger mais termos num dos relatórios, é o GPT-5 que revela maior consistência global, assegurando uma correspondência mais equilibrada entre relatórios.

Ainda que esta análise permita obter uma noção da abrangência terminológica e da correspondência com a lista de *keywords* definida, importa sublinhar que estas métricas não capturam o contexto ou a semântica das frases. Ou seja, o simples facto de uma palavra surgir num relatório, não garante que esteja a ser utilizada no enquadramento clínico adequado, o que limita fortemente a possibilidade de tirar conclusões sólidas a partir destes resultados.

5.2.2. LLM-as-a-judge

Com o objetivo de mitigar a limitação anteriormente identificada, recorreu-se à estratégia de avaliação *LLM-as-a-judge*. Esta abordagem consiste em utilizar um modelo de linguagem para avaliar os relatórios gerados por outro, em diferentes dimensões, atribuindo uma pontuação numa escala de 1 a 5. As dimensões definidas para avaliação foram: clareza, precisão científica, adequação ao pedido e completude.

A relação entre avaliador e avaliado foi estabelecida da seguinte forma:

- Relatório gerado pelo GPT-5 → avaliado pelo Claude Sonnet 4
- Relatório gerado pelo Claude Sonnet 4 → avaliado pelo GPT-5
- Relatório gerado pelo Gemini 2.5 Pro → avaliado pelo GPT-5

Esta escolha deveu-se ao desempenho dos 3 modelos em análises anteriores.

Tabela 17 - Pontuações atribuídas aos relatórios gerados nas 4 dimensões definidas (valores médios)

Modelo de Geração	Modelo Juíz	Clareza	Precisão Científica	Adequação	Completude
Claude Sonnet 4	GPT-5	4	3	3	2
Gemini 2.5 Pro	GPT-5	4	3	2	2
GPT-5	Claude Sonnet 4	4	4	4	3

De forma global, observa-se que os relatórios gerados pelo GPT-5 obtiveram, consistentemente, uma melhor avaliação, alcançando valores médios de 4 em todas as dimensões, exceto completude, com um valor de 3. Os relatórios produzidos pelo Claude Sonnet 4 estabilizaram numa média geral de 3, com o Gemini 2.5 Pro a apresentar um desempenho relativamente inferior, menos robusto.

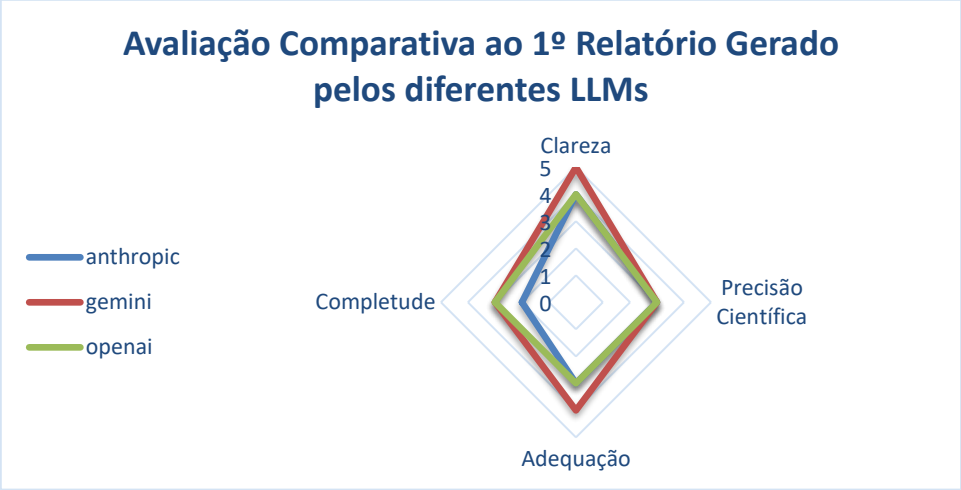


Gráfico 5 - Pontuações atribuídas ao 1º relatório gerado pelos diferentes LLMs

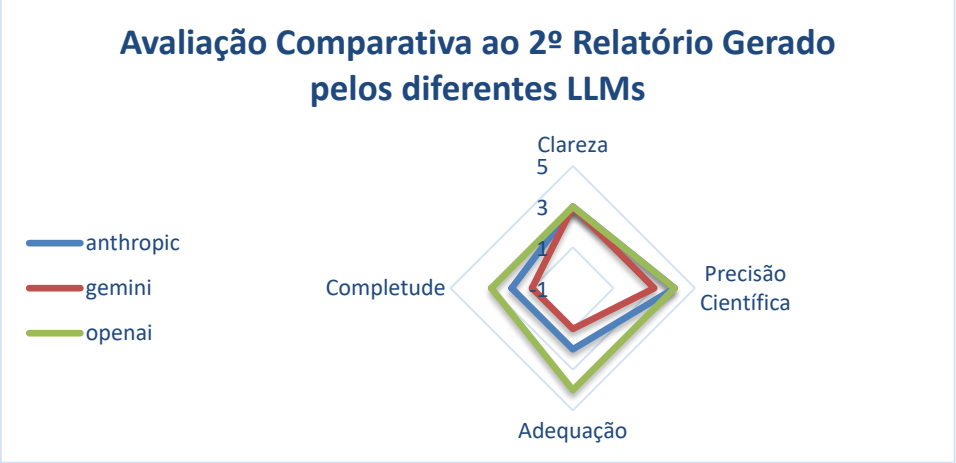


Gráfico 6 - Pontuações atribuídas ao 2º relatório gerado pelos diferentes LLMs

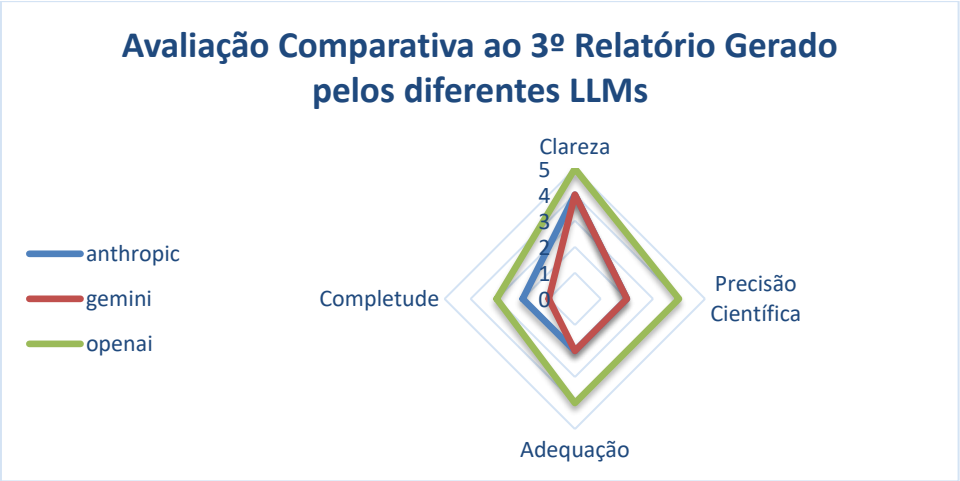


Gráfico 7 - Pontuações atribuídas ao 3º relatório gerado pelos diferentes LLMs

Quando a análise é detalhada por parecer, surgem algumas nuances. No Parecer 1, as diferenças entre modelos não são tão marcantes, com uma superioridade do Gemini, especificamente no que diz respeito à clareza (5) e adequação do pedido (4), com um Claude a apresentar uma menor completude (2). No Parecer 2, o desempenho do Gemini é claramente inferior (notas de 1 em adequação e completude), evidenciando dificuldade em responder de forma minimamente satisfatória ao pedido. Já o GPT-5, ainda que não tenha atingido níveis máximos, mostrou maior consistência (valores entre 3 e 4), demonstrando capacidade de adaptação ao parecer (4). No Parecer 3, repete-se a tendência: Claude e Gemini voltam a registar avaliações menores (notas de 2 em precisão e adequação), enquanto o GPT-5 assegura resultados equilibrados, com elevada clareza (5), precisão e adequação bastante satisfatórias e uma completude razoável (3).

De notar que, juntamente com as pontuações, o modelo providenciou justificações para as suas atribuições, que permitiram garantir a ausência de alucinações. Importa ainda reconhecer que a abordagem *LLM-as-a-judge* apresenta limitações, uma vez que capta apenas parcialmente a qualidade real dos relatórios e pode refletir enviesamentos do próprio modelo avaliador.

Apesar da aparente superioridade do GPT-5, os resultados não atingem a robustez necessária para uma área tão exigente como a FC. Torna-se, por isso, evidente a necessidade de melhorias significativas no processo de geração, de modo a aumentar a confiança e a fiabilidade dos *outputs*. Apenas com este reforço será possível garantir que os relatórios automatizados constituem um apoio efetivo e seguro ao trabalho dos profissionais da UFC, assegurando rigor científico e utilidade prática na tomada de decisão.

5.2.3. Checklists para validação manual

Como já mencionado anteriormente, no âmbito deste projeto, este método de avaliação manual, com o auxílio de profissionais de farmacologia, não foi explorado a fundo. Ainda assim, foi desenvolvido um exemplo de *checklist* aplicada ao relatório nº 1 (Tabela 18, Tabela 19 e Tabela 20), que poderá servir como ponto de partida para trabalhos futuros. Esta *checklist* pretende ilustrar de que forma é possível estruturar a avaliação manual com base em critérios clínicos objetivos, permitindo verificar se o relatório cobre todos os pontos essenciais solicitados para responder ao pedido da CFT.

Tabela 18 - *Checklist* para validação da Secção "Fundamentação | Farmacologia"

Bloco Principal	*Pontos essenciais*	*Relevância*
Identificação do medicamento	- **Nome comercial**: Tagrisso; - **Titular da AIM** (Autorização de Introdução no Mercado): AstraZenca	- Garantia da **correta identificação do fármaco**.
Indicações terapêuticas	- **Indicações terapêuticas para as quais o medicamento é aprovado pela:** **1. EMA** (*European Medicines Agency*); **2. FDA** (*Foods and Drugs Administration*);	- Verificação do **enquadramento do pedido realizado pela CFT nas indicações terapêuticas já aprovadas (*on-label*) ou se falamos de um uso extrapolado (*off-label*)**. Tal será crítico na fundamentação necessária e/ou restrição da decisão.
Mecanismo de ação/Propriedades farmacodinâmicas	- **Inibidor irreversível de EGFR-TKI.** Por outras palavras, falamos da ação (irreversível) do fármaco no bloqueio da atividade de uma proteína específica, que estimula o crescimento e sobrevivência das células tumorais**;** **- Ativo contra mutações sensibilizantes** (caso clínico em questão) **e T790M** (após algum tempo de tratamento, os tumores desenvolvem uma mutação (T790M) que os torna resistentes. O fármaco em questão bloqueia, também, esta mutação de resistência).	- Demonstração do **racional do uso do medicamento em doentes com mutações do gene EGFR** (Recetor do Fator de Crescimento Epidérmico), no caso clínico em questão, **a deleção do exão 19**.
Perfil farmacocinético	- **Depuração plasmática (**velocidade com que o fármaco é eliminado do plasma)**:** 14,3 L/h **- Volume de distribuição** (extensão da distribuição do fármaco pelos tecidos do organismo)**:** 918 L **- Semivida** (tempo necessário para que a concentração plasmática do fármaco se reduza para metade): ~44h	- Importante para **prevenir interações medicamentosas**, - **Ajustar exposição** do idoso em questão; - **Garantir segurança no uso prolongado**; - Definição da **posologia**.
Posologia recomendada	- **Posologia:** 80mg/dia em adultos	- **Ajustar exposição** do idoso em questão; - **Garantir segurança no uso prolongado**;
Segurança	- **Reações adversas (**diarreia, erupação cutânea, pele seca, toxicidade das unhas, estomatite, fadiga e diminuição do apetite)**;** **- Contraindicações (**hipersensibilidade)**;**	- Avaliação da **tolerabilidade ao fármaco num perfil polimedicado e idoso**, como é o caso.

Tabela 19 - Checklist para validação da Secção "Fundamentação | Estudos Relevantes"

Bloco Principal	*Pontos essenciais*	*Relevância*
Identificação do estudo principal	- **Estudo LAURA**: 1. Fase III (nível mais elevado de evidência clínica, antes da aprovação regulatória); 2. Randomizado; 3. Duplo-cego; 4. Placebo-controlado; 5. População-alvo: doentes com CPLNPC localmente avançado, irressecável, com mutações EGFR sensibilizantes (exão 19 del ou L858R), sem progressão após QT/RT baseada em platina.	- Garantia da **elevada qualidade científica da evidência utilizada, mostrando direta aplicabilidade ao caso clínico em questão** (doente com EGFR mutado, pós-QT/RT (Quimioterapia e Radioterapia).
Critérios de inclusão/exclusão	**Critérios de Inclusão:** **- Idade:** ≥18 anos. População incluída: mediana ~62 anos; 17% ≥70 anos *Paciente - 75 anos*; **- Estado funcional (ECOG Performance Status):** 0 ou 1. Distribuição no estudo: 64% ECOG 1, 36% ECOG 0; *Paciente - ECOG 1* **- Doença de base:** CPNPC localmente avançado, irressecável, estádio III (confirmado por avaliação histológica/citológica) *Paciente - estadio IIIB (T3N2M0) confirmado por biópsia* **- Mutações EGFR sensibilizantes:** Deleção exão 19 ou L858R no exão 21. Distribuição no estudo: 63% del19, 37% L858R. *Paciente - apresenta deleção do exão 19* **- Situação clínica após tratamento local:** Ausência de progressão após quimiorradioterapia (QT/RT) baseada em platina. Tipo de QT/RT: 69% concomitante, 31% sequencial. *Paciente - completou QT/RT baseada em platina, sem progressão* **Critérios de Exclusão (+ relevantes):** **- Histologia não confirma adenocarcinoma.** *Paciente - adenocarcinoma confirmado* - **Doenças pulmonares intersticiais (DPI) ou pneumonite** *Paciente - não apresenta DPI nem pneumonite*	- Verificação da **correspondência do doente em questão no perfil estudado**; - **Garantia de aplicabilidade.**

	- **Progressão da doença durante/antes da QT/RT (**necessidade de atingir estabilização**)** *Paciente - redução tumoral pós QT/RT, não teve, portanto, progressão.*	
Características da população	**Nº total de doentes**; **Distribuição por sexo/idade, ECOG, estadiamento, mutação específica, tipo de QT/RT.**	- Validação na **extrapolação dos resultados para o doente em questão.**
Resultados de eficácia	**Endpoint primário** **- Sobrevicência livre de progressão (PFS):** 31.9 meses com o fármaco vs 5,6 meses com placebo; - **HR (Hazard Rate):** 0.16 (IC95%; p < 0.001) → redução de 84% no risco de progressão ou morte; **Endpoints secundários:** **- Taxa de PFS aos 12 meses:** 74% (osimertinib) vs 22% (placebo). - **Taxa de PFS aos 24 meses:** 65% (osimertinib) vs 13% (placebo). - **Taxa de resposta objetiva (ORR):** 57% vs 33%. - **Duração mediana da resposta (DOR):** 36,9 meses vs 6,5 meses. **Benefício consistente em todos os subgrupos analisados (**idade, sexo, ECOG, tipo de mutação e tipo de QT/RT)	- Núcleo da decisão, na medida em que **quantifica o benefício clínico**; - Os **subgrupos em estudo dão confiança para que o fármaco seja aplicado dado o perfil do doente em questão.**
Resultados de segurança	**Eventos adversos (EA) de qualquer grau:** 98% dos doentes, no grupo osimertinib vs 88% placebo; **EA mais comuns:** **- Penumonite rádica** (48% com osimertinib vs. 38% com placebo); **- Diarreia** (36% com osimertinib vs. 14% com placebo); **- Rash** (24% com osimertinib vs. 14% com placebo). **Tolerabilidade considerada aceitável face ao ganho clínico obtido.**	- **Necessário no peso do risco/benefício, especialmente no contexto em questão** (idosa polimedicada).

Tabela 20 - *Checklist* para validação da Secção "Considerações Finais"

Bloco Principal	*Pontos essenciais*	*Relevância*
Caracterização resumida do caso clínico	- Menção do **sexo, idade, ECOG 1, adenocarcinoma pulmonar IIIB, mutação EGFT del19, PD-L1 1-50%, pós QT/RT, proposta de adjuvância com osimertinib**;	- **Resumo do perfil do doente**; - **Enquadramento do perfil do doente no perfil estudado** (critério-chave para extrapolação dos resultados do LAURA)
Situação regulatória (EMA e FDA)	- EMA não aprovou o osimertinib para adjuvância em CPNPC irressecável pós QT/RT; - FDA aprovou, em setembro de 2024, com base nos resultados do estudo LAURA;	- **Enquadramento regulatório no que toca as indicações terapêuticas do fármaco**, para perceber se o pedido é *off-label* na Europa (maior exigência de fundamentação) ou se já existe aprovação noutra agência de referência, como a FDA.
Evidência científica	**Eficácia:** **-** PFS: 39,1 vs 5,6 meses (HR 0,16); - OS: tendência favorável (dados imaturos, não significativos); - ORR: 57% vs 33%; - DOR: 36,9 vs 6,5 meses; - Progressão no SNC: 9% vs 36% aos 12 meses; **Segurança:** **-** EA qualquer grau: 98% vs 88%; - EA atribuídos ao tratamento ≥10%: paroníquia, diarreia, pele seca, rash, prurido. - EA ≥ grau 3: 35% vs 12%; principais: pneumonia, pneumonite rádica, diarreia, gastroenterite, ↑ CPK.	- Sustento da decisão, na medida em que **demonstram o benfício clínico substancial (PFS e SNC) e um perfil de toxicidade esperado mas manejável**; - Crucial para **avaliar o balanço risco-benefício do doente em questão**
Diretrizes internacionais	**European Society For Medical Oncology** (ESMO)**:** não inclui recomendação; **American Society of Clinical Oncology (ASCO) e National Comprehensive Cancer Network (NCCN):** recomendam osimertinib no contexto do doente (EGFR del19/L858R, pós QT/RT definitiva); **National Institute for Health and Care Excellence** (**NICE):** previsão de guideline para agosto de 2025.	- **Evidência científica já reconhecida por algumas sociedades** (ASCO, NCCN), mas ainda não por todas (ESMO, não, NICE em preparação); - **Importante indicador de uma tendência global para adoção.**

6. Conclusão e Trabalhos Futuros

O trabalho desenvolvido alcançou o seu objetivo central: desenvolver um PoC, recorrendo a *Large Language Models*, *Retrieval-Augmented Generation* (RAG), agentes, *Model Context Protocol* (MCP) e métodos adjacentes, com o intuito de automatizar, tanto quanto possível, a criação e a articulação de relatórios técnicos no domínio da FC, em resposta a pedidos realizados pela Comissão de Farmácia e Terapêutica (CFT). Este objetivo foi sempre norteado pela ambição de aproximar o sistema proposto ao *modus operandi* da UFC do Centro Hospitalar Universitário de São João, no Porto.

Primeiramente, procurou-se compreender, em profundidade, o processo, as fontes de informação utilizadas, o raciocínio e os critérios metodológicos utilizados pelos profissionais; de seguida, mapeou-se a estrutura dos pareceres, identificando componentes passíveis de automatização e selecionando técnicas e ferramentas existentes com potencial para cumprir, ainda que parcialmente, as exigências do processo.

No cerne do protótipo desenvolvido está o processo de recuperação de informação (RI), concebido para recuperar, de modo eficiente e contextual, informação relevante a partir de múltiplas fontes documentais e integrá-la, com minúcia, nos relatórios técnicos. Foi dedicada particular atenção ao processamento do documento "Resumo das Características do Medicamento (RCM)", documento regulamentar que sustenta a secção "Fundamentação: Farmacologia". Dado que a utilidade das diferentes secções do RCM varia com o pedido clínico, as características do doente e o fármaco em avaliação, a arquitetura proposta (*embeddings* + indexação + pesquisa semântica) pretende privilegiar apenas os excertos estritamente necessários à decisão. Considerando que diferentes modelos de *embedding* capturam relações semânticas de forma desigual, compararam-se várias alternativas em cenários realistas, suportados por *datasets* próprios: pares de *query–answer* e *gold standards,* por fármaco, todos eles validados manualmente, onde cada *chunk* recuperado foi rotulado como relevante ou não. A avaliação recorreu a métricas clássicas de IR (MAP, MRR, NDCG@3, *Recall@3* e *Precision@3*), tendo revelado, de forma consistente, a superioridade do text-embedding-ada-002 na ordenação (MAP = 0,832; MRR = 0,841; NDCG = 0,849), enquanto o gemini-embedding-001 se

destacou na cobertura (*Recall* = 0,955) com *precision* mais modesta (0,428). Modelos como voyage-3.5, medialbertina-pt-pt-1.5b e o *encoder* albertina-pt-pt-1.5b apresentaram desempenhos intermédios a inferiores. Dado o espaço de contexto limitado dos LLMs e a elevada extensão dos documentos, privilegiou-se a correta ordenação dos primeiros resultados (MAP, MRR e NDCG@3), fator determinante para reduzir ruído no *prompting* que alimenta o LLM. Assim, avançou-se com o text-embedding-ada-002 como *baseline* para as restantes etapas. Ainda se explorou o *finetuning* do *encoder* Albertina com os *datasets* construídos. Embora insuficientes para superar o desempenho do modelo de referência, os ganhos das experiências realizadas foram bastante promissores.

No eixo de pesquisa bibliográfica, operacionalizou-se a automatização da secção "Fundamentação | Estudos Relevantes" por via de um MCP biomédico (BioMCP) que, a partir de *queries* em linguagem natural, consulta o PubMed e outras fontes clínicas/médicas, devolvendo artigos e uma síntese da evidência. Entre as alternativas avaliadas, o BioMCP destacou-se pelo leque abrangente de *tools*, pela transparência do raciocínio e pela proximidade ao método crítico seguido por profissionais de farmacologia. Com base em três pedidos clínicos e *gold standards* de PMIDs validados, comparou-se o comportamento de diferentes LLMs na sua interação com o MCP. Os valores médios nas métricas de estabilidade mostraram o GPT-5 como o mais consistente (maior similaridade de Jaccard e RBO@5, menor volatilidade), contrastando com Claude Sonnet 4 e Gemini 2.5 Pro, que apresentaram maior diversidade de PMIDs, mas também maior instabilidade de ordenação e, até, episódios pontuais de erros (identificação do PMID errado e, no caso do Gemini, alucinação de um artigo). Não sendo as métricas de estabilidade suficientemente relevantes por si só foram também analisadas métricas de qualidade dos resultados recuperados. Também estas métricas permitiram convergir na conclusão de que o GPT-5 oferece melhor compromisso entre consistência e precisão. O Claude, eventualmente, poderá ser útil em fases exploratórias pela maior diversidade, enquanto o Gemini mostrou-se menos robusto para este caso de uso.

Todo o fluxo foi encapsulado num protótipo em Streamlit, organizado em cinco secções que espelham a estrutura do relatório final, tendo sido gerados 3 relatórios distintos, utilizando os 3 LLMs já mencionados como base. A avaliação dos relatórios gerados seguiu uma estratégia faseada: primeiramente, calcularam-se métricas quantitativas de texto (no caso, frequência e cobertura lexical de *keywords* clínicas), seguidamente, utilizou-se o método *LLM-as-a-judge,* com a comparação face a relatórios de referência redigidos por profissionais, e, por fim, *checklists* estruturadas para uma validação manual a ser levado a cabo pelos profissionais de farmacologia (trabalho futuro). Relativamente às métricas linguísticas, tanto no cálculo dos termos chave relevantes como na cobertura lexical global, registou-se uma variabilidade entre relatórios, ainda que com uma aparente supremacia do GPT-5 em mencionar as *keywords* com mais frequência. Ainda assim, a simples presença de *keywords* não garante qualidade semântica. Relativamente às avaliações provenientes do método *LLM-as-a-judge*, observou-se que o GPT-5 produziu, sistematicamente, um relatório com maior completude e melhor distribuição de termos pelas diversas categorias nucleares (fármacos, contexto, doença, biomarcadores, posologia, caso clínico, monitorização), com o Claude Sonnet 4 a apresentar

desempenho intermédio e o Gemini resultados mais inconsistentes. Contudo, são claras as limitações de ambos os métodos, reforçando-se a necessidade de avaliação clínica por parte de profissionais de farmacologia.

Por fim, a prova de conceito demonstra, de um modo prático, a viabilidade técnica e potencial: é possível automatizar partes substanciais do *pipeline* (recuperação, síntese, estruturação do parecer) com um bom desempenho e com relatórios que, numa análise preliminar, se aproximam do expectável.

Contudo, não se atingiu ainda a robustez necessária para uso autónomo num domínio tão exigente como a FC. Persistem desafios que requerem aprofundamento: mitigação de alucinações e ruído, relevância e completude nas secções redigidas, reprodutibilidade e auditoria do raciocínio, e, sobretudo, validação contínua por especialistas em cenários reais.

Trabalhos futuros poderão incidir em várias frentes:

1. Iterar processo de *finetuning* de modelos pt-pt: Prosseguir o *finetuning* de *encoders* portugueses e biomédicos com *corpus* curados (RCM, bulas, pareceres históricos, pareceres de sociedades médicas e científicas), recorrendo a aprendizagem contrastiva (semelhante ao realizado, com *hard negatives*);
2. Expandir o uso de agentes e MCP *Servers:* avançar para a automatização da secção "Pareceres de Sociedades Médicas e Científicas", através de agentes web capazes de aceder e estruturar informação proveniente de fontes variadas;
3. Qualidade, segurança e avaliação contínua: incorporar mecanismos de validação sistemática e um forte componente de *human-in-the-loop*, garantindo rigor científico, consistência e adaptabilidade às exigências clínicas em constante evolução. Desenvolver *checklists* para validação dos *outputs* gerados;
4. Melhorar a UI: Refinar a UI do protótipo, permitindo mais interação com o utilizador (ex.: ajuste de *queries* ao BioMCP, regeneração de secções com *inputs* adicionais, etc..);

Estas linhas de evolução, articuladas, permitem ambicionar um sistema cada vez mais fiável, auditável e útil no apoio à decisão da UFC. Espera-se que esta dissertação contribua para clarificar requisitos e definir padrões na aplicação de RAG, MCP e agentes na elaboração de relatórios técnicos de FC, disponibilizar *datasets* e métricas alinhados com tarefas reais e na demonstração, com resultados quantificados e *trade-offs* transparentes, e que a automatização é exequível, desde que acompanhada de validação humana e de um programa robusto de melhoria contínua. Espera-se, igualmente, que o futuro esteja, assim, delineado: transformar esta prova de conceito num sistema clínico de confiança, capaz de acelerar o trabalho da UFC, elevar a qualidade dos pareceres e, em última instância, contribuir para decisões terapêuticas mais informadas, consistentes e seguras.

Referências

Aamer Baig, Lareina Yee, Alex Singla, & Alexander Sukharevsky. (2024, abril 2). *What is ChatGPT, DALL-E, and generative AI? | McKinsey*. https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai

Abid All Awan. (2024, novembro 22). *What is Tokenization? Types, Use Cases, Implementation*. https://www.datacamp.com/blog/what-is-tokenization

Acosta-Enriquez, B. G., Ramos Farroñan, E. V., Villena Zapata, L. I., Mogollon Garcia, F. S., Rabanal-León, H. C., Angaspilco, J. E. M., & Bocanegra, J. C. S. (2024). Acceptance of artificial intelligence in university contexts: A conceptual analysis based on UTAUT2 theory. *Heliyon*, *10*(19), e38315. https://doi.org/10.1016/j.heliyon.2024.e38315

Adam Zewe. (2023, novembro 9). *Explained: Generative AI*. MIT News | Massachusetts Institute of Technology. https://news.mit.edu/2023/explained-generative-ai-1109

Alex Singla, Alexander Sukharevsky, Lareina Yee, Michael Chui, & Bryce Hall. (2025, março 12). *The State of AI: Global survey | McKinsey*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

Amanatullah. (2023, setembro 1). Transformer Architecture explained. *Medium*. https://medium.com/@amanatulla1606/transformer-architecture-explained-2c49e2257b4c

Anderson, M. (2021, fevereiro 25). *Famous Graphics Chips: Nvidia's GeForce 256*. IEEE Computer Society. https://www.computer.org/publications/tech-news/chasing-pixels/nvidias-geforce-256/

Anthropic. (2025, fevereiro 24). *Claude 3.7 Sonnet and Claude Code*. https://www.anthropic.com/news/claude-3-7-sonnet

B. J. Copeland. (2025, fevereiro 24). *DENDRAL | Artificial Intelligence, Machine Learning & Expert Systems | Britannica*. https://www.britannica.com/technology/DENDRAL

Bairaktaris, J. A., & Johannssen, A. (2025). Outsmarting algorithms: A comparative battle between Reinforcement Learning and heuristics in Atari Tetris. *Expert Systems with Applications*, *277*, 127251. https://doi.org/10.1016/j.eswa.2025.127251

Banjara, B. (2023, agosto 2). Fine-Tuning Large Language Models. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2023/08/fine-tuning-large-language-models/

Basheer, K. C. S. (2025, abril 9). What are LLM Benchmarks? *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2025/04/what-are-llm-benchmarks/

Beena Ammanath, Francisco Barroso, Sulabh Soral, Nitin Mittal, Costi Perricos, Deborshi Dutt, & Lynne Sterrett. (2023). *Generative AI use cases by type and industry*. Deloitte United States. https://www2.deloitte.com/us/en/pages/consulting/articles/gen-ai-use-cases.html

Beigel, J. H., Nam, H. H., Adams, P. L., Krafft, A., Ince, W. L., El-Kamary, S. S., & Sims, A. C. (2019). Advances in respiratory virus therapeutics – A meeting report from the 6th isirv Antiviral Group conference. *Antiviral Research*, *167*, 45–67. https://doi.org/10.1016/j.antiviral.2019.04.006

Biever, C. (2023). ChatGPT broke the Turing test—The race is on for new ways to assess AI. *Nature*, *619*(7971), 686–689. https://doi.org/10.1038/d41586-023-02361-7

Brilliant.org. (sem data). *Markov Chains | Brilliant Math & Science Wiki*. Obtido 4 de abril de 2025, de https://brilliant.org/wiki/markov-chains/

Broad Institute. (2010, setembro 13). *What is Mass Spectrometry?* @broadinstitute. https://www.broadinstitute.org/technology-areas/what-mass-spectrometry

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (No. arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Brunello, A., Fabris, G., Gasparetto, A., Montanari, A., Saccomanno, N., & Scalera, L. (2025). A survey on recent trends in robotics and artificial intelligence in the furniture industry. *Robotics and Computer-Integrated Manufacturing*, *93*, 102920. https://doi.org/10.1016/j.rcim.2024.102920

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4* (No. arXiv:2303.12712). arXiv. https://doi.org/10.48550/arXiv.2303.12712

Buchanan, B. G., & Feigenbaum, E. A. (1978). Dendral and meta-dendral: Their applications dimension. *Artificial Intelligence*, *11*(1), 5–24. https://doi.org/10.1016/0004-3702(78)90010-3

Caixeta, M. C. B. F., & Fabricio, M. M. (2018). Métodos e instrumentos de apoio ao codesign no processo de projeto de edifícios. *Ambiente Construído*, *18*(1), Artigo 1.

Campbell, M., Hoane, A. J., & Hsu, F. (2002). Deep Blue. *Artificial Intelligence*, *134*(1), 57–83. https://doi.org/10.1016/S0004-3702(01)00129-1

Casonatto, R. A., De Pádua Grillo Souza, T., & Mariano, A. M. (2024). Quality and Risk Management in Data Mining: A CRISP-DM Perspective. *Procedia Computer Science*, *242*, 161–168. https://doi.org/10.1016/j.procs.2024.08.257

Catarina Moreira. (2013). Neurónio. *Ciência Elementar*, *1*(1), 3.

Chandra, A. L. (2018, julho 24). McCulloch-Pitts Neuron—Mankind's First Mathematical Model Of A Biological Neuron. *TDS Archive*. https://medium.com/data-science/mcculloch-pitts-model-5fdf65ac5dd1

Chen, B., Wu, Z., & Zhao, R. (2023). From fiction to fact: The growing role of generative AI in business and finance. *Journal of Chinese Economic and Business Studies*, *21*(4), 471–496. https://doi.org/10.1080/14765284.2023.2245279

Chen, D., Liu, Y., Guo, Y., & Zhang, Y. (2024). The revolution of generative artificial intelligence in psychology: The interweaving of behavior, consciousness, and ethics. *Acta Psychologica*, *251*, 104593. https://doi.org/10.1016/j.actpsy.2024.104593

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., … Zaremba, W. (2021). *Evaluating Large Language Models Trained on Code* (No. arXiv:2107.03374). arXiv. https://doi.org/10.48550/arXiv.2107.03374

Choi, S.-Y., Arya, V., & Reynolds, K. (2021). Clinical Pharmacology-Informed Development of COVID-19 Therapeutics: Regulatory Experience. *Clinical Pharmacology and Therapeutics*, *109*(4), 810–812. https://doi.org/10.1002/cpt.2210

Chollet, F. (2019). *On the Measure of Intelligence* (No. arXiv:1911.01547). arXiv. https://doi.org/10.48550/arXiv.1911.01547

Cinlar, E. (2013). *Introduction to Stochastic Processes* (Ilustrado edição). Dover Publications.

Cole Stryker & Mark Scapicchio. (2024, março 22). *What is Generative AI? | IBM*. https://www.ibm.com/think/topics/generative-ai

Collie, R. J., & Martin, A. J. (2024). Teachers' motivation and engagement to harness generative AI for teaching and learning: The role of contextual, occupational, and background factors. *Computers and Education: Artificial Intelligence*, *6*, 100224. https://doi.org/10.1016/j.caeai.2024.100224

Cost, B. (2025, abril 4). *Terrifying study reveals AI robots have passed 'Turing test'—And are now indistinguishable from humans, scientists say*. https://nypost.com/2025/04/04/tech/terrifying-study-reveals-ai-robots-have-passed-turing-test-and-are-now-indistinguishable-from-humans-scientists-say/

Currie, G. M., Hawk, K. E., & Rohren, E. M. (2024). Generative Artificial Intelligence Biases, Limitations and Risks in Nuclear Medicine: An Argument for Appropriate Use Framework and Recommendations. *Seminars in Nuclear Medicine*. https://doi.org/10.1053/j.semnuclmed.2024.05.005

Databricks. (2023, outubro 18). *Retrieval Augmented Generation*. Databricks. https://www.databricks.com/glossary/retrieval-augmented-generation-rag

Dave Bergmann & Cole Stryker. (2024, novembro 25). *What is Backpropagation? | IBM*. https://www.ibm.com/think/topics/backpropagation

Edd Gent. (2024, julho 27). *12 game-changing moments in the history of artificial intelligence (AI)*. Livescience.Com. https://www.livescience.com/technology/artificial-intelligence/12-game-changing-moments-in-the-history-of-ai

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., & Larson, J. (2025). *From Local to Global: A Graph RAG Approach to Query-Focused Summarization* (No. arXiv:2404.16130). arXiv. https://doi.org/10.48550/arXiv.2404.16130

Evidently AI. (2025, fevereiro 20). *20 LLM evaluation benchmarks and how they work*. https://www.evidentlyai.com/llm-guide/llm-benchmarks

Falconer, N., Abdel-Hafez, A., Scott, I. A., Marxen, S., Canaris, S., & Barras, M. (2021). Systematic review of machine learning models for personalised dosing of heparin. *British Journal of Clinical Pharmacology*, *87*(11), 4124–4139. https://doi.org/10.1111/bcp.14852

Fanjia Yan, Huanzhi Mao, Charlie Cheng-Jie Ji, Ion Stoica, Joseph E. Gonzalez, Tianjun Zhang, & Shishir G. Patil. (2024, agosto 19). *Berkeley Function Calling Leaderboard V3 (aka Berkeley Tool Calling Leaderboard V3)*. https://gorilla.cs.berkeley.edu/leaderboard.html

Feigenbaum, E. A. (1969). *HEURISTIC DENDRAL: A Program for Generating Explanatory Hypotheses in Organic Chemistry*. https://purl.stanford.edu/yj802bw3458

Foster, D., & Friston, K. (2023). *Generative deep learning: Teaching machines to paint, write, compose and play* (2nd ed). O'Reilly.

Foung, D., Lin, L., & Chen, J. (2024). Reinventing assessments with ChatGPT and other online tools: Opportunities for GenAI-empowered assessment practices. *Computers and Education: Artificial Intelligence*, *6*, 100250. https://doi.org/10.1016/j.caeai.2024.100250

García-Peñalvo, F., & Vázquez-Ingelmo, A. (2023). What Do We Mean by GenAI? A Systematic Mapping of The Evolution, Trends, and Techniques Involved in Generative AI. *International Journal of Interactive Multimedia and Artificial Intelligence*, *8*(4), 7. https://doi.org/10.9781/ijimai.2023.07.006

Goda, Y., Yamada, M., Matsukawa, H., Hata, K., & Yasunami, S. (2014). Conversation with a Chatbot before an Online EFL Group Discussion and the Effects on Critical Thinking. *The Journal of Information and Systems in Education*, *13*(1), 1–7. https://doi.org/10.12937/ejsise.13.1

Gulmez, S. E., Aydin, V., & Akici, A. (2020). Footprints of Clinical Pharmacology in Turkey: Past, Present, and Future. *Clinical Therapeutics*, *42*(2), 351–362. https://doi.org/10.1016/j.clinthera.2019.12.014

H20.ai. (2022). *What is Self-attention?* https://h2o.ai/wiki/self-attention/

Hayden Wolff. (2023, dezembro 18). *RAG 101: Demystifying Retrieval-Augmented Generation Pipelines*. NVIDIA Technical Blog. https://developer.nvidia.com/blog/rag-101-demystifying-retrieval-augmented-generation-pipelines/

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). *Measuring Massive Multitask Language Understanding* (No. arXiv:2009.03300). arXiv. https://doi.org/10.48550/arXiv.2009.03300

Hervé Jegou, Matthijs Douze, & Jeff Johnson. (2017, março 29). Faiss: A library for efficient similarity search. *Engineering at Meta*. https://engineering.fb.com/2017/03/29/data-infrastructure/faiss-a-library-for-efficient-similarity-search/

Hotz, N. (2024, dezembro 9). *What is CRISP DM?* Data Science PM. https://www.datascience-pm.com/crisp-dm-2/

IBM. (2021a, agosto 17). *SPSS Modeler*. https://www.ibm.com/docs/pt-br/spss-modeler/saas?topic=dm-crisp-help-overview

IBM. (2021b, outubro 6). *O que são redes neurais convolucionais? | IBM*. https://www.ibm.com/br-pt/think/topics/convolutional-neural-networks

IBM Research. (2021a, fevereiro 9). *Foundation Models*. IBM Research. https://research.ibm.com/topics/foundation-models

IBM Research. (2021b, fevereiro 9). *What is retrieval-augmented generation (RAG)?* IBM Research. https://research.ibm.com/blog/retrieval-augmented-generation-RAG

IBM Research. (2023, novembro 2). *What Are Large Language Models (LLMs)? | IBM*. https://www.ibm.com/think/topics/large-language-models

Ismail Eruysaliz. (2024, setembro 15). *Top 7 Challenges with Retrieval-Augmented Generation*. https://www.valprovia.com/en/blog/top-7-challenges-with-retrieval-augmented-generation

Javier Canales Luna. (2024, abril 23). *O que são modelos de fundação?* https://www.datacamp.com/blog/what-are-foundation-models

Jeremy Norman. (2021). *McCulloch & Pitts Publish the First Mathematical Model of a Neural Network: History of Information*. HistoryofInformation.com. https://www.historyofinformation.com/detail.php?entryid=782

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, *55*(12), 248:1-248:38. https://doi.org/10.1145/3571730

Jiang, J. (2024). When generative artificial intelligence meets multimodal composition: Rethinking the composition process through an AI-assisted design project. *Computers and Composition*, *74*, 102883. https://doi.org/10.1016/j.compcom.2024.102883

Jim Holdsworth & Matthew Kosinski. (2024, julho 29). *What Is A Vector Database? | IBM*. https://www.ibm.com/think/topics/vector-database

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019). *PubMedQA: A Dataset for Biomedical Research Question Answering* (No. arXiv:1909.06146). arXiv. https://doi.org/10.48550/arXiv.1909.06146

Joanna Goodrich. (2021, janeiro 25). *How IBM's Deep Blue Beat World Champion Chess Player Garry Kasparov—IEEE Spectrum*. https://spectrum.ieee.org/how-ibms-deep-blue-beat-world-champion-chess-player-garry-kasparov

John Nosta. (2025, abril 2). *AI Beat the Turing Test by Being a Better Human | Psychology Today*. https://www.psychologytoday.com/us/blog/the-digital-self/202504/ai-beat-the-turing-test-by-being-a-better-human

Johnson, M., Patel, M., Phipps, A., van der Schaar, M., Boulton, D., & Gibbs, M. (2023). The potential and pitfalls of artificial intelligence in clinical pharmacology. *CPT: Pharmacometrics & Systems Pharmacology*, *12*(3), 279–284. https://doi.org/10.1002/psp4.12902

Jovanović, M., & Campbell, M. (2022). Generative Artificial Intelligence: Trends and Prospects. *Computer*, *55*(10), 107–112. https://doi.org/10.1109/MC.2022.3192720

Kalita, D. (2022, março 11). What is Recurrent Neural Networks (RNN)? *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2022/03/a-brief-overview-of-recurrent-neural-networks-rnn/

Kanakala, G. C., Devata, S., Chatterjee, P., & Priyakumar, U. D. (2024). Generative artificial intelligence for small molecule drug design. *Current Opinion in Biotechnology*, *89*, 103175. https://doi.org/10.1016/j.copbio.2024.103175

Kaneko, M., Bollegala, D., Okazaki, N., & Baldwin, T. (2024). *Evaluating Gender Bias in Large Language Models via Chain-of-Thought Prompting* (No. arXiv:2401.15585). arXiv. https://doi.org/10.48550/arXiv.2401.15585

Kevin Gibbs. (2004, dezembro 10). I've got a suggestion. *Official Google Blog*. https://googleblog.blogspot.com/2004/12/ive-got-suggestion.html

Kim, S. (2023, setembro 30). List of Open Sourced Fine-Tuned Large Language Models (LLM). *Medium*. https://sungkim11.medium.com/list-of-open-sourced-fine-tuned-large-language-models-llm-8d95a2e0dc76

Kingma, D. P., & Welling, M. (2022). *Auto-Encoding Variational Bayes* (No. arXiv:1312.6114). arXiv. https://doi.org/10.48550/arXiv.1312.6114

Korf, R. E. (1990). Real-time heuristic search. *Artificial Intelligence*, *42*(2), 189–211. https://doi.org/10.1016/0004-3702(90)90054-4

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, *25*. https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* (No. arXiv:2005.11401). arXiv. https://doi.org/10.48550/arXiv.2005.11401

Lin, S., Hilton, J., & Evans, O. (2022). *TruthfulQA: Measuring How Models Mimic Human Falsehoods* (No. arXiv:2109.07958). arXiv. https://doi.org/10.48550/arXiv.2109.07958

Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A., & Lederberg, J. (1993). DENDRAL: A case study of the first expert system for scientific hypothesis formation. *Artificial Intelligence*, *61*(2), 209–261. https://doi.org/10.1016/0004-3702(93)90068-M

Liz Gannes. (2013, agosto 23). Nearly a Decade Later, the Autocomplete Origin Story: Kevin Gibbs and Google Suggest. *AllThingsD*. https://allthingsd.com/20130823/nearly-a-decade-later-the-autocomplete-origin-story-kevin-gibbs-and-google-suggest/

Löfström, T. (2009). *Utilizing Diversity and Performance Measures for Ensemble Creation*. https://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-2920

Luiza Pereira. (2024, junho). *O que é o Algoritmo Perceptron?* Asimov Academy. https://hub.asimov.academy/tutorial/o-que-e-o-algoritmo-perceptron/

Luo, Z., Shi, X., Lin, X., & Gao, J. (2025). *Evaluation Report on MCP Servers* (No. arXiv:2504.11094). arXiv. https://doi.org/10.48550/arXiv.2504.11094

Lyakhova, U. A., & Lyakhov, P. A. (2024). Systematic review of approaches to detection and classification of skin cancer using artificial intelligence: Development and prospects. *Computers in Biology and Medicine*, *178*, 108742. https://doi.org/10.1016/j.compbiomed.2024.108742

Maël Fabien. (2018, novembro 20). *The Rosenblatt's Perceptron*. https://maelfabien.github.io/deeplearning/Perceptron/

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, *27*(4), Artigo 4. https://doi.org/10.1609/aimag.v27i4.1904

McCulloch, W. S., & Pitts, W. (1990). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, *52*(1), 99–115. https://doi.org/10.1016/S0092-8240(05)80006-0

Menzli, A. (2022, julho 21). *Tokenization in NLP: Types, Challenges, Examples, Tools*. Neptune.Ai. https://neptune.ai/blog/tokenization-in-nlp

Merry, C., & Flexner, C. W. (2008). CHAPTER 16—Pharmacology of Antiretroviral Drugs. Em P. A. Volberding, M. A. Sande, W. C. Greene, J. M. A. Lange, J. E. Gallant, & C. C. Walsh (Eds.), *Global HIV/AIDS Medicine* (pp. 171–179). W.B. Saunders. https://doi.org/10.1016/B978-1-4160-2882-6.50020-4

Meta AI. (2025, abril 5). *The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation*. Meta AI. https://ai.meta.com/blog/llama-4-multimodal-intelligence/

Michael Chui, Eric Hazan, Roger Roberts, Alex Singla, Kate Smaje, Alex Sukharevsky, Lareina Yee, & Rodney Zemmel. (2023, junho 14). *Economic potential of generative AI | McKinsey*. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-AI-the-next-productivity-frontier#work-and-productivity

Minsky, M., & Papert, S. (1969). *(1969) Marvin Minsky and Seymour Papert, Perceptrons, Cambridge, MA: MIT Press, Introduction, pp. 1-20, and p. 73 (figure 5.1)*. https://doi.org/10.7551/mitpress/4943.003.0015

Mojan Javaheripi. (2023, dezembro 12). *Phi-2: The surprising power of small language models—Microsoft Research*. https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/

Nentidis, A., Katsimpras, G., Krithara, A., López, S. L., Farré-Maduell, E., Gasco, L., Krallinger, M., & Paliouras, G. (2023). *Overview of BioASQ 2023: The eleventh BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering* (Vol. 14163, pp. 227–250). https://doi.org/10.1007/978-3-031-42448-9_19

Nielsen, M. A. (2015). *Neural Networks and Deep Learning*. http://neuralnetworksanddeeplearning.com

NVIDIA. (2024). *What is Generative AI?* NVIDIA. https://www.nvidia.com/en-us/glossary/generative-ai/

Ojha, V., & Nicosia, G. (2022). Backpropagation Neural Tree. *Neural Networks*, *149*, 66–83. https://doi.org/10.1016/j.neunet.2022.02.003

Open AI Platform. (2025). *Vector embeddings—OpenAI API*. https://platform.openai.com

OpenAI. (2022, novembro 30). *Introducing ChatGPT*. https://openai.com/index/chatgpt/

OpenAI. (2025, fevereiro 27). *Introducing GPT-4.5*. https://openai.com/index/introducing-gpt-4-5/

OpenAI Research. (2023). https://openai.com/research/index/

Oppy, G., & Dowe, D. (2021). The Turing Test. Em E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entriesuring-test/

Pallottino, F., Violino, S., Figorilli, S., Pane, C., Aguzzi, J., Colle, G., Nerio Nemmi, E., Montaghi, A., Chatzievangelou, D., Antonucci, F., Moscovini, L., Mei, A., Costa, C., & Ortenzi, L. (2025). Applications and perspectives of Generative Artificial Intelligence in agriculture. *Computers and Electronics in Agriculture*, *230*, 109919. https://doi.org/10.1016/j.compag.2025.109919

Ramos, S. (2011). Clara Pereira Coutinho. 2011. Metodologia da Investigação em Ciências Sociais e Humanas: Teoria e Prática. *Interações: Sociedade e as novas modernidades*, *20*, Artigo 20. https://www.interacoes-ismt.com/index.php/revista/article/view/285

Raschka, S. (2023, fevereiro 9). *Understanding and Coding the Self-Attention Mechanism of Large Language Models From Scratch*. Sebastian Raschka, PhD. https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html

Rasul, T., Nair, S., Kalendra, D., Balaji, M. S., Santini, F. de O., Ladeira, W. J., Rather, R. A., Yasin, N., Rodriguez, R. V., Kokkalis, P., Murad, M. W., & Hossain, M. U. (2024). Enhancing academic integrity among students in GenAI Era:A holistic framework. *The International Journal of Management Education*, *22*(3), 101041. https://doi.org/10.1016/j.ijme.2024.101041

Rawlins, M. D. (2013). Clinical pharmacology in health care, teaching and research. *British Journal of Clinical Pharmacology*, *75*(5), 1219–1220. https://doi.org/10.1111/bcp.12074

Rina Caballar & Cole Stryker. (2024, junho 25). *What Are LLM Benchmarks? | IBM*. https://www.ibm.com/think/topics/llm-benchmarks

Rockwell Anyoha. (2017, agosto 28). *The History of Artificial Intelligence – Science in the News*. https://sites.harvard.edu/sitn/2017/08/28/history-artificial-intelligence/

Rodler, S., Ganjavi, C., De Backer, P., Magoulianitis, V., Ramacciotti, L. S., De Castro Abreu, A. L., Gill, I. S., & Cacciamani, G. E. (2024). Generative artificial intelligence in surgery. *Surgery*, *175*(6), 1496–1502. https://doi.org/10.1016/j.surg.2024.02.019

Rolls, E. T. (2024). The memory systems of the human brain and generative artificial intelligence. *Heliyon*, *10*(11), e31965. https://doi.org/10.1016/j.heliyon.2024.e31965

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Saeed, M. (2022, setembro 19). A Gentle Introduction to Positional Encoding in Transformer Models, Part 1. *MachineLearningMastery.Com*. https://www.machinelearningmastery.com/a-gentle-introduction-to-positional-encoding-in-transformer-models-part-1/

Salah, A., & Hayette, G. (2025). A meta-analysis of supervised and unsupervised machine learning algorithms and their application to active portfolio management. *Expert Systems with Applications*, *271*, 126611. https://doi.org/10.1016/j.eswa.2025.126611

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (Eds.). (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Vol. 11700). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6

SAS Insights. (sem data). *Redes Neurais: O que são? Porque são importantes?* Obtido 10 de abril de 2025, de https://www.sas.com/pt_pt/insights/analytics/neural-networks.html

Schmidt, B., & Hildebrandt, A. (2024). From GPUs to AI and quantum: Three waves of acceleration in bioinformatics. *Drug Discovery Today*, *29*(6), 103990. https://doi.org/10.1016/j.drudis.2024.103990

Solaiman, B. (2024). Generative artificial intelligence (GenAI) and decision-making: Legal & ethical hurdles for implementation in mental health. *International Journal of Law and Psychiatry*, *97*, 102028. https://doi.org/10.1016/j.ijlp.2024.102028

Stockemer, D. (2019). *Quantitative Methods for the Social Sciences: A Practical Introduction with Examples in SPSS and Stata*. Springer International Publishing. https://doi.org/10.1007/978-3-319-99118-4

Stöffelbauer, A. (2023, outubro 24). How Large Language Models Work. *Data Science at Microsoft*. https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f

Sunil Ramlochan. (2023, dezembro 12). *Openness in Language Models: Open Source vs Open Weights vs Restricted Weights*. Prompt Engineering Institute. https://promptengineering.org/llm-open-source-vs-open-weights-vs-restricted-weights/

Suvojit. (2023, março 13). What are Large Language Models(LLMs)? *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2023/03/an-introduction-to-large-language-models-llms/

Tan, X., Cheng, G., & Ling, M. H. (2025). Artificial intelligence in teaching and teacher professional development: A systematic review. *Computers and Education: Artificial Intelligence*, *8*, 100355. https://doi.org/10.1016/j.caeai.2024.100355

Teo, Z. L., Quek, C. W. N., Wong, J. L. Y., & Ting, D. S. W. (2024). Cybersecurity in the generative artificial intelligence era. *Asia-Pacific Journal of Ophthalmology*, *13*(4), 100091. https://doi.org/10.1016/j.apjo.2024.100091

The ELIZA Effect. (2019, dezembro 11). *99% Invisible*. https://99percentinvisible.org/episode/the-eliza-effect/

Tripathi, G. P. (2023, maio 16). How to Evaluate a Large Language Model (LLM)? *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2023/05/how-to-evaluate-a-large-language-model-llm/

TURING, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Ulla, M. B., Advincula, M. J. C., Mombay, C. D. S., Mercullo, H. M. A., Nacionales, J. P., & Entino-Señorita, A. D. (2024). How can GenAI foster an inclusive language classroom? A critical language pedagogy perspective from Philippine university teachers. *Computers and Education: Artificial Intelligence*, *7*, 100314. https://doi.org/10.1016/j.caeai.2024.100314

Vanna. (2025, fevereiro 18). *What is self-attention? | IBM*. https://www.ibm.com/think/topics/self-attention

Vashisth, V. (2025, abril 7). Decoding LLMs: When to Use Prompting, Fine-tuning, AI Agents, and RAG Systems. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2025/04/llm-approach/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (No. arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

vellum. (2024, setembro 11). *LLM Benchmarks: Overview, Limits and Model Comparison*. https://www.vellum.ai/blog/llm-benchmarks-overview-limits-and-model-comparison

Wang, H., Dang, A., Wu, Z., & Mac, S. (2024). Generative AI in higher education: Seeing ChatGPT through universities' policies, resources, and guidelines. *Computers and Education: Artificial Intelligence*, *7*, 100326. https://doi.org/10.1016/j.caeai.2024.100326

Wang, J., Yi, X., Guo, R., Jin, H., Xu, P., Li, S., Wang, X., Guo, X., Li, C., Xu, X., Yu, K., Yuan, Y., Zou, Y., Long, J., Cai, Y., Li, Z., Zhang, Z., Mo, Y., Gu, J., … Xie, C. (2021). Milvus: A Purpose-Built Vector Data Management System. *Proceedings of the 2021 International Conference on Management of Data*, 2614–2627. https://doi.org/10.1145/3448016.3457550

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (No. arXiv:2201.11903). arXiv. https://doi.org/10.48550/arXiv.2201.11903

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). *Ethical and social risks of harm from Language Models* (No. arXiv:2112.04359). arXiv. https://doi.org/10.48550/arXiv.2112.04359

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Willsmore, M. (2024, janeiro 17). *5 Challenges Implementing Retrieval Augmented Generation (RAG)*. Pureinsights. https://pureinsights.com/blog/2024/five-common-challenges-when-implementing-rag-retrieval-augmented-generation/

Wu, Y. (2023, junho 7). *Why You Shouldn't Invest In Vector Databases?* Medium. https://blog.det.life/why-you-shouldnt-invest-in-vector-databases-c0cd3f59d23c

Yang, C. (2025, março 25). *DeepSeek v3—Advanced AI & LLM Model Online*. DeepSeek V3. https://deepseekv3.org/

Yao, Y., Li, Z., & Zhao, H. (2024). *Beyond Chain-of-Thought, Effective Graph-of-Thought Reasoning in Language Models* (No. arXiv:2305.16582). arXiv. https://doi.org/10.48550/arXiv.2305.16582

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena* (No. arXiv:2306.05685). arXiv. https://doi.org/10.48550/arXiv.2306.05685

Anexos

Anexo A. Parecer exemplo redigido por profissionais da UFC (resumido)

Parecer nº 06/2022: Guselcumab na psoríase em placas grave refratária

ENQUADRAMENTO

A Comissão de Farmácia e Terapêutica (CFT) do Centro Hospitalar Universitário de S. João (CHUSJ) vem por este meio solicitar à Unidade de Farmacologia Clínica (UFC) parecer relativo à utilização do biotecnológico guselcumab numa doente com diagnóstico de psoríase em placas, grave e refratária a tratamentos prévios.

CASO CLÍNICO

Paciente: XXX

Antecedentes patológicos: arritmia e hipertensão arterial pós-ciclosporina (não medicada).
Medicação habitual: contracetivo oral.
História da doença atual: Doente com psoríase em placas grave, com atingimento de couro cabeludo, tronco, membros e região genital (PASI 20), refratária a vários tratamentos tópicos prévios, com intolerância a ciclosporina e indisponibilidade para fototerapia. Para além disso, apresenta agravamento recente da doença, com impacto significativo na qualidade de vida (DLQI >10) e implicações a nível psicossocial e estético associadas. Neste sentido, proposta para terapêutica biológica com guselcumab.

FUNDAMENTAÇÃO

FARMACOLOGIA

Nome Comercial e Titular da AIM: Tremfya®, Janssen-Cilag International NV, Bélgica
Indicações on-label:
O guselcumab encontra-se aprovado pela European Medicines Agency (EMA) no tratamento de adultos com:

- psoríase em placas, moderada a grave, que são candidatos a terapêutica sistémica;

- artrite psoriática ativa, em monoterapia ou em combinação com metotrexato (MTX), em doentes que tiveram uma reposta inadequada ou que demonstraram ser intolerantes a uma terapêutica prévia com medicamentos anti-reumáticos modificadores da doença (DMARD).(1)

O guselcumab encontra-se aprovado pela U.S. Food and Drug Administration (FDA) no tratamento de adultos com:

- psoríase em placas, moderada a grave, que são candidatos a terapêutica sistémica ou fototerapia;
- artrite psoriática ativa.(2)

Propriedades Farmacodinâmicas: O guselcumab é um anticorpo monoclonal de IgG1λ humano que se liga de forma seletiva à proteína IL-23 com alta especificidade e afinidade. A IL-23, uma citocina reguladora, afeta a diferenciação, expansão e sobrevivência de subconjuntos de células T e subconjuntos de células imunes inatas que provocam doenças inflamatórias. Os níveis de IL-23 presentes na pele dos doentes com psoríase em placas encontram-se elevados, pelo que o guselcumab exerce efeitos clínicos na psoríase em placas e na artrite psoriática através do bloqueio da via da citocina IL-23. Nos humanos, o bloqueio seletivo da IL-23 mostrou normalizar a produção de citocinas mediadas pela IL-23 e, nos modelos in vitro, demonstrou inibir a bioatividade da IL-23, ao bloquear a sua interação com o recetor de superfície celular da IL-23, interrompendo a sinalização, a ativação e as cascatas de citocinas.(1)

Perfil Farmacocinético: Absorção: Após uma injeção única subcutânea de 100 mg em indivíduos saudáveis, o guselcumab atingiu uma concentração sérica máxima (Cmax) média (± DP) de 8,09 ± 3,68 µg/ml em, aproximadamente, 5,5 dias após a administração, e apresentou uma biodisponibilidade absoluta estimada em, aproximadamente, 49%. Distribuição: O volume de distribuição médio após uma administração única intravenosa em indivíduos saudáveis variou entre, aproximadamente, 7 e 10 L. Biotransformação: A via exata através da qual o guselcumab é metabolizado não se encontra caracterizada. Sendo um mAc IgG humano, espera-se que o guselcumab seja degradado em pequenos péptidos e aminoácidos através de vias catabólicas de forma semelhante à da IgG endógena. Eliminação: Após uma administração única intravenosa em indivíduos saudáveis, a depuração sistémica (Cl) média variou entre 0,288 e 0,479 L/dia e o tempo de semivida (t½) médio do guselcumab foi de, aproximadamente, 17 dias. As análises farmacocinéticas populacionais indicaram que o uso concomitante de AINEs, corticosteróides orais e DMARDcs como o metotrexato, não afetou a depuração de guselcumab.(1)

Considerações posológicas: Na psoríase em placas, a dose recomendada é de 100 mg por injeção subcutânea nas semanas 0 e 4, seguida de uma dose de manutenção a cada 8 semanas. A descontinuação do tratamento deve ser considerada em doentes que não apresentem qualquer resposta após 16 semanas de tratamento. (1)

Reações adversas: As mais frequentes foram as infeções do trato respiratório. Com menor frequência foram também relatadas cefaleia, diarreia, artralgia, reações no local de injeção e transaminases aumentadas. Foram relatadas com baixa frequência infeções por herpes simplex, tinha, gastroenterite, hipersensibilidade, anafilaxia, urticária, erupção cutânea e neutropenia.(1)

Contraindicações: Hipersensibilidade grave à substância ativa ou a qualquer um dos excipientes.(1)

ESTUDOS RELEVANTES

Guselcumab vs. Adalimumab

Os estudos de eficácia e segurança mais relevantes no que concerne à comparação entre as terapêuticas com guselcumab vs. adalimumab e que tiveram impacto na aprovação do guselcumab no tratamento da psoríase em placas foram:

- VOYAGE 1 e VOYAGE 2 (2016)(3, 4) – Nível de Evidência 2 de Oxford: Estudos de fase III, randomizados, com dupla ocultação, que avaliaram a eficácia e segurança do guselcumab vs. placebo ou adalimumab em 1829 doentes, ao longo do período de 48 semanas. No estudo VOYAGE 2, após a semana 28, os doentes sob guselcumab ou adalimumab com melhoria ≥90% na pontuação do PASI [PASI90] foram novamente randomizados para guselcumab ou placebo; todos os doentes que não responderam/ tiveram perda de resposta, receberam guselcumab.

    - Os outcomes primários avaliados foram: Avaliação Global do Investigador [IGA]; Índice de Gravidade e Área da Psoríase [PASI]; Índice de Qualidade de Vida em Dermatologia; Diário de Sinais e Sintomas de Psoríase; e segurança.
    - VOYAGE 1: O guselcumab foi superior ao ($p<0,001$) ao placebo na semana 16 relativamente à pontuação IGA (85,1% vs. 6,9%) e à melhoria de 90% na pontuação PASI [PASI90] (73,3% vs. 2,9%]. O guselcumab também foi superior ($p<0,001$) ao adalimumab para IGA (85,1% vs. 65,9%) e PASI90 (73,3% vs. 49,7%) na semana 16, na semana 24 (84,2% vs. 61,7% e 80,2% vs.53,0%) e na semana 48 (80,5% vs. 55,4% e 76,3% vs. 47,9%). Para além das melhorias significativas nas medidas da atividade da doença, o guselcumab também melhorou significativamente os resultados relatados pelos doentes até a semana 48. As taxas de eventos adversos foram comparáveis entre os tratamentos.
    - VOYAGE 2: O guselcumab foi superior ao placebo relativamente a IGA e PASI90 na semana 16 e superior ao adalimumab para IGA, PASI75 e PASI90 na semana

16 e para IGA, PASI90 e PASI100 na semana 24 (p<0,001). Das semanas 28 a 48, foi observada melhor taxa de persistência de resposta ao tratamento no grupo de manutenção em relação ao grupo retirado do guselcumab (p<0,001). Dos doentes que não responderam ao adalimumab e mudaram para guselcumab, 66,1% alcançaram PASI 90 na semana 48. O guselcumab também melhorou os resultados relatados pelos doentes. Os eventos adversos foram comparáveis entre os grupos.

- A eficácia e a segurança do guselcumab foram demonstradas independentemente da idade, género, raça, peso corporal, localização das placas, intensidade basal do PASI, artrite psoriática concomitante e tratamento prévio com uma terapêutica biológica.(5-8)

Outros estudos que comparam as terapêuticas em questão:

- Yang et al. (2020)(9) – Nível de Evidência 1 de Oxford: Meta-análise baseada em ensaios controlados e randomizados que investigou a eficácia e segurança de guselcumab no tratamento da psoríase em placas moderada a grave. Foram incluídos 7 artigos, com um total de 1206 doentes tratados com guselcumab, 585 com placebo e 1250 com adalimumab.
  - Os outcomes primários avaliados foram as pontuações de PASI, IGA e DLQI e incidência de eventos adversos (AEs) e eventos adversos graves (SAEs).
  - O guselcumab teve melhor eficácia comparativamente com placebo ou adalimumab conforme demonstrado pelo PASI75 (OR 61,37; IC 95%: 31,15 a 120,91 e OR 3,08; IC 95%: 2,35 a 4,06, respetivamente para placebo e adalimumab), pela pontuação da IGA (OR 65,75; IC 95%; 45,54 a 94,95 e OR 2,79; IC 95%; 2,17 a 3,59) e pela pontuação do DLQI (OR 29,64; IC95%: 18,80 a 46,73 e OR 1,86; IC 95%; 1,50 a 2,31).
  - O guselcumab teve segurança semelhante com placebo ou adalimumab sobre a incidência de EA (OR 1,05; IC 95%: 0,86 a 1,29 e OR 0,97; IC 95%: 0,79 a 1,19) e SAEs (OR 1,03; IC 95%: 0,47 a 2,27 e OR 0,91; IC 95%: 0,44 a 1,87).
  - A meta-análise apresentou uma heterogeneidade moderada (0% < I2 < 57%).

(…)

PARECERES DE SOCIEDADES MÉDICAS E CIENTÍFICAS

De acordo com as mais recentes diretrizes relativas à abordagem e ao tratamento da psoríase com agentes biológicos, elaboradas em conjunto pela American Academy of Dermatology (AAD) e pela National Psoriasis Foundation, e publicadas em 2020, os agentes biológicos, em

monoterapia ou em combinação com outros medicamentos tópicos ou sistémicos, têm uma alta relação risco-benefício, sendo, por isso, indicados para o tratamento da psoríase moderada a grave, em doentes incapazes de controlar adequadamente a doença apenas com medicamentos tópicos e/ou fototerapia.(13) Não são referidas linhas de orientação terapêutica que distingam quais agentes são preferidos como 1ª linha de terapêutica biológica ou linhas subsequentes.

PARECER DO NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE (NICE)

De acordo com as diretrizes da NICE subordinadas ao tema "Psoriasis: assessment and management", atualizadas em 2017, é previsto o uso de terapêutica biológica em doentes adultos com psoríase severa, sendo enunciadas como alternativas possíveis adalimumab, bimequizumab, brodalumab, certolizumab pegol, etanercept, guselcumab, ixekisumab, risanquizumab, secucinumab, tildraquizumab e ustecinumab, na mesma linha terapêutica.(14)

PARECER DO INFARMED

O guselcumab foi submetido a uma avaliação de financiamento público, da qual resultou, em junho de 2020, o deferimento da sua utilização no tratamento da psoríase em placas, moderada a grave, em adultos que são candidatos a terapêutica sistémica. Relativamente à avaliação farmacoterapêutica, o guselcumab apresentou valor terapêutico acrescentado não quantificável, face ao comparador, o adalimumab, na população com psoríase em placas moderada a grave, elegível para terapêutica biológica, e em que o tratamento sistémico não biológico não é eficaz, não é tolerado, ou está contraindicado. Quanto à avaliação económica, o custo da terapêutica com guselcumab foi inferior ao custo da terapêutica com adalimumab.(15)

(…)

CONSIDERAÇÕES FINAIS

Tendo em consideração a doente em questão, e de acordo com o protocolo institucional existente, verificamos que estamos perante uma doente com indicação formal para o uso de terapêutica botecnológica para o tratamento da sua condição clínica, podendo o mesmo passar pelo uso de fármacos anti-IL23 (rizanquizumab, guselcumab ou tildraquizumab), ou, apesar da doente apresentar uma idade correspondente ao limite mínimo para a sua indicação, pelo uso de adalimumab.

No que concerne à evidência científica apresentada, foi possível verificar que o guselcumab (fármaco proposto) tem vindo a apresentar resultados favoráveis quando comparado com adalimumab, nomeadamente, no alcance de PASI75, PASI90, PASI100 e IGA (Grau de Evidência: Nível 1 de Oxford).

comparações indiretas entre os três agentes, mas também entre os agentes guselcumab e tildracizumab, não tendo este último permitido concluir da superioridade de um tratamento em relação a outro (Grau de Evidência: Nível 1 de Oxford).

As sociedades científicas não fornecem linhas orientadoras suficientemente específicas para a eleição de um tratamento biológico em detrimento de outro. Apesar de fornecerem condições para a utilização dos agentes biológicos tanto na sua generalidade como no que toca a algumas condições de utilização de vários agentes, estas são muitas vezes sobreponíveis dentro da mesma classe de fármaco e também entre classes.

Já no que concerne aos pareceres da agência reguladora nacional, o INFARMED, com a colaboração da CNFT, o guselcumab apresentou uma avaliação prévia hospitalar favorável em comparação com adalimumab, tanto no que diz respeito ao seu valor terapêutico como na avaliação económica. Contudo, foram lançadas posteriormente a esta avaliação, algumas considerações enunciando que o guselcumab deverá ser considerado como uma 2ª linha de tratamento (no nível dos medicamentos biológicos e equivalentes), a par com os restantes agentes anti-IL23, a não ser que seja devidamente justificada a sua utilização em 1ª linha. Para além disso, levando em consideração a relação de preços atual, pela disponibilidade de medicamentos biossimilares anti-TNFα, é recomendado iniciar o tratamento por um medicamento desse grupo sempre que for considerado adequado à situação clínica do doente dado tratar-se de uma opção mais custo-efetiva para o SNS.

CONCLUSÃO

Atendendo aos vários aspetos farmacoterapêuticos e farmacoeconómicos mencionados ao longo do presente parecer, tanto no que concerne ao guselcumab como a outras opções terapêuticas biológicas com potencial de tratamento em doente com psoríase grave e refratária, verificamos que o guselcumab mostrou uma tendência para um maior benefício clínico quando comparado com adalimumab (opção terapêutica que, segundo o protocolo institucional implementado, se colocaria apenas em doentes com idade >25 anos); no entanto, quando comparado com outros fármacos da mesmo classe, fármacos anti- IL23, este benefício não se mostrou tão claro. Para além disso, apesar do guselcumab apresentar enquadramento de acordo com o protocolo institucional para o caso clínico em causa neste documento, também os fármacos risanquizumab e tildracizumab se encontram na mesma linha terapêutica, com igual enquadramento. Posto isto, e considerando o recentemente enunciado pelas autoridades regulamentares nacionais, prevendo-se resposta satisfatória com qualquer dos agentes da classe mencionada, é recomendado privilegiar a opção farmacológica com melhor perfil farmacoeconómico.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Agency EM. Resumo das Características do Medicamento: Tremfya® (guselcumab) 2017 [Available from: https://www.ema.europa.eu/en/documents/product-information/tremfya-epar-product- information_pt.pdf.

2. Administration FaD. Prescribing Information: Tremfya® (guselkumab) 2017 [Available from: https://www.accessdata.fda.gov/drugsatfda_docs/label/2020/761061s007lbl.pdf.

3. Blauvelt A, Papp KA, Griffiths CE, Randazzo B, Wasfi Y, Shen YK, et al. Efficacy and safety of guselkumab, an anti-interleukin-23 monoclonal antibody, compared with adalimumab for the continuous treatment of patients with moderate to severe psoriasis: Results from the phase III, double-blinded, placebo- and active comparator-controlled VOYAGE 1 trial. J Am Acad Dermatol. 2017;76(3):405-17.

(…)

Anexo B. Categorização do histórico de pareceres

ID Parecer	Título	Ano	Classificação
6	Guselcumab na psoríase em placas grave refratária	2022	Evidência científica do medicamento em determinado contexto clínico/doença
8	Associação do durvalumab à quimioterapia no tratamento do colangiocarcinoma metastizado	2023	Evidência científica do medicamento em determinado contexto clínico/doença
13	Sacituzumab govitecano no carcinoma da mama triplo negativo	2023	Evidência científica do medicamento em determinado contexto clínico/doença
17	Lanadelumab na profilaxia de longa duração no Angioedema Hereditário do Tipo I	2023	Evidência científica do medicamento em determinado contexto clínico/doença
18	Risdiplam na atrofia muscular espinhal tipo 3	2023	Evidência científica do medicamento em determinado contexto clínico/doença
26	Eficácia e segurança de omalizumab com mepolizumab num doente com micose broncopulmonar alérgica associada a pneumonia eosinofílica	2023	Evidência científica do medicamento em determinado contexto clínico/doença
31	Rechallenge de quimioterapia em associação com Pembrolizumab em doente com recidiva de Adenocarcinoma Pulmonar	2023	Evidência científica do medicamento em determinado contexto clínico/doença
32	Eficácia e segurança de lorlatinib com selpercatinib num doente com carcinoma pulmonar não pequenas células estádio IV	2023	Evidência científica do medicamento em determinado contexto clínico/doença
38	Filgotinib no tratamento da pancolite ulcerosa em doentes jovens em idade reprodutiva	2023	Evidência científica do medicamento em determinado contexto clínico/doença
43	Tratamento adjuvante com ribociclib em contexto de carcinoma da mama precoce, com RH positivo _ HER2 negativo	2023	Evidência científica do medicamento em determinado contexto clínico/doença
3	Belimumab para doente com Lupus Eritematoso Juvenil, nefrite lúpica e provável imunodeficiência primária	2024	Evidência científica do medicamento em determinado contexto clínico/doença

6	Avaliação de segurança e eficácia no tratamento de melanoma estádio IV com terapêutica combinada com pembrolizumab e lenvatinib	2024	Evidência científica do medicamento em determinado contexto clínico/doença
12	Pembrolizumab no Adenocarcinoma do Cólon Metastizado	2024	Evidência científica do medicamento em determinado contexto clínico/doença
13	Tratamento off-label com inibidor da PARP no adenocarcinoma do cólon metastizado	2024	Evidência científica do medicamento em determinado contexto clínico/doença
14	Utilização off-label de dupilumab no tratamento da alopécia universalis	2024	Evidência científica do medicamento em determinado contexto clínico/doença
16	Canabidiol na Encefalopatia Epilética relacionada com CDKL5	2024	Evidência científica do medicamento em determinado contexto clínico/doença
1	Osimertinib adjuvante no tratamento do adenocarcinoma do pulmão localmente avançado	2025	Evidência científica do medicamento em determinado contexto clínico/doença
4	Ustecinumab versus Tildracizumab off-label, na Artrite Psoriática	2023	Eficácia e segurança de um medicamento vs comparadores ou outras terapêuticas disponíveis,Evidência de eficácia e/ou segurança no tratamento de determinada doença vs standard of care
11	letermovir em detrimento de ganciclovir valganciclovir em doentes com candidatos a alotransplante de dador relacionado com serologia negativa para citomegalovirus (CMV) e recetor positivo	2023	Eficácia e segurança de um medicamento vs comparadores ou outras terapêuticas disponíveis,Evidência de eficácia e/ou segurança no tratamento de determinada doença vs standard of care
45	Abiraterona vs. Apalutamida no Carcinoma da Próstata Hormonossensível Metastizado	2023	Evidência de eficácia e/ou segurança no tratamento de determinada doença vs standard of care
3	Tratamento off-label com anacinra no síndrome autoinflamatório	2023	Opções terapêuticas no tratamento de determinada doença
35	Opções de tratamento de primeira linha com imunoterapia em contexto de melanoma metastático BRAF wild-type	2023	Opções terapêuticas no tratamento de determinada doença
2	Guselcumab vs Ustecinumab no tratamento da artrite psoriática e psoríase em placas	2023	Eficácia e segurança de um medicamento vs comparadores ou outras terapêuticas disponíveis

4	Ustecinumab versus Tildracizumab off-label, na Artrite Psoriática	2023	Eficácia e segurança de um medicamento vs comparadores ou outras terapêuticas disponíveis,Evidência de eficácia e/ou segurança no tratamento de determinada doença vs standard of care
11	letermovir em detrimento de ganciclovir valganciclovir em doentes com candidatos a alotransplante de dador relacionado com serologia negativa para citomegalovirus (CMV) e recetor positivo	2023	Eficácia e segurança de um medicamento vs comparadores ou outras terapêuticas disponíveis,Evidência de eficácia e/ou segurança no tratamento de determinada doença vs standard of care
23	Risco de malignidade com o uso de terapêutica anti-TNFalfa comaprada com inibidores IL-23	2023	Eficácia e segurança de um medicamento vs comparadores ou outras terapêuticas disponíveis
29	Selinexor vs Trametinib no Glioblastoma Recorrente	2023	Eficácia e segurança de um medicamento vs comparadores ou outras terapêuticas disponíveis
52	Efetividade e segurança do rituximab (off-label) versus ocrelizumab no tratamento da Esclerose Múltipla Primária Progressiva	2023	Eficácia e segurança de um medicamento vs comparadores ou outras terapêuticas disponíveis
20	Ponatinib versus dasatinib no tratamento da Leucemia Linfoblástica Aguda Philadelphia positivo	2024	Eficácia e segurança de um medicamento vs comparadores ou outras terapêuticas disponíveis
15	Ustecinumab em regime posológico off-label na Doença Inflamatória Intestinal	2023	Critérios de monitorização de um fármaco
16	Metodologias de doseamento dos fármacos anti-TNF alfa e dos anticorpos antifármaco na Doença Inflamatória Intestinal	2023	Critérios de monitorização de um fármaco
19	Tratamento com Burosumab no adulto com raquitismo	2023	Critérios de iniciação/suspensão de um fármaco; Critérios de monitorização de um fármaco
7	Limiar clínico de doença cardíaca para uso de enzalutamida em detrimento de abiraterona em doentes com carcinoma da próstata hormonorresistente	2023	Critérios de iniciação/suspensão de um fármaco
19	Tratamento com Burosumab no adulto com raquitismo	2023	Critérios de iniciação/suspensão de um fármaco; Critérios de monitorização de um fármaco
28	Risco tromboembólico e cardiovascular de Baricitinib em doente com dermatite atópica, obesidade, sob contracepção oral	2023	Critérios de iniciação/suspensão de um fármaco

37	Reinício da terapêutica imunossupressora ou início de tratamento com antifibrótico na Pneumonite de Hipersensibilidade	2023	Critérios de iniciação/suspensão de um fármaco
39	Continuação do tratamento com omalizumab na urticária crónica espontânea	2023	Critérios de iniciação/suspensão de um fármaco
42	Suspensão do tratamento com omalizumab em doentes com urticária crónica espontânea	2023	Critérios de iniciação/suspensão de um fármaco

Anexo C. Principais características do BioMCP

Dimensão	Resumo
Licença e maturidade	Projeto **open-source (MIT)**, com documentação clara e *releases* frequentes; integra-se por STDIO (local/IDE) e HTTP "*streamable*" para uso em agentes/LLMs.
Ferramentas núcleo	1. *search* (pesquisa unificada por campos/domínios); 2. *fetch* (obter detalhes por PMID/DOI/ID).
Ferramentas por domínio	**- Artigos:** *article_searcher*, *article_getter* ; - **Ensaios clínicos:** ***trial_searcher, trial_getter, trial_protocol_getter, trial_outcomes_getter***, *trial_locations_getter, trial_references_getter*. - **Genómica/biomarcadores:** *variant_searcher, variant_getter* (e *getters* de gene/doença/fármaco).
Fontes de dados integradas	**- Literatura:** PubMed/PubTator3, Europe PMC, bioRxiv/medRxiv; - **Ensaios:** ClinicalTrials.gov e NCI CTS; - **Genómica:** MyVariant, MyGene, MyDisease, cBioPortal; - **Regulatório/Safety:** OpenFDA (FAERS, rotulagem, MAUDE).
Linguagem de *query*	**- Unificada por campos** (ex.: "*articles.date: >2018 AND disease*: *Non-small cell lung cancer AND drug: osimertinib")* e **por domínio**; - Suporte a sinónimos e filtros por tipo de estudo, data, população, etc.
Formatos de saída	**-** JSON estruturado (para integração programática) e respostas textuais/markdown para consumo humano; - Tenta enriquecer com metadados adicionais quando disponíveis.
***Workflows* típicos**	1. Recuperar **Top-N artigos** relevantes por fármaco/doença/biomarcador; 2. Verificar elegibilidade/localizações de ensaios clínicos; 3. Analisar variantes/genética clínica com ligação a literatura e bases biomédicas.

Anexo D. *Queries* para o BioMCP relativos aos relatórios 1 e 2

Relatório	Enquadramento	Caso Clínico	*Query* formulada para o BioMCP
Ustecinumab em regime posológico off-label na Doença Inflamatória Intestina	… no âmbito do tratamento da Doença Inflamatória Intestinal, parecer acerca da evidência científica atual para a utilização de ustecinumab em regimes posológicos off-label, definição de critérios clínicos e laboratoriais que possam justificar a intensificação e critérios de monitorização terapêutica.	[vazio]	*"Search for high-quality studies (randomized controlled trials, systematic reviews/meta-analyses, clinical practice guidelines, and large observational cohorts/registries) evaluating ustekinumab in inflammatory bowel disease (IBD: Crohn's disease and ulcerative colitis) with off-label dosing and dose-intensification strategies (e.g., subcutaneous 90 mg every 4 weeks or every 6 weeks vs standard every 8 weeks; IV re-induction or higher induction doses). Emphasize evidence defining clinical and laboratory criteria that justify intensification (primary non-response, secondary loss of response, persistent symptoms, elevated C-reactive protein, fecal calprotectin, and endoscopic activity) and therapeutic drug monitoring (TDM) (trough concentrations/exposure–response relationships, proposed target levels, and anti-ustekinumab antibodies/immunogenicity)."*
Tratamento com Burosumab no adulto com raquitismo	… vem por este meio solicitar, à Unidade de Farmacologia Clínica (UFC), parecer por forma a ser estabelecido um plano de monitorização do tratamento com Burosumab no adulto, nomeadamente com os critérios que poderão consubstanciar a interrupção do tratamento.	**Paciente:** XXXX **Antecedentes pessoais**: Diagnóstico de raquitismo hipofosfatémico ligado ao cromossoma X, aos 2 anos. Submetida a cirurgia ortopédica em 2002 para correção de varismo, apresentando ainda deformidade ossea (genus valgus) nos membros inferiores. Espessamento do nervo ótico à direita, com amaurose quase completa ipsilateral. **(…)**	*"Search for high-quality studies (randomized controlled trials, systematic reviews/meta-analyses, clinical practice guidelines/consensus statements, long-term extension studies, and real-world cohorts/registries) evaluating burosumab for adult X-linked hypophosphatemia (XLH)/hypophosphatemic rickets/osteomalacia. Emphasize adult dosing and titration regimens and evidence-based monitoring frameworks (serum phosphate targets, TmP/GFR, alkaline phosphatase, calcium, 1,25-dihydroxyvitamin D, PTH, renal function/urinalysis) and criteria for treatment interruption/discontinuation (persistent hyperphosphatemia, worsening nephrocalcinosis or other ectopic/vascular calcifications, renal impairment, lack of biochemical or clinical response, pregnancy)."*

Anexo E. *Gold Standard* de evidência científica relativa aos 3 relatórios técnicos em análise

Relatório	Título	Autores	Data	Tipologia	Relevância Gradual	Justificação_Inclusão
P1	Osimertinib after Chemoradiotherapy in Stage III EGFR-Mutated NSCLC (LAURA)	Lu S et al.	ago/24	Ensaio clínico aleatorizado (RCT)	3	Evidência pivotal que suporta consolidação com osimertinib.
P1	Management of Stage III Non-Small Cell Lung Cancer: ASCO Guideline Rapid Recommendation Update	Daly ME et al.	set/24	Guideline / Consenso	3	Guideline de prática clínica diretamente aplicável ao cenário em questão.
P1	LAURA: análises de eficácia no SNC e progressão distante	Lu S et al.	dez/24	Outro	3	Relevante para endpoints críticos (SNC) no cenário de consolidação.
P1	LAURA Completes the Osimertinib Treatment Jigsaw Puzzle of EGFR+ NSCLC from Stage IB to IV: Adjuvant Osimertinib Significantly Improves PFS and CNS Progression in Unresectable Stage III EGFR-Mutated NSCLC Compared to Placebo (LAURA, NCT03521154)	Luo et al.	abr/25	Revisão narrativa	3	Resumo científico e interpretação dos resultados do ensaio LAURA, útil como enquadramento.
P1	Comparison of treatment regimens for unresectable stage III epidermal growth factor receptor (EGFR) mutant non-small cell lung cancer	Dai et al.	jul/25	Meta-análise	2	Revisão comparativa de esquemas pós-QT/RT em EGFR+, incluindo TKI vs durvalumab
P1	Multidisciplinary approach for locally advanced non-small cell lung cancer (NSCLC): 2023 expert consensus of the Spanish Lung Cancer Group GECP	Naidoo et al.	jul/24	Outro	2	Consenso nacional com recomendações práticas aplicáveis a EGFR+ estádio III
P1	Osimertinib Maintenance After Definitive Chemoradiation in Patients With Unresectable EGFR Mutation Positive Stage III Non-small-cell Lung Cancer: LAURA Trial in Progress	Lu et al.	jul/21	Ensaio clínico aberto/não aleatorizado	2	Publicação de protocolo/desenho do LAURA, relevante para compreender endpoints e critérios.

P1	Exciting progress in targeted therapy innovation for unresectable stage III EGFR-mutated NSCLC: the phase III LAURA study	Tong et al.	dez/24	Revisão narrativa	2	Revisão narrativa que sumariza impacto do ensaio LAURA e terapias alvo no cenário.
P1	LAURA (NCT03521154) – protocolo do ensaio clínico	ClinicalTrials.gov	jan/24	Outro	2	Descrição detalhada do desenho e endpoints — útil para protocolo e monitorização.
P1	Efficacy and safety of tyrosine kinase inhibitors with thoracic radiotherapy for patients with oncogene-mutated non-small cell lung cancer: a meta-analysis	Li et al.	nov/24	Meta-análise	1	Meta-análise que avalia integração de TKIs com RT, relevante como enquadramento de segurança/eficácia.
P1	Optimization of treatment strategies for elderly patients with advanced non-small cell lung cancer	Chen et al.	ago/24	Revisão narrativa	1	Contexto específico de idosos EGFR+, complementa decisão em paciente de 75 anos.
P1	Brief Report: Durvalumab After Chemoradiotherapy in Unresectable Stage III EGFR-Mutant NSCLC: A Post Hoc Subgroup Analysis From PACIFIC	Naidoo et al.	mai/23	Outro	1	Dados do comparador (durvalumab) em doentes EGFR+. Importante, ainda que não mencione diretamente a substância ativa em análise (osimertinib)
P1	Targeted therapies for unresectable stage III non-small cell lung cancer	Remon & Hendriks	set/21	Revisão narrativa	1	Revisão de terapias alvo, menciona osimertinib mas com menor profundidade.
P1	Targeted treatment for unresectable EGFR mutation-positive stage III non-small cell lung cancer: Emerging evidence and future perspectives	Kato et al.	jan/24	Revisão narrativa	1	Perspetiva futura sobre integração de osimertinib em doença localmente avançada
P2	Effectiveness of Reinduction and/or Dose Escalation of Ustekinumab in Crohn's Disease: Systematic Review and Meta-analysis	Meserve J et al.	jan/22	Meta-análise	3	Síntese quantitativa mais robusta sobre intensificação de UST.

P2	Effectiveness and Safety of Ustekinumab Intensification at 90 mg Every Four Weeks in Crohn's Disease: Multicenter Study	Fumery M et al.	set/20	Estudo de coorte retrospetivo	3	Evidência pivotal de encurtamento para q4s em vida real.
P2	Systematic Review With Meta-Analysis: Loss of Response and Requirement of Ustekinumab Dose Escalation in Inflammatory Bowel Diseases	Yang H et al.	abr/22	Meta-análise	3	Confirma robustamente eficácia de escalada em IBD.
P2	The Real-World Effectiveness and Safety of Ustekinumab in the Treatment of Crohn's Disease: Results From the SUCCESS Consortium	Johnson AM et al.	fev/23	Estudo de coorte retrospetivo	3	Evidência de mundo real multicêntrica com amostras grandes.
P2	ACG Clinical Guideline: Management of Crohn's Disease in Adults	Lichtenstein GR et al.	jun/25	Guideline / Consenso	3	Guideline atual com recomendação explícita de intensificação.
P2	Efficacy of Dose Escalation of Ustekinumab in Inflammatory Bowel Disease: A Systematic Review and Meta-Analysis	Rehman et al.	ago/25	Revisão sistemática	3	Revisão sistemática robusta sobre escalada de dose de UST em DII.
P2	Systematic Review and Meta-analysis: The Association Between Serum Ustekinumab Trough Concentrations and Treatment Response in Inflammatory Bowel Disease	Vasudevan et al.	abr/24	Revisão sistemática	3	Síntese de evidência ligando níveis de vale a resposta clínica/endoscópica.
P2	Effectiveness of Ustekinumab Dose Escalation in Patients With Crohn's Disease	Ollech JE et al.	jan/21	Estudo de coorte retrospetivo	2	Confirma benefício do encurtamento de intervalo em prática real.
P2	Long-Term Outcomes After Ustekinumab Dose Intensification for Inflammatory Bowel Diseases	Dalal RS et al.	mai/23	Estudo de coorte retrospetivo	2	Evidência de durabilidade e segurança no longo prazo.
P2	Effectiveness of Ustekinumab Dose Escalation in Crohn's Disease Patients With Insufficient Response to Standard-Dose Subcutaneous Maintenance Therapy	Kopylov U et al.	jul/20	Estudo de coorte retrospetivo	2	Suporte consistente em prática real.
P2	Efficacy of Ustekinumab Optimization by 2 Initial Intravenous Doses in Adult Patients With Severe Crohn's Disease	Ren H et al.	ago/24	Estudo de coorte retrospetivo	2	Explora reindução/alta dose em casos graves.

P2	Predictors and Outcomes of Ustekinumab Dose Intensification in Ulcerative Colitis: A Multicenter Cohort Study	Dalal RS et al.	out/22	Estudo de coorte retrospetivo	2	Evidência específica para UC, apoio real-world.
P2	Relationship Between Ustekinumab Trough Concentrations and Clinical, Biochemical and Endoscopic Outcomes in Crohn's Disease: TARGET Study	Shehab M et al.	jul/24	Estudo de coorte retrospetivo	2	Informa alvos TDM e racional para otimização.
P2	Long-Term Efficacy of Therapeutic Drug Monitoring-Guided Optimization of Ustekinumab Maintenance Therapy for Crohn's Disease Patients	Ren H et al.	fev/25	Estudo de coorte retrospetivo	2	Apoia TDM como estratégia para intensificação racional.
P2	Extra intravenous Ustekinumab reinduction is an effective optimization strategy for patients with refractory Crohn's disease	Yao et al.	jul/24	Estudo de coorte retrospetivo	2	Mostra benefício da reindução IV em doentes refratários.
P2	Intravenous ustekinumab maintenance treatment in patients with loss of response to subcutaneous dosing	Argüelles-Arias et al.	ago/23	Estudo de coorte retrospetivo	2	Evidência de manutenção IV como estratégia off-label após falha SC.
P2	Ustekinumab trough levels predicting laboratory and endoscopic remission in patients with Crohn's disease	Hirayama et al.	abr/22	Estudo de coorte prospectivo	2	Define potenciais cut-offs de vale associados a remissão clínica/endoscópica.
P2	Effectiveness and safety of ustekinumab dose escalation in Crohn's disease: a multicenter observational study	Olmedo Martín et al.	dez/23	Estudo de coorte retrospetivo	2	Dados multicêntricos de vida real sobre intensificação posológica.
P2	Impact of immediate drug optimization of ustekinumab on medium term targets in Crohn's disease: results from the multicentre retrospective real-life study MUST	Hupé et al.	set/25	Estudo de coorte retrospetivo	2	Mostra eficácia clínica da otimização imediata em prática real.
P2	IBD Patients with Primary or Secondary Nonresponse to Ustekinumab Benefit from Dose Escalation or Reinduction	Vernia et al.	jul/24	Revisão narrativa	2	Suporta benefício clínico da escalada/reindução em não respondedores.
P2	How to Optimize Treatment With Ustekinumab in Inflammatory Bowel	Gutiérrez & Rodríguez-Lago	ago/21	Revisão sistemática	2	Revisão que compila evidência de ensaios e prática real sobre otimização.

	Disease: Lessons Learned From Clinical Trials and Real-World Data					
P2	Ustekinumab Drug Clearance Is Better Associated with Disease Control than Serum Trough Concentrations in a Prospective Cohort of Inflammatory Bowel Disease	Yarur et al.	fev/25	Estudo de coorte prospectivo	2	Dados de PK mostrando clearance como marcador mais fiável para resposta.
P2	Monthly intravenous maintenance treatment with ustekinumab regains clinical response in patients with Crohn's disease who no longer respond to the drug when administered subcutaneously	López-Sáez et al.	ago/25	Estudo de coorte retrospetivo	2	Demonstra eficácia da manutenção IV mensal em falhados SC.
P2	MOdel-Informed Precision Dosing of Ustekinumab and VEdolizumab in Inflammatory Bowel Disease (MOVE-IT)	Odense University Hospital	jul/25	Ensaio clínico aleatorizado (RCT)	2	Estudo em curso que explora doseamento personalizado com base em modelos PK.
P2	Dose Escalation Patterns of Advanced Therapies in Crohn's Disease and Ulcerative Colitis: A Systematic Literature Review	Panaccione et al.	mai/23	Revisão sistemática	1	Revisão mais abrangente, inclui UST entre várias terapias.
P2	A Review of Therapeutic Drug Monitoring in Patients with Inflammatory Bowel Disease Receiving Combination Therapy	Patel & Yarur	out/23	Revisão narrativa	1	Revisão geral de TDM, inclui implicações para UST em doentes combinados.
P2	Treatment escalation and de-escalation decisions in Crohn's disease: Delphi consensus recommendations from Japan, 2021	Nakase et al.	mar/23	Guideline / Consenso	1	Consenso especialista com recomendações sobre escalada/desescalada incluindo UST.
P2	Real-world long-term effectiveness of ustekinumab in ulcerative colitis: results from a spanish open-label cohort	Iborra et al.	mar/24	Estudo de coorte prospectivo	1	Dados de UC em vida real, inclui durabilidade do tratamento com possibilidade de intensificação.
P3	A Randomized, Double-Blind, Placebo-Controlled, Phase 3 Trial Evaluating the Efficacy of Burosumab in Adults With XLH (Week 24)	Insogna KL et al.	ago/18	Ensaio clínico aleatorizado (RCT)	3	RCT fase 3 em adultos com XLH — evidência central para eficácia/segurança.

P3	X-Linked Hypophosphatemia (GeneReviews®)	Laurent MR et al.	dez/23	Guideline / Consenso	3	Fonte prática para dose, titulação e critérios de suspensão; altamente aplicável à monitorização.
P3	Real-World Effectiveness of Burosumab Versus Oral Phosphate and Active Vitamin D in Adults With XLH	Florenzano P et al.	jul/25	Estudo de coorte retrospetivo	3	Evidência comparativa direta em adultos; altamente aplicável.
P3	Burosumab Efficacy and Safety in XLH: Systematic Review and Meta-Analysis of Real-World Data	Kiafzezi D et al.	set/24	Meta-análise	3	Síntese quantitativa focalizada em mundo real; reforça decisão terapêutica e monitorização.
P3	Systematic Review: Efficacy of Medical Therapy on Outcomes Important to Adult Patients With XLH	Ali DS et al.	mai/25	Revisão sistemática	3	Síntese transversal centrada no adulto; aplicabilidade clínica para PROs/qualidade de vida.
P3	X-Linked Hypophosphatemia Management in Adults: An International Working Group Clinical Practice Guideline	Khan et al.	jul/25	Guideline / Consenso	3	Guideline internacional específica para adultos; fornece recomendações diretas de monitorização e critérios de suspensão.
P3	Clinical practice recommendations for the diagnosis and management of X-linked hypophosphataemia	Haffner et al.	jul/19	Guideline / Consenso	3	Documento de consenso abrangente; inclui diagnóstico, seguimento e monitorização em adultos.
P3	Burosumab treatment in adults with XLH: 96-week patient-reported outcomes and ambulatory function from a randomised phase 3 trial and extension	Briot K et al.	set/21	Ensaio clínico aberto/não randomizado	2	Extensão do RCT com dados de longo prazo clinicamente relevantes.
P3	Continued Beneficial Effects of Burosumab in Adults with XLH: 24-week continuation after a 24-week double-blind period	Portale AA et al.	set/19	Ensaio clínico aberto/não randomizado	2	Evidência longitudinal complementar após o RCT.
P3	Burosumab Improved Histomorphometric Measures of Osteomalacia in Adults With XLH: Phase 3, Single-Arm, International Trial	Insogna KL et al.	dez/19	Ensaio clínico aberto/não randomizado	2	Complementa RCT com endpoints de histologia óssea relevantes.

P3	Predictors of Response to Burosumab in Adults With XLH: Real-World Data From an Italian Cohort	Arcidiacono GP et al.	mai/25	Estudo de coorte retrospetivo	2	Apoia estratificação/monitorização com base em fatores preditivos.
P3	Long-Term Burosumab Administration Is Safe and Effective in Adults With XLH	Weber TJ et al.	jan/22	Estudo de coorte prospectivo	2	Suporta segurança/eficácia a longo prazo — relevante para decisão contínua.
P3	Burosumab treatment of X-linked hypophosphatemia patients: interim analysis of the SUNFLOWER longitudinal, observational cohort study	Michigami et al.	jun/24	Estudo de coorte prospectivo	2	Dados observacionais do mundo real em adultos; complementa ensaios clínicos com evidência prática.
P3	Efficacy of Burosumab in Adults with X-linked Hypophosphatemia (XLH): A Post Hoc Subgroup Analysis of a Randomized Double-Blind Placebo-Controlled Phase 3 Study	Brandi et al.	out/22	Ensaio clínico aleatorizado (RCT)	2	Subanálise de RCT; reforça eficácia e segurança em adultos.
P3	What are the benefits of the anti-FGF23 antibody burosumab on the manifestations of X-linked hypophosphatemia in adults in comparison with conventional therapy? A review	Lafage-Proust	fev/22	Revisão narrativa	2	Revisão comparativa com terapêutica convencional; útil para contextualizar monitorização.
P3	Asia-Pacific Consensus Recommendations on X-Linked Hypophosphatemia: Diagnosis, Multidisciplinary Management, and Transition From Pediatric to Adult Care	Munns et al.	mai/23	Guideline / Consenso	2	Consenso regional; aborda seguimento e transição adulto, com recomendações de monitorização.
P3	X-Linked Hypophosphatemia: A New Era in Management	Dahir et al.	out/20	Revisão narrativa	1	Revisão narrativa; enquadra novas abordagens terapêuticas, incluindo burosumab, mas menos prática para critérios de suspensão.

Anexo F. Artigos recuperados para o 1º Relatório, utilizando o BioMCP orquestrado pelo GPT-5

Nº Execução	Título	Autores	Data	Tipologia	Relevância	Integridade Metadados	Notas
1	Osimertinib after Chemoradiotherapy in Stage III EGFR-Mutated NSCLC (LAURA)	Lu S et al.	ago/24	Ensaio clínico aleatorizado (RCT)	1	1	
1	LAURA (NCT03521154) – protocolo do ensaio clínico	AstraZeneca	jan/24	Ensaio clínico aleatorizado (RCT)	1	1	
1	LAURA Completes the Osimertinib Treatment Jigsaw Puzzle of EGFR+ NSCLC from Stage IB to IV: Adjuvant Osimertinib Significantly Improves PFS and CNS Progression in Unresectable Stage III EGFR-Mutated NSCLC Compared to Placebo (LAURA, NCT03521154)	Luo et al.	abr/25	Revisão narrativa	1	1	
1	Comparison of treatment regimens for unresectable stage III epidermal growth factor receptor (EGFR) mutant non-small cell lung cancer	Dai et al.	jul/25	Meta-análise	1	1	
1	Efficacy and safety of tyrosine kinase inhibitors with thoracic radiotherapy for patients with oncogene-mutated non-small cell lung cancer: a meta-analysis	Li et al.	nov/24	Meta-análise	1	1	
2	Osimertinib after Chemoradiotherapy in Stage III EGFR-Mutated NSCLC (LAURA)	Lu S et al.	ago/24	Ensaio clínico aleatorizado (RCT)	1	1	
2	LAURA (NCT03521154) – protocolo do ensaio clínico	AstraZeneca	jan/24	Ensaio clínico aleatorizado (RCT)	1	1	
2	A Phase II Trial on Osimertinib as a First-Line Treatment for EGFR Mutation-Positive Advanced NSCLC in Elderly Patients: The SPIRAL-0 Study	Chihara et al.	nov/22	Ensaio clínico aberto/não aleatorizado	0	1	metastático, não pós-QT/RT

2	Five-year outcomes with gefitinib induction and chemoradiotherapy in EGFR-mutant stage III non-small-cell lung cancer: LOGIK0902/OLCSG0905 phase II study	Hotta et al.	mai/25	Ensaio clínico aberto/não aleatorizado	0	1	fase II, não osimertinib, não cenário pós-QT/RT definitivo
2	Optimization of treatment strategies for elderly patients with advanced non-small cell lung cancer	Chen et al.	ago/24	Revisão narrativa	1	1	
3	Osimertinib after Chemoradiotherapy in Stage III EGFR-Mutated NSCLC (LAURA)	Lu S et al.	ago/24	Ensaio clínico aleatorizado (RCT)	1	1	
3	LAURA (NCT03521154) – protocolo do ensaio clínico	AstraZeneca	jan/24	Ensaio clínico aleatorizado (RCT)	1	1	
3	LAURA Completes the Osimertinib Treatment Jigsaw Puzzle of EGFR+ NSCLC from Stage IB to IV: Adjuvant Osimertinib Significantly Improves PFS and CNS Progression in Unresectable Stage III EGFR-Mutated NSCLC Compared to Placebo (LAURA, NCT03521154)	Luo et al.	abr/25	Revisão narrativa	1	1	
3	Optimization of treatment strategies for elderly patients with advanced non-small cell lung cancer	Chen et al.	ago/24	Revisão narrativa	1	1	
3	Brief Report: Durvalumab After Chemoradiotherapy in Unresectable Stage III EGFR-Mutant NSCLC: A Post Hoc Subgroup Analysis From PACIFIC	Naidoo et al.	mai/23	Outro	1	1	comparador direto ao fármaco
4	Osimertinib after Chemoradiotherapy in Stage III EGFR-Mutated NSCLC (LAURA)	Lu S et al.	ago/24	Ensaio clínico aleatorizado (RCT)	1	1	
4	LAURA (NCT03521154) – protocolo do ensaio clínico	AstraZeneca	jan/24	Ensaio clínico aleatorizado (RCT)	1	1	
4	Five-year outcomes with gefitinib induction and chemoradiotherapy in EGFR-mutant stage III non-small-cell lung cancer: LOGIK0902/OLCSG0905 phase II study	Hotta et al.	mai/25	Ensaio clínico aberto/não aleatorizado	0	1	fase II, não osimertinib, não cenário pós-QT/RT definitivo
4	Optimization of treatment strategies for elderly patients with advanced non-small cell lung cancer	Chen et al.	ago/24	Revisão narrativa	1	1	

4	Multidisciplinary approach for locally advanced non-small cell lung cancer (NSCLC): 2023 expert consensus of the Spanish Lung Cancer Group GECP	Naidoo et al.	jul/24	Outro	1	1	
5	Osimertinib after Chemoradiotherapy in Stage III EGFR-Mutated NSCLC (LAURA)	Lu S et al.	ago/24	Ensaio clínico aleatorizado (RCT)	1	1	
5	LAURA (NCT03521154) – protocolo do ensaio clínico	AstraZeneca	jan/24	Ensaio clínico aleatorizado (RCT)	1	1	
5	A Phase II Trial on Osimertinib as a First-Line Treatment for EGFR Mutation-Positive Advanced NSCLC in Elderly Patients: The SPIRAL-0 Study	Chihara et al.	nov/22	Ensaio clínico aberto/não aleatorizado	0	1	metastático, não pós-QT/RT
5	Five-year outcomes with gefitinib induction and chemoradiotherapy in EGFR-mutant stage III non-small-cell lung cancer: LOGIK0902/OLCSG0905 phase II study	Hotta et al.	mai/25	Ensaio clínico aberto/não aleatorizado	0	1	fase II, não osimertinib, não cenário pós-QT/RT definitivo
5	Optimization of treatment strategies for elderly patients with advanced non-small cell lung cancer	Chen et al.	ago/24	Revisão narrativa	1	1	

Anexo G. Cálculo da métrica *Rank Biased Overlap (RBO)*

RBO é uma métrica para comparar listas ordenadas (rankings), dando mais peso aos primeiros itens da lista.

Explicação:

Com duas listas ordenadas de resultados:

- RBO calcula a semelhança cumulativa entre essas listas, com um determinado fator de decaimento p (ou seja, não trata todas as posições de forma igual):
    - os primeiros lugares têm muito peso;
    - os lugares mais abaixo contribuem cada vez menos.
- O parâmetro p controla esse decaimento:
    - $p = 0{,}9$ significa que cada posição pesa apenas 90% da posição anterior. Logo, quanto menor o valor de p, mais será a "importância" dada aos primeiros elementos.
    - O cálculo do decaimento é, por isto, uma distribuição geométrica:

$$w_k = (1 - p).p^{\{k-1\}}$$

- A este peso w multiplica-se a percentagem de itens iguais na lista até à posição k, obtendo, assim, o contributo a dar para o RBO até essa posição;
- O RBO final consiste nos contributos acumuladas de todas as posições.

Interpretação:

- Um RBO de 0,6 quer dizer que as listas ainda são razoavelmente parecidas, sobretudo no topo, mas há rotatividade significativa de artigos mais abaixo.

Anexo H. *System prompts* utilizados nas diferentes etapas de elaboração do relatório

Prompt para identificação do(s) fármaco(s) a serem analisados, bem como a classificação do relatório a ser gerado em uma das seis categorias possíveis

```
# Identifiying the medicine being analyzed and classifying the report into on of the existing categories
medicine_identification = """Com base no título e enquadramento do relatório:

Título: {title}
Enquadramento: {context}

1. Identifique o(s) medicamento(s) em análise.
2. Classifique o pedido realizado em uma das seguintes categorias:

{categories}

{format_instructions}"""
```

Prompt para geração de *queries* a serem feitas à base de dados vetorial que contém indexadas secções do RCM

```
# Generating queries based on similar reports to then query the vector database containing the RCMs (retrieving relevant sections of them)
farmacologia_query_generation = """Considerando:

- Os medicamentos identificados ({medicines});
- Contexto ({context});
- Caso clínico ({clinical_case}),

liste os principais aspectos farmacológicos que devem ser pesquisados nas bulas/RCMs para redigir uma secção semelhante.
```

```
Exemplo de secção:
---
{example_section}
---

{format_instructions}"""
```

Prompt para geração da secção “Fundamentação | Farmacologia”

```
# Generating the section "FUNDAMENTAÇÃO: FARMACOLOGIA" based on the
example section and the RCMs sections retrieved
farmacologia_section_generation = """Escreva a secção "FUNDAMENTAÇÃO:
FARMACOLOGIA" de um relatório técnico.

Use como referência:
1. A secção "FUNDAMENTAÇÃO: FARMACOLOGIA" do relatório exemplo:
---
{example_section}
---

2. As informações extraídas das bulas (RCMs) relevantes:
---

2.1. Informações de prescrição da FDA:
{fda_indications_usage}

2.2. Resumo das Características do Medicamento aprovado pela EMA:
{relevant_chunks_rcm}


---

3. O pedido em questão:
Título: {title}
```

```
Contexto: {context}
Caso clínico: {clinical_case}

Redija uma nova seção clara, precisa e objetiva baseada nas
informações acima, sem incluir informações irrelevantes ou
redundantes."""
```

Prompt para geração de *um prompt* otimizado para o BioMCP procurar evidência científica relevante que responda ao pedido

```
# Summarization prompt for later use in the MCP server
summarization_prompt = """
You are assisting with generating a highly specific PubMed search
query for an AI agent.

Given the following inputs:
- **Title**: {title}
- **Context**: {context}
- **Clinical Case**: {clinical_case}

Your task is to synthesize a precise PubMed search prompt. It must:
- Prioritize study types such as randomized controlled trials, meta-
analyses, systematic reviews, clinical guidelines, and prospective
cohort studies
- Explicitly incorporate the medicine(s), condition(s),
mutations/biomarkers, and clinical setting provided in the inputs. Do
not retrieve articles that do not meet these criteria.

---

Use this prompt as an example for inspiration, adapting it precisely
to the inputs:
```

```
'Search for high-quality studies (randomized controlled trials, meta-analyses, clinical guidelines, and cohort studies) evaluating osimertinib as consolidation therapy after definitive chemoradiotherapy in unresectable, locally advanced non-small cell lung cancer (NSCLC),
specifically adenocarcinoma, with EGFR exon 19 deletion mutations (and other EGFR mutations if evidence is limited).
Focus on patients with stage III disease (e.g., IIIB, T3N2M0) who did not progress following platinum-based chemoradiotherapy with curative intent,
and consider data on efficacy (progression-free survival, overall survival, response rates) and safety/tolerability
in older patients (around 70–80 years, ECOG 1).'

Output only the final PubMed search prompt, with no additional comments or instructions.
"""
```

Prompt para geração da secção “Fundamentação | Estudos Relevantes”

```
# Generating the section "FUNDAMENTAÇÃO: ESTUDOS RELEVANTES" based on the example section and articles retrieved
estudos_section = """Com base nas informações fornecidas abaixo, elabore uma versão original para uma seção de relatório técnico intitulada **"FUNDAMENTAÇÃO: ESTUDOS RELEVANTES"**.

Use as seguintes referências como base:

1. **Secção de exemplo retirada de um relatório anterior semelhante** (use como inspiração de estrutura e estilo):
{example_section}
```

```
2. **Contexto do Relatório**:
{context}

3. **Resumo dos Artigos Científicos Relevantes**:
{articles_summary}


---

### Instruções:

- Estruture a secção de forma coesa, explicando como os estudos selecionados se relacionam com o caso clínico e o uso do(s) medicamento(s), detalhadamente.
- Mantenha uma linguagem objetiva, impessoal e apropriada para relatórios médicos/técnicos.
- Caso não haja estudos relevantes, apenas mencione a ausência de evidências específicas para o caso, sem entrar em detalhes desnecessários.
- Evite incluir comentários sobre o processo de escrita ou instruções adicionais.

---

### Formato de Saída:
- **Título do Artigo** (link):
  - Descrição detalhada do estudo, evidenciando a sua relevância para o contexto do relatório e o caso clínico em questão

- **Título do Artigo** (link):
  - Descrição detalhada do estudo, evidenciando a sua relevância para o contexto do relatório e o caso clínico em questão
```

```
...

**Resumo** das evidências científicas apresentadas, destacando a
importância dos estudos para a fundamentação do uso do(s)
medicamento(s) no caso clínico analisado.

---

Retorne apenas o texto final da secção, sem comentários ou instruções
adicionais.
"""
```

Prompt para geração da secção “Considerações Finais”

```
# Generating the section "CONSIDERAÇÕES FINAIS" based on the entire
report
consideracoes_finais_section = """Com base no conteúdo completo do
relatório, redija a seção "CONSIDERAÇÕES FINAIS", sintetizando de
forma clara, precisa e objetiva:

- As principais conclusões do caso clínico analisado;
- A relevância e os efeitos do(s) medicamento(s) utilizado(s);
- As evidências científicas apresentadas (estudos e fundamentos
farmacológicos);
- Os pareceres emitidos por sociedades médicas e científicas
pertinentes.

Use as seguintes informações como base para a redação:

1. Título do Relatório: {title}
2. Contexto Geral: {context}
3. Descrição do Caso Clínico: {clinical_case}
```

```
4. Fundamentação Farmacológica: {farmacologia_section}
5. Estudos Científicos Relevantes: {estudos_section}
6. Pareceres de Sociedades Médicas/Científicas: {society_opinions}

Utilize o exemplo a seguir como modelo de estilo, estrutura e tom:

{example_section}

Importante: A nova seção deve ser original e redigida com base nas
informações fornecidas, sem incluir comentários sobre o próprio
processo de escrita ou instruções.
"""
```

Prompt para geração secção "Conclusão"

```
# Generating the section "Conclusão" with the final suggestions on the
question at hand
conclusao_section = """Com base no relatório completo, redija a seção
"Conclusão" de forma clara e objetiva, abordando:

  - As principais recomendações como resposta à questão levantada;
  - A relevância dos achados para a prática clínica;

  Use o exemplo a seguir como modelo de estilo, estrutura e tom:

  {example_section}

  Importante: A nova seção deve ser original e redigida com base nas
informações fornecidas, sem incluir comentários sobre o próprio
processo de escrita ou instruções.
  """
```

Anexo I. *Keywords* definidas para cada relatório

***Keywords* Relatório 1:**

"**Fármaco**": [
"osimertinib", "tagrisso",
"tki", "egfr-tki", "tki do egfr", "tki do receptor egfr",
"inibidor de tirosina-quinase", "inibidor da tirosina-quinase",
"inibidor de tirosina quinase", "inibidor da tirosina quinase",
"inibidor de tirosina cinase", "inibidor da tirosina cinase",
"inibidores de tirosina-quinase", "inibidores de tirosina quinase",
"tyrosine kinase inhibitor", "tyrosine-kinase inhibitor"
],

"**Contexto**": [
"adjuvância", "adjuvante", "terapia adjuvante", "tratamento adjuvante",
"quimiorradioterapia", "quimio-radioterapia", "radioquimioterapia",
"quimiorradiação", "crt", "qt/rt", "rt/qt",
"concomitante", "em concomitância",
"sequencial", "em sequência"
],

"**Doença**": [
"cancro do pulmão de não pequenas células",
"carcinoma do pulmão de não pequenas células",
"carcinoma pulmonar de não pequenas células",
"cpnpc", "nsclc", "non-small cell lung cancer",
"adenocarcinoma", "adenocarcinomas",
"localmente avançado", "doença localmente avançada",
"irressecável", "irressecavel", "não ressecável", "nao ressecavel", "inoperável", "inoperavel",
"estadio iii", "estágio iii", "estagio iii", "stage iii",
"estadio iiia", "estadio iiib", "estadio iiic",
"iiia", "iiib", "iiic",
"tnm", "t3n2m0", "t3", "n2", "m0"
],

"**Biomarcadores e mutações**": [
"egfr", "egfrm", "egfr mutado", "receptor do fator de crescimento epidérmico",
"deleção do exon 19", "deleção do exão 19", "delecao do exon 19", "delecao do exao 19",
"exon 19", "exão 19", "exon19", "exão19",
"ex19del", "del19", "19del",
"l858r", "l-858r",
"t790m",

"mutação sensibilizante", "mutações sensibilizantes", "sensitizing mutation", "sensitizing mutations",
"pd-l1", "pdl1", "pd l1", "expressão de pd-l1", "tps"
],

"**Ensaio clínico**": [
"laura",
"fase iii", "fase 3", "phase iii", "phase 3",
"dupla-ocultação", "duplo-cego", "double-blind",
"placebo",
"randomizado", "aleatorizado", "randomização", "randomizacao", "randomization",
"estratificação", "estratificar", "estratos", "stratification"
],

"**Endpoints e estatística**": [
"pfs", "sobrevida livre de progressão", "sobrevida sem progressão", "ssp", "progression-free survival",
"os", "sobrevida global", "sg", "overall survival",
"orr", "taxa de resposta objetiva", "tro", "objective response rate",
"dor", "duração de resposta", "duration of response",
"hazard ratio", "hr", "razão de risco", "razão de hazards",
"mediana", "intervalo de confiança", "ic 95%", "95% ci", "p-valor", "p valor", "p-value"
],

"**Eventos adversos**": [
"eventos adversos", "eventos adversos graves", "eag", "grau 3-4", "grau ≥3", "grau >=3",
"ctcae", "ctcae v5", "ctcae v5.0",
"pneumonite rádica", "pneumonite por radiação", "pneumonite induzida por rt",
"doença pulmonar intersticial", "dpi", "interstitial lung disease", "ild",
"diarreia", "diarréia",
"rash", "erupção cutânea", "erupcao cutanea",
"paroníquia", "paroniquia",
"pele seca", "xerose", "xerose cutânea", "xerose cutanea",
"contagem leucocitária", "leucócitos", "leucocitos", "leucopenia",
"pneumonia", "gastroenterite",
"creatina fosfocinase", "creatinofosfoquinase", "ck", "cpk"
],

"**Tratamento prévio**": [
"platina", "duplo de platina", "derivados de platina",
"cisplatina", "carboplatina",
"pemetrexedo", "pemetrexed",
"radioterapia 60 gy", "rt 60 gy", "≥60 gy", ">=60 gy", "radioterapia definitiva"

```
 ],

 "Posologia e Farmacocinética": [
   "dose 80 mg", "80 mg qd", "80 mg diariamente", "dose diária 80 mg", "dose diaria 80 mg",
   "semivida 44 horas", "meia-vida 44 horas", "meia vida 44 horas", "t1/2 44 h", "t1/2 ~44 h",
   "depuração 14,3 l/h", "depuracao 14,3 l/h", "depuração 14.3 l/h", "clearance 14.3 l/h", "cl
14.3 l/h",
   "volume de distribuição 918 l", "volume de distribuicao 918 l", "vd 918 l", "volume aparente
918 l"
 ],

 "Diretrizes / agências": [
   "esmo", "diretrizes esmo", "esmo guidelines",
   "asco", "diretrizes asco", "asco guidelines",
   "nccn", "diretrizes nccn", "nccn guidelines",
   "ema", "european medicines agency",
   "fda", "food and drug administration",
   "nice", "nice guidance", "nice guideline"
 ]
}
```

***Keywords* Relatório 2:**

```
 "Fármaco": [
   "ustekinumab", "ustecinumab",  # grafia correta + comum erro
   "stelara",
   "anticorpo monoclonal", "anticorpos monoclonais",
   "igg1k", "igg1κ", "igg1 kappa", "igg1-kappa",
   "il-12", "il 12", "interleucina 12", "interleukin 12",
   "il-23", "il 23", "interleucina 23", "interleukin 23",
   "p40", "subunidade p40", "p-40",
   "anti il-12", "anti il-23", "anti il12", "anti il23", "anti-il12/23", "anti il-12/23"
 ],

 "Doença": [
   "doença inflamatória intestinal", "doencas inflamatorias intestinais", "dii",
   "inflammatory bowel disease", "ibd",
   "doença de crohn", "crohn", "crohn's disease",
   "colite ulcerosa", "colite ulcerativa", "ulcerative colitis",
```

"moderada a grave", "moderada a severa", "moderada-grave", "moderada-severa"
],

"**Regimes on-label (RCM) e posologia**": [
"indução intravenosa", "induçao intravenosa", "inducao intravenosa", "indução iv", "induçao iv", "inducao iv",
"intravenosa", "iv",
"6 mg/kg", "6mg/kg", "6 mg kg", "6mg kg",
"130 mg", "130mg",
"subcutânea 90 mg", "subcutanea 90 mg", "sc 90 mg", "sc90 mg", "90 mg sc", "90mg sc",
"semana 8", "s8", "week 8", "wk8",
"manutenção", "manutencao", "manutenção sc", "manutencao sc",
"q12 semanas", "q12sem", "q12 sem", "q12w", "a cada 12 semanas", "de 12 em 12 semanas",
"q8 semanas", "q8sem", "q8 sem", "q8w", "a cada 8 semanas", "de 8 em 8 semanas",
"interrupção aos 16 semanas", "interrupcao aos 16 semanas", "semana 16", "s16", "week 16", "wk16"
],

"**Estratégias off-label / intensificação**": [
"off-label", "uso off label", "fora do rcm", "fora de bula",
"intensificação de dose", "intensificacao de dose", "intensificação", "intensificacao",
"aumento de frequência", "encurtamento de intervalo",
"q6 semanas", "q6sem", "q6 sem", "q6w", "a cada 6 semanas",
"q4 semanas", "q4sem", "q4 sem", "q4w", "a cada 4 semanas",
"reindução intravenosa", "re-indução intravenosa", "reinducao intravenosa", "re-inducao intravenosa", "reindução iv", "reinduçao iv", "reinducao iv",
"estratégia combinada", "estrategia combinada", "combinação terapêutica", "combinacao terapeutica",
"perda de resposta", "loss of response", "lor",
"resposta inadequada", "inadequate response", "partial response only"
],

"**Ensaios pivotais**": [
"uniti-1", "uniti 1", "uniti1",
"uniti-2", "uniti 2", "uniti2",
"im-uniti", "im uniti", "imuniti"
],

"**Marcadores clínicos e laboratoriais**": [
"hbi", "harvey-bradshaw", "harvey bradshaw", "harvey–bradshaw index", "harvey bradshaw index",
"cdai", "crohn's disease activity index",

"pcr", "proteína c reativa", "proteina c reativa", "proteína c reativa de alta sensibilidade", "crp", "c-reactive protein",
"calprotectina fecal", "calprotectina nas fezes", "calprotectina", "fecal calprotectin", "cal",
"atividade endoscópica", "atividade endoscopica", "atividade endoscópica da mucosa", "atividade endoscopica da mucosa",
"úlceras", "ulceras", "lesões ulceradas", "ulcerations",
"mayo score", "escala de mayo", "mayo clinic score", "partial mayo", "mayo parcial",
"patient reported outcomes", "pros", "relatos de doente", "resultados reportados pelo doente",
"retorragias", "sangue retal", "rectorragias", "sangramento retal",
"tenesmo", "tenesmus",
"dor abdominal", "dor abdominais", "abdominal pain",
"diarreia", "diarréia", "diarrhea",
"sinais de doença ativa em imagem", "imagem compatível com atividade", "achados de atividade em imagem"
],

"**Tratamentos prévios**": [
"falência anti-tnf", "falencia anti-tnf", "falha anti-tnf", "loss of response to anti-tnf",
"intolerância anti-tnf", "intolerancia anti-tnf",
"anti-tnf", "anti tnf", "anti-tnfα", "anti-tnfa",
"corticosteroides", "corticosteroides sistémicos", "corticosteroides sistemicos", "cct", "esteroides", "esteroides sistêmicos", "esteroides sistemicos",
"imunomoduladores", "imunossupressores",
"aminosalicilatos", "5-asa", "mesalazina", "mesalamina",
"antibióticos", "antibioticos"
],

"**Outcomes**": [
"resposta clínica", "resposta clinica", "clinical response",
"remissão clínica", "remissao clinica", "clinical remission",
"remissão endoscópica", "remissao endoscopica", "endoscopic remission",
"resposta endoscópica", "resposta endoscopica", "endoscopic response",
"cicatrização da mucosa", "cicatrizacao da mucosa", "mucosal healing",
"remissão sem corticosteroides", "remissao sem corticosteroides", "steroid-free remission",
"remissão sem cct", "remissao sem cct", "cct-free remission"
],

"**Estatísticas**": [
"ic 95%", "intervalo de confiança de 95%", "95% ci", "confidence interval 95%",
"i2", "i^2", "i-squared",
"heterogeneidade", "heterogeneity",

"modelo de efeitos aleatórios", "random-effects", "random effects", "dersimoni an laird", "dl"
],

"**Segurança e contraindicações**": [
"eventos adversos", "eventos adversos graves", "eag", "grau 3-4", "grau ≥3", "grau >=3",
"nasofaringite", "nasofaringites",
"eritema no local da injecção", "eritema no local da injeção", "eritema no local de injecao", "reação no local da injeção", "reacao no local da injecao",
"candidíase", "candidiase", "candidiasis",
"candidíase vulvovaginal", "candidiase vulvovaginal",
"bronquite", "bronchitis",
"prurido", "coceira",
"infeção do trato urinário", "infecção do trato urinário", "infeccao do trato urinario", "itu", "uti", "urinary tract infection",
"sinusite", "sinusitis",
"vómitos", "vomitos", "vomiting",
"contraindicações", "contraindicacoes",
"infeção ativa", "infecção ativa", "infeccao ativa", "active infection",
"tuberculose ativa", "tb ativa", "active tuberculosis"
],

"**Abreviaturas**": [
"ust", "ustk", # incluir variante reforçada
"iv", "intravenosa",
"sc", "subcutanea", "subcutânea",
"q4sem", "q4 sem", "q4w",
"q6sem", "q6 sem", "q6w",
"q8sem", "q8 sem", "q8w",
"q12sem", "q12 sem", "q12w"
]
}

Keywords **Relatório 3:**

"**Fármaco**": [
"burosumab", "crysvita",
"kyowa kirin", "kyowa-kirin",
"anticorpo monoclonal", "anticorpos monoclonais",
"igg1", "igg-1", "igg 1",
"fgf23", "fgf-23", "fgf 23",
"fator de crescimento de fibroblastos 23", "fibroblasto 23", "fibroblast growth factor 23",

```
    "medicamento órfão", "medicamento orfao", "orphan drug",
    "ema", "european medicines agency",
    "fda", "food and drug administration"
  ],

"Doença": [
    "xlh", "xlhr", "x-linked hypophosphatemia", "hipofosfatemia ligada ao x",
    "raquitismo hipofosfatémico", "raquitismo hipofosfatemico",
    "osteomalacia", "osteomalacias",
    "tio", "tumor-induced osteomalacia", "osteomalacia induzida por tumor",
    "tumores mesenquimatosos", "tumor mesenquimatoso"
  ],

"Caso clínico": [
    "phex", "mutação phex", "gene phex",
    "pseudofraturas", "pseudo-fraturas", "pseudofratura", "pseudo-fratura", "looser zones",
    "cintilograma ósseo", "cintilograma osseo", "bone scan", "bone scintigraphy",
    "ecografia renal", "ultrassonografia renal", "ultrassom renal", "renal ultrasound",
    "nefrocalcinose", "nephrocalcinosis"
  ],

"Posologia": [
    "1 mg/kg", "1mg/kg", "1 mg kg",
    "90 mg", "90mg",
    "q4 semanas", "q4 sem", "q4w", "a cada 4 semanas", "de 4 em 4 semanas",
    "jejum", "em jejum", "fasting",
    "descontinuar fosfato oral", "suspender fosfato oral", "parar fosfato oral",
    "análogos ativos da vitamina d", "analogos ativos da vitamina d",
    "calcitriol", "vitamina d ativa", "vitamina d inativa", "inactive vitamin d",
    "fosfato sérico", "fosfato serico", "serum phosphate",
    "diminuição da dose", "diminuicao da dose", "redução da dose", "reducao da dose",
    "suspender dose", "interromper dose",
    "metade da dose", "reduzir para metade da dose",
    "2 semanas após a dose", "2 semanas apos a dose", "duas semanas após a dose",
    "pós-prandial", "pos-prandial", "pós prandial", "pos prandial", "após refeição", "apos
refeicao", "postprandial"
  ],

"Monitorização": [
    "fosfato sérico", "fosfato serico", "serum phosphate",
    "fosfatase alcalina", "alcaline phosphatase", "alp",
    "cálcio", "calcio", "serum calcium",
    "creatinina", "serum creatinine",
```

"pth", "paratormona", "parathyroid hormone",
"fosfato urinário", "fosfato urinario", "urinary phosphate",
"tmp/gfr", "tmp gfr", "trp", "tubular reabsorption of phosphate"
],

"**Endpoints**": [
"patient reported outcomes", "pros", "resultados reportados pelo doente",
"brief pain inventory", "bpi",
"brief fatigue inventory", "bfi",
"womac", "western ontario and mcmaster universities osteoarthritis index",
"six minute walk test", "6mwt", "teste da marcha de 6 minutos", "teste da caminhada de 6 minutos",
"qualidade de vida", "quality of life", "qol",
"fraturas cicatrizadas", "cicatrização de fraturas", "healed fractures"
],

"**Marcadores**": [
"p1np", "procollagen type 1 n-terminal propeptide",
"ctx", "c-terminal telopeptide",
"balp", "bone-specific alkaline phosphatase"
],

"**Reações adversas**": [
"eventos adversos", "eventos adversos graves", "eag",
"lombalgia", "dor lombar", "low back pain",
"reação no local de injeção", "reacao no local da injecao", "local injection site reaction", "reação no local da administração",
"cefaleias", "cefaléia", "headache",
"infeção dentária", "infecção dentária", "infeccao dentaria", "tooth infection", "dental infection",
"diminuição da vitamina d", "diminuicao da vitamina d", "redução da vitamina d", "vitamin d decreased",
"espasmos musculares", "muscle spasms",
"síndrome das pernas inquietas", "sindrome das pernas inquietas", "restless legs syndrome", "rls",
"tonturas", "vertigens", "dizziness",
"obstipação", "obstipacao", "prisão de ventre", "prisao de ventre", "constipação intestinal", "constipacao intestinal", "constipation"
],

"**Contraindicações e precauções**": [
"hipersensibilidade", "hipersensitivity",

"compromisso renal grave", "insuficiência renal grave", "insuficiencia renal grave", "severe renal impairment",
"doença renal terminal", "doenca renal terminal", "end-stage renal disease", "esrd",
"hiperfosfatemia", "hyperphosphatemia",
"mineralização ectópica", "mineralizacao ectopica", "ectopic mineralization", "ectopic calcification"
]
}