Politécnico de Leiria
Escola Superior de Tecnologia e Gestão
Departamento de Engenharia Informática
Mestrado em Ciência de Dados

APLICAÇÃO DE RAG EM MODELOS LLM COM BASES DE DADOS VETORIAIS

RUBEN ALEXANDRE DIAS MARQUES

Leiria, Novembro de 2024

Politécnico de Leiria
Escola Superior de Tecnologia e Gestão
Departamento de Engenharia Informática
Mestrado em Ciência de Dados

APLICAÇÃO DE RAG EM MODELOS LLM COM BASES DE DADOS VETORIAIS

RUBEN ALEXANDRE DIAS MARQUES
Número: 2220128

Projeto realizado sob orientação do Professor Doutor Ricardo Malheiro (ricardo.malheiro@ipleiria.pt) e Professora Doutora Maria Beatriz Piedade (beatriz.piedade@ipleiria.pt).

Leiria, Novembro de 2024

AGRADECIMENTOS

A realização deste projeto contou com apoios importantes sem os quais não se teria tornado uma realidade e aos quais estarei eternamente grato.

Ao Doutor Ricardo Malheiro, orientador deste projeto, expresso o meu profundo agradecimento pela sua incansável disponibilidade e paciência, proporcionando-me as condições necessárias para o desenvolvimento deste trabalho.

À Professora Maria Beatriz Piedade, agradeço pela orientação na fase inicial deste projeto e pela disponibilidade e interesse demonstrado no acompanhamento da sua elaboração.

Ao Doutor Bruno Miguel Lopes e Silva, o meu sincero obrigado por todo o apoio ao longo deste processo, pelo exemplo de excelência e pelos valiosos momentos de reflexão que contribuíram significativamente para todo o percurso.

Aos meus pais, expresso a minha eterna gratidão por sempre terem feito o possível para me tornarem numa pessoa de valores e princípios. O vosso amor e apoio incondicional foram a base de todo o meu percurso.

Ao João Ramos, um agradecimento especial por sempre ter acreditado em mim e por todo o companheirismo ao longo deste percurso. A tua presença e apoio constantes foram essenciais para superar alguns dos desafios encontrados.

A todos os que, de alguma forma, contribuíram para a realização deste trabalho, o meu sincero agradecimento.

RESUMO

A Geração Aumentada por Recuperação (RAG) é uma estrutura que permite aos Modelos de Linguagem de Grande Escala (LLMs) melhorar a precisão e a relevância de respostas dos modelos, através de integração de bases de conhecimento externas.

Neste trabalho, apresenta-se a implementação de um sistema RAG integrado com LLMs e bases de dados vetoriais (VecDBS) de forma a otimizar a utilização de Inteligência Artificial Generativa em áreas complexas do ponto de vista de conhecimento técnico, como a certificação energética em Portugal.

Realizou-se extração de conhecimento através do manual SCE da ADENE, entidade reguladora, e construiu-se a estratégia do sistema RAG integrado com LLMs, implementado o modelo Gemma 7B e a base de dados vetorial ChromaDB, dando acesso aos profissionais da área de terem informações relativamente a processos, cálculos e elementos legislativos, de forma muito mais eficiente, eliminando o tempo de pesquisa associada a este processo de certificação.

Avaliou-se o projeto através de uma análise comparativa entre o sistema RAG e os métodos tradicionais, focando na precisão, relevância e clareza das respostas geradas. As metodologias de avaliação empíricas demonstram que o sistema melhora significativamente as capacidades de resposta a este tema complexo, dando mais clareza, integridade e relevância na informação gerada e aumentando a eficiência dos profissionais da área.

Simultaneamente, os resultados demonstraram ainda uma redução de 92,5% nos custos para preparação e utilização do sistema, em comparação com as abordagens de *fine-tuning* tradicionais, e uma melhoria consistente na precisão e relevância das respostas, reduzindo ainda os custos associados, sendo eles financeiros, computacionais e temporais.

Termos-chave: Geração Aumentada por Recuperação, Modelo de Linguagem de Grande Escala, Inteligência Artificial Generativa, Bases de Dados Vetoriais.

ABSTRACT

Retrieval-Augmented Generation (RAG) is a framework that allows Large Language Models (LLMs) to improve the accuracy and relevance of their responses by integrating external knowledge bases.

This work presents the implementation of a RAG system integrated with LLMs and vector databases (VecDBs) to optimize the use of Generative Artificial Intelligence in areas with complex technical knowledge, such as energy certification in Portugal.

Knowledge extraction was performed using the SCE manual from ADENE, the regulatory entity, and a strategy for the RAG system integrated with LLMs was constructed. The system was implemented using the Gemma 7B model and the ChromaDB vector database, providing professionals in the field with much more efficient access to information regarding processes, calculations, and legislative elements, eliminating the research time associated with this certification process.

The project was evaluated through a comparative analysis between the RAG system and traditional methods, focusing on the accuracy, relevance, and clarity of the generated responses. Empirical evaluation methodologies demonstrate that the system significantly improves the response capabilities to this complex topic, providing more clarity, completeness, and relevance in the generated information and increasing the efficiency of professionals in the field.

Simultaneously, the results also demonstrated a 92.5% reduction in costs for system preparation and usage, compared to traditional fine-tuning approaches, and a consistent improvement in the accuracy and relevance of the responses, further reducing associated costs, both financial and in terms of time.

Keywords: Retrieval-Augmented Generation, Large Language Model, Generative Artificial Intelligence, Vector Databases.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

BERT	Bidirectional Encoder Representations from Transformers.
BGE	BAAI General Embeddings.
CNN	Convolutional Neural Networks.
FAISS	Facebook AI Similarity Search.
GPT	Generative Pre-trained Transformer.
HITL	Human-in-the-Loop.
HNSW	Hierarchical Navigable Small World.
IA	Inteligência Artificial.
KL	Kullback-Leibler.
LLM	Modelos de Linguagem de Grande Escala.
LSH	Locality-Sensitive Hashing.
LSTM	Long Short-Term Memory.
LTR	Learning to Rank.
PLN	Processamento de Linguagem Natural.
RAG	Geração Aumentada por Recuperação.
RPH	Taxa de Renovação do Ar.

SCE	Sistema de Certificação Energética de Edifícios.
SHAP	SHapley Additive exPlanations.
TIC	Tecnologias de Informação e Comunicação.
U	Coeficiente de Transmissão Térmica.
VecDB	Base de Dados Vetorial.
VecDBs	Bases de Dados Vetoriais.
WMD	Word Mover's Distance.
XAI	Explainable AI.

1

INTRODUÇÃO

O desenvolvimento e a democratização da Inteligência Artificial Generativa, especialmente no que diz respeito aos Modelos de Linguagem de Grande Escala (LLM), representam um marco significativo na evolução tecnológica dos últimos cinco anos. Segundo dados publicados pelo Eurostat em 2021, 24% das empresas já utilizavam software ou sistemas de Inteligência Artificial (IA) para a segurança das Tecnologias de Informação e Comunicação (TIC), por exemplo, através de técnicas de aprendizagem automática para a deteção e prevenção de ciberataques. Similarmente, 23% das empresas aplicaram estas tecnologias na automação de processos administrativos, como planeamento e assistência virtual em negócios. (Eurostat, 2022)

Contudo, o uso de software ou sistemas de Inteligência Artificial (IA) para a gestão de recursos humanos ou recrutamento foi o menos implementado, com apenas 8% das empresas a reportar a utilização desta ferramenta. Este dado sugere uma hesitação ou possíveis barreiras na implementação de IA para funções que envolvam a necessidade de julgamento humano.

Importa destacar que a aplicação de IA nas empresas varia significativamente com o tamanho da organização. Por exemplo, 39% das grandes empresas usaram IA para segurança das Tecnologias de Informação e Comunicação (TIC), comparativamente a apenas 20% das pequenas empresas. Esta tendência repete-se na produção, com 33% das grandes empresas e 17% das pequenas, e na logística, com 18% das grandes contra 8% das pequenas. (Eurostat, 2022).

Na realidade, apesar do sucesso notável dos LLM, esta tecnologia enfrenta limitações consideráveis, especialmente em tarefas que requerem conhecimento específico ou são abrangentes em conhecimento. Entre os principais problemas transversais a modelos tanto rudimentares quanto avançados, como o Generative Pre-trained Transformer (GPT), nomeadamente o GPT-3, GPT-4 [1], ou o mais recente desenvolvimento da *Google* à data deste artigo, o *Gemma 7b*, destacam-se a geração de "alucinações"— respostas inventadas ou factualmente incorretas — e a obsolescência

1 Desenvolvidos pela OpenAI, GPT-3 e GPT-4 são modelos de linguagem que utilizam milhões de parâmetros para gerar texto de forma contextualizada. Detalhes adicionais estão disponíveis no site oficial da *OpenAI*: https://openai.com/research/gpt-3 e https://openai.com/research/gpt-4.

acelerada devido à rápida evolução tecnológica. Estes desafios acontecem devido à incapacidade dos modelos terem acesso informações fora do seu conjunto de dados de treino, conforme destacado em alguns estudos mais recentes (Jing et al., 2024).

Adicionalmente, a exigência de um elevado poder computacional para treinar LLM, cujos modelos possuem biliões de parâmetros, envolve custos significativos. Isso reflete as barreiras económicas e técnicas que ainda precisam de ser superadas para permitir uma aplicação destes modelos em ambientes de produção. Estas questões sublinham a necessidade de avanços contínuos na tecnologia de LLM para garantir a eficácia, a precisão e a viabilidade económica da tecnologia (Alizadeh et al., 2024).

A indústria enfrenta uma crescente necessidade de processamento mais rápido e eficiente, o que destaca a importância de otimizar custos e facilmente adaptar-se a novas informações nos modelos de IA. Com esta exigência, surge da necessidade de manter a competitividade e a eficácia operacional, de forma a garantir o desenvolvimento de tecnologias que facilitem a integração de atualizações contínuas e acelerem o processamento de dados. Assim sendo, as soluções de IA adotadas além de ágeis e escaláveis, devem também permitir ajustes rápidos com capacidade de responder a novas perguntas com informações atualizadas. A implementação destes sistemas pode resultar em vantagens significativa, como a simplificação de temas complexos e com grande abrangência de conhecimento. (Yan et al., 2023)

Estes desafios enfrentados pelos LLM tornam-se particularmente críticos quando são aplicados a domínios associados a alta regulamentação, como é o caso da certificação energética em Portugal, que é o tema motivador central descrito na secção de *Motivação*, do presente Capítulo deste documento.

No contexto da certificação energética é importante garantir a precisão e atualização das informações, uma vez que afetam diretamente decisões importantes relacionadas com a eficiência energética de edifícios e estão associadas ao cumprimento de normas regulatórias que podem sofrer alterações. Neste cenário, a aplicação de LLM poderia oferecer uma solução de suporte de valor acrescentado, através do auxílio na interpretação de normas, na realização de cálculos e na geração de relatórios detalhados após o processo de certificação. No entanto, as limitações atuais dos LLM, como a dificuldade a incorporar rapidamente novas informações, representam riscos significativos para este campo de conhecimento. Pretendem-se ultrapassar estes desafios no desenvolvimento do presente projeto através da implementação de um sistema de Geração Aumentada por Recuperação (RAG). É importante destacar que um erro na interpretação de uma norma ou a utilização

de dados desatualizados pode resultar em certificações incorretas, com potenciais consequências do ponto de vista legal e financeiro para as empresas que utilizem a tecnologia, e ainda, para os próprios proprietários dos imóveis.

1.1 MOTIVAÇÃO

No contexto da certificação energética em Portugal, os profissionais enfrentam um desafio constante para se manterem atualizados com base num abrangente conjunto de normas e regulamentos, que estão sujeitos a alterações e evoluções com alguma frequência. Este processo torna-se ainda mais desafiante quando se considera a necessidade de garantir que as avaliações realizadas sejam sempre precisas e em conformidade com a legislação.

De um ponto de vista objetivo, esta realidade revela a necessidade de existir uma solução que facilite o trabalho dos profissionais, permitindo-lhes aceder rapidamente a informação atualizada e garantindo que as todas as decisões tomadas se baseiem sempre nos dados mais recentes. A ineficiência do processo atual, onde os profissionais precisam de investir um tempo significativo na consulta de documentos extensos e na interpretação de normas, pode levar a erros de interpretação e até originar atrasos, apresentando um forte impacto na qualidade das certificações emitidas.

É precisamente nesta intersecção entre a necessidade do setor e os avanços tecnológicos recentes que surge a motivação para o presente projeto. O desenvolvimento de tecnologias no campo do PLN, aliado a métodos inovadores como a RAG e VecDBs, possibilita a criação de soluções capazes de responder de forma eficaz aos desafios identificados.

A integração de RAG com VecDBs e LLM permite a construção de um sistema de Inteligência Artificial Generativa que responde a questões complexas de forma contextualizada, e que ainda se atualiza de acordo com as mudanças dos documentos normativos. Com esta abordagem pretende-se transformar a forma como os profissionais interagem com a informação normativa, reduzindo grande parte do tempo gasto na pesquisa e interpretação de documentos.

A aplicação prática deste projeto não se limita à inovação tecnológica, mas abrange também a parte operacional da engenharia, uma vez que se traduz em benefícios concretos para os profissionais da certificação energética. A ferramenta proposta tem o potencial de revolucionar o setor, ao permitir que os profissionais

foquem na análise crítica e na tomada de decisões, em vez de perderem tempo em tarefas repetitivas.

Por fim, este projeto alinha-se com os objetivos de modernização e digitalização do país, contribuindo para que Portugal se posicione na vanguarda da aplicação de tecnologias disruptivas no setor da sustentabilidade e eficiência energética.

Ao criar uma ferramenta que responde às necessidades específicas do setor, ao mesmo tempo que capitaliza os mais recentes avanços em IA, este projeto resolve desafios técnicos e ainda estabelece um novo padrão para a interação com informação normativa complexa, com potencial de replicação em outras áreas de conhecimento.

Em suma, a motivação para este trabalho surge da conjugação entre uma necessidade real e urgente no setor da certificação energética, os avanços tecnológicos disponíveis, e a oportunidade de contribuir para os objetivos de inovação a nível nacional.

1.2 OBJETIVOS DO PROJETO

Este projeto tem como objetivo principal desenvolver e avaliar um sistema de RAG integrado com LLM e VecDBs, tendo como foco principal a aplicação desta tecnologia no domínio da certificação energética em Portugal. Para tal, propõe-se a cumprir os seguintes pontos:

- Implementar um sistema RAG-LLM adaptado ao contexto da certificação energética em Portugal. Este sistema integrará o modelo Gemma 7B com a Base de Dados Vetorial (VecDB) *ChromaDB*[2], utilizando como fonte primária de informação o Manual Sistema de Certificação Energética de Edifícios (SCE) da ADENE. O sistema será capaz de processar perguntas em português relacionadas com regulamentações e procedimentos de certificação energética, fornecendo respostas precisas e contextualizadas.
- Desenvolver um mecanismo de atualização contínua do conhecimento do sistema. Este mecanismo permitirá a incorporação regular de novas informações no domínio da certificação energética, sem necessidade de treinar o modelo completo. O objetivo é manter o sistema atualizado com normas e práticas mais recentes definidas pela entidade reguladora como adequadas ao processo de certificação.

2 ChromaDB é uma base de dados vetorial utilizada no projeto para armazenar e recuperar informações específicas. Informações adicionais sobre ChromaDB podem ser consultadas no site oficial: https://chromadb.com.

- Realizar uma avaliação comparativa do desempenho do sistema RAG-LLM em relação às abordagens tradicionais de acesso à informação no contexto da certificação energética. Esta avaliação incluirá métricas quantitativas de precisão, relevância e tempo de resposta, bem como uma análise qualitativa da utilidade e aplicabilidade das respostas através de perguntas reais no cenário da certificação energética.
- Quantificar o impacto do sistema na eficiência do trabalho dos profissionais de certificação energética. Isto envolverá a medição da redução no tempo necessário para a disponibilidade tecnológica para os profissionais, bem como a avaliação da melhoria na precisão e consistência das interpretações regulamentares. O objetivo é demonstrar uma redução significativa no tempo de processamento e um aumento na precisão das avaliações.
- Elaborar uma breve análise de custo-benefício da implementação do sistema em empresas e instituições do setor de certificação energética. Esta análise considerará os custos de desenvolvimento e manutenção do sistema, bem como os potenciais benefícios em termos de eficiência operacional, precisão das certificações e conformidade regulatória. O objetivo é fornecer uma base sólida para a tomada de decisões sobre a adoção desta tecnologia no setor.
- Desenvolver e avaliar um protótipo de interface de utilizador que permita aos profissionais do setor interagir eficientemente com o sistema RAG-LLM. Esta interface será projetada considerando as necessidades específicas dos certificadores energéticos, facilitando a formulação de consultas complexas e a interpretação das respostas geradas pelo sistema.

De forma a guiar o projeto e a documentar este conjunto de objetivos de forma eficaz, o estudo propõe-se a responder às seguintes questões-chave:

- Como é que a integração de RAG com VecDBs melhora a precisão, relevância e atualidade das respostas fornecidas por LLM no contexto específico da certificação energética em Portugal?
- De que forma o sistema RAG-LLM proposto aumenta a agilidade e adaptabilidade dos profissionais do setor face às constantes atualizações regulatórias e técnicas?
- Quais são os desafios técnicos, éticos e práticos na implementação e manutenção de um sistema RAG-LLM para uso em ambientes profissionais reais, e como podem ser superados?

- Como pode este sistema servir de modelo para a aplicação de tecnologias de IA em outros domínios regulatórios complexos em Portugal?

Ao abordar estes objetivos e questões, este trabalho tem como objetivo não apenas contribuir para o avanço científico na área de inteligência artificial aplicada ao domínio da engenharia, mas também oferecer uma solução prática para otimizar os processos e simplificar os desafios enfrentados pelos profissionais do setor de certificação energética em Portugal.

O projeto pretende estabelecer um precedente para a aplicação responsável e eficaz da tecnologia de Inteligência Artificial Generativa em domínios altamente regulamentados da engenharia, potencialmente influenciando futuras políticas e práticas no setor.

1.3 ESTRUTURA DO DOCUMENTO

O presente documento está organizado em seis capítulos principais, cada um com foco assente nos aspetos específicos do projeto de aplicação de RAG em modelos LLM com Bases de Dados Vetoriais no contexto da certificação energética em Portugal.

O Capítulo 1, Introdução, apresenta o contexto geral do projeto, destacando a relevância da Inteligência Artificial Generativa e dos LLM no cenário tecnológico, seguindo-se da motivação para o projeto, os objetivos específicos propostos e uma breve visão geral da estrutura do documento.

O segundo Capítulo, Estado da Arte, oferece uma revisão abrangente das tecnologias e conceitos fundamentais para o projeto. Aborda-se o panorama geral dos LLM, o conceito de RAG, a importância das VecDBs, e detalha-se ainda os pilares fundamentais associados ao RAG. Além disso, este Capítulo explora em detalhe o modelo Gemma 7B e as métricas de análise e avaliação utilizadas no projeto.

O Capítulo 3, Metodologia e Arquitetura, detalha a abordagem técnica adotada no projeto. Apresenta-se uma visão geral do sistema, seguida pela descrição detalhada da arquitetura proposta, incluindo os componentes principais e o fluxo de operação. Este Capítulo também aborda a metodologia de implementação, a análise e utilização dos dados, bem como considerações sobre desempenho e otimização.

O Capítulo 4, apresenta e discute os resultados obtidos através da implementação do sistema proposto. Inclui-se uma análise comparativa detalhada entre o sistema RAG e métodos tradicionais, bem como uma avaliação da eficiência computacional

e custos associados. Este capítulo aborda ainda os testes empíricos realizados e as métricas utilizadas para avaliar o desempenho do sistema.

O Capítulo 5, Conclusões, sintetiza os principais resultados e descobertas do projeto. Reflete-se sobre o cumprimento dos objetivos iniciais, as implicações dos resultados obtidos e as contribuições do projeto no processo de certificação energética.

O Capítulo 6, Trabalho Futuro, explora as possíveis direções para expansão e aprimoramento do projeto. São discutidas potenciais melhorias na eficiência do sistema, propostas para o desenvolvimento de interfaces mais intuitivas e considerações sobre questões éticas e de transparência na aplicação da tecnologia.

Esta estrutura foi concebida para proporcionar uma progressão lógica da evolução do projeto, ajudando a correta compreensão desde todo os fundamentos teóricos que suportam os desenvolvimentos até às aplicações práticas e discussão dos resultados obtidos, terminando com uma reflexão sobre as implicações e possibilidades futuras do trabalho realizado.

2

ESTADO DA ARTE

A inteligência artificial tem sofrido avanços sem precedentes nas últimas décadas, criando um impacto determinante em vários setores da sociedade e da economia. Com o fenómeno da democratização da Inteligência Artificial Generativa, particularmente dos LLM, registou-se um ponto de inflexão na tecnologia ao longo dos últimos 5 anos. Esta evolução deve-se principalmente à pesquisa e desenvolvimento da inovação, que tem levado ao desenvolvimento de soluções cada vez mais disruptivas.

Um dos marcos mais significativos deste campo tecnológico é o aparecimento e a rápida evolução dos LLM. A tecnologia, exemplificada por modelos como GPT, Bidirectional Encoder Representations from Transformers (BERT), e mais recentemente, o *Gemma*, têm demonstrado capacidades únicas para compreensão e criação de texto, aproximando-se em muitos aspetos do desempenho do ser humano. Estes modelos têm sido soluções disruptivas em casos como a tradução automática, a análise de sentimentos e até na geração de conteúdo.

Dada a centralidade dos LLM no avanço recente da IA e a relevância centrada no presente projeto, é necessário que se examine em detalhe a trajetória evolutiva destes modelos, bem como os desafios que atualmente enfrentam. Com esta análise pretende-se contextualizar o estado atual da tecnologia e fundamentar a necessidade de abordagens inovadoras como a tecnologia de Geração Aumentada por Recuperação (RAG), que será explorado nas seções seguintes.

2.1 PANORAMA GERAL DOS LLM E A NECESSIDADE DE EVOLUÇÃO

Os LLM têm revolucionado o campo da IA através de capacidades nunca antes atingidas no que se relaciona à geração de texto, compreensão de linguagem natural e interação humano-computador. Estes modelos, com base tecnológica nas redes neurais treinadas com grande corpus de texto, compreendem, preveem e geram respostas linguísticas com um nível de sofisticação que se aproxima, e por vezes supera, a performance humana (Naveed et al., 2024).

Historicamente, os LLM evoluíram de métodos estatísticos simples e baseados em regras para os sistemas mais complexos que se fundamentaram em técnicas avançadas de estatística e aumento da capacidade computacional. Esta transformação registou o primeiro marco significativo com o desenvolvimento dos modelos *Transformer* em 2017, altura em que se tornaram fundamentais para os LLM devido ao mecanismo de atenção eficaz que permite focar diferentes partes de um texto durante o processamento da linguagem (Vaswani et al., 2023). Dada a importância desta arquitetura, é abordada na secção seguinte com maior detalhe.

2.2 ARQUITETURA TRANSFORMERS

A arquitetura Transformer, introduzida por Vaswani et al., 2023, é o modelo que deu base aos LLM. A arquitetura distingue-se das abordagens anteriores, como as redes Long Short-Term Memory (LSTM) e Convolutional Neural Networks (CNN), por ter eliminado a necessidade de mecanismos recorrentes e convolucionais. Por exemplo, os mecanismos recorrentes, utilizados em LSTM, processam sequências de dados passo a passo, e utilizam a saída de um passo como entrada para o próximo, o que pode limitar a eficiência e dificultar a captura de dependências de longo alcance. Já os mecanismos convolucionais, aplicam filtros nas sequências de dados para extrair características, mas também apresentam limitações na captura de dependências durante a sequência de passos. A eliminação destas dependências permitiu aos *Transformers* melhorar a eficiência e a capacidade de capturar dependências de longo alcance em sequências de texto.

Para esta tecnologia funcionar, é utilizado um mecanismo de atenção. Este mecanismo permite ao modelo atribuir diferentes pesos aos *tokens* de entrada, com base na relevância para a tarefa em questão. A arquitetura inclui múltiplas cabeças de atenção, e cada uma processa a sequência de forma independente, mas em paralelo. Isto faz com que o modelo detete as relações contextuais em simultâneo, e melhora a precisão na interpretação do modelo.

A estrutura do *Transformer* baseia-se num modelo codificador-descodificador. O codificador processa a sequência de entrada e gera uma representação interna que encapsula a informação contextual. O descodificador, por sua vez, utiliza esta representação para produzir a saída, que pode ser, por exemplo, uma sequência de texto traduzido ou um resumo. Cada camada do codificador e do descodificador integra uma subcamada de atenção e uma rede *feed-forward*, que normaliza o processo de treino, e que facilitam a propagação dos gradientes.

Uma característica importante dos *Transformers* é a capacidade de lidar com a ordem das palavras numa sequência. Isto é conseguido através de codificações posicionais, e é extremamente importante para o projeto desenvolvido, uma vez que o foco é precisamente a utilização de sequências que formam o documento.

Como o *Transformer* processa todos os *tokens* em paralelo, é necessário preservar a ordem sequencial. Para manter a coerência do contexto, são utilizadas então as codificações posicionais - vetores adicionados às representações dos *tokens* e que permitem ao modelo captar a ordem relativa dos mesmos.

Após o desenvolvimento da arquitetura *Transformer*, surgiram modelos como o GPT e o BERT, que contribuíram bastante para o estado da arte em Processamento de Linguagem Natural (PLN). Estes modelos, treinados em conjuntos de dados massivos e assentes em fortes recursos computacionais, demonstraram capacidades quase humanas numa variedade de tarefas linguísticas (Devlin et al., 2019a). Por exemplo, o GPT-3, com 175 biliões de parâmetros, demonstrou a possibilidade de gerar texto coerente e contextualmente relevante numa vasta gama de estilos e tópicos. (Radford et al., 2019)

A sofisticação destes modelos abriu caminho para várias aplicações práticas - como assistentes virtuais com capacidade para manter conversas fluídas (Tyen et al., 2022) até sistemas de recomendação que compreendem variações contextuais (Shah et al., 2024), os LLM têm sido disruptivos na transformação da maior parte dos setores. Na análise de sentimentos, por exemplo, modelos como o *RoBERTa* demonstraram precisão superior a 95% em *benchmarks* de análise (Q. Wu et al., 2023) - método comparativo para modelos de IA.

No entanto, apesar dos avanços impressionantes, os LLM enfrentam desafios significativos. Um dos mais salientes é a necessidade de manter o conhecimento atualizado em áreas dinâmicas. A natureza estática do treino inicial dos LLM significa que podem rapidamente tornar-se desatualizados, um problema grave em domínios com possibilidade de alterações, como é o caso da engenharia (Gekhman et al., 2024).

A necessidade de atualizar LLM sem submetê-los a ciclos de treino intensivos e dispendiosos apresenta-se assim como um dos maiores desafios da tecnologia. Existe alguma literatura recente que estima que o treino de um LLM pode custar milhões de dólares e produzir uma quantidade significativa de emissões de carbono (Strubell et al., 2019). Este desafio levanta claramente implicações económicas e ambientais, e afeta diretamente a relevância e clareza das respostas dadas pelos modelos, colocando em causa a integridade dos resultados do modelo.

Ao observar com detalhe estes desafios significativos enfrentados pelos LLM, torna-se evidente a necessidade de soluções disruptivas que possam manter a relevância e clareza destes modelos sem sofrer custos proibitivos de retreino constante. É neste contexto que o RAG aparece como uma solução disruptiva. Ao combinar a capacidade generativa dos LLM com a flexibilidade de acesso a informações externas e atualizadas, o RAG representa um paradigma que tem a capacidade de mudar a forma como os sistemas de Inteligência Artificial Generativa lidam com conhecimento em constante evolução. Esta tecnologia além de mitigar os problemas de desatualização dos LLM, também abre novas possibilidades para aplicações em domínios que exigem informações atuais.

2.3 GERAÇÃO AUMENTADA POR RECUPERAÇÃO (RAG)

A metodologia de RAG representa uma inovação significativa na operação dos LLM, visto que integra informações atualizadas no processo de funcionamento do modelo, sem a necessidade de se submeter a um novo processo iterativo de treino.

Inicialmente proposta por *Patrick Lewis* na literatura cientifica *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, a abordagem RAG evoluiu rapidamente e tornou-se uma metodologia usada para melhorar a precisão e atualidade das respostas dos LLM (Lewis et al., 2021).

O RAG difere fundamentalmente do método tradicional de *fine-tuning* ao introduzir um sistema intermédio de pesquisa a bases de dados externas, frequentemente implementadas como Bases de Dados Vetoriais (VecDBs). Assim, superam-se as limitações do conhecimento supramencionadas e incorpora-se informações recentes ou específicas do domínio de forma dinâmica (Akkiraju et al., 2024).

A eficácia do RAG tem sido demonstrada em vários domínios que exigem informações atualizadas. Por exemplo, *Gautier Izacard* e Touvron et al., 2023 mostraram que modelos RAG superaram os LLM tradicionais em tarefas de pergunta-resposta de domínio aberto, alcançando uma melhoria de até 30% em métricas padrão como *Exact Match* e *F1-score* (Izacard e Grave, 2021).

Por outro lado, na área da medicina, apesar de ser um tema ainda sensível para a implementação da IA, provou-se que um sistema RAG forneceu recomendações médicas mais acertadas em comparação com LLM convencionais, reduzindo erros de informação em aproximadamente 25%. (Ke et al., 2024).

Apesar dos avanços promissores e da tecnologia disruptiva, a implementação do RAG também enfrenta ainda desafios significativos:

- Latência: A recuperação e integração de informações externas podem provocar atrasos significativos na geração de respostas. No entanto, existem já estudos científicos que propuseram técnicas de indexação avançadas que reduziram a latência em até 40% em comparação com métodos RAG tradicionais (Khattab, Santhanam et al., 2023).
- Relevância da Informação Recuperada: Garantir que as informações recuperadas sejam relevantes de acordo com a pergunta do LLM é necessário para que haja uma utilização sem atritos desta tecnologia. Alguma literatura recente explora técnicas de *re-ranking* semântico que melhoraram a precisão da recuperação em até 15% (Hwang et al., 2024).
- Integração Eficiente: A conjugação eficaz das informações recuperadas da base de dados externa com o conhecimento do modelo base continua a ser um impedimento à implementação. No estudo *Penetrative AI: Making LLM Comprehend the Physical World* propôs-se um mecanismo de atenção adaptativa que demonstrou melhorar a coerência das respostas geradas em 20% (Xu et al., 2024).

A integração de tecnologias complementares, como as VecDBs, com os sistemas RAG aparece para ultrapassar precisamente estes desafios acima descritos. As VecDBs otimizadas para consultas de similaridade rápidas, como o Facebook AI Similarity Search (FAISS) [1] (Johnson et al., 2017) e o Hierarchical Navigable Small World (HNSW)[2], (Malkov e Yashunin, 2018) , têm sido utilizadas na implementação de RAG, de forma a permitir a recuperação de informações com latência reduzida e alta precisão.

Existem ainda alguns estudos recentes focadas nas abordagens híbridas que combinam RAG com outras técnicas de atualização de conhecimento. Por exemplo, no estudo de *Siriwardhana* (Siriwardhana et al., 2023), propôs-se um sistema que integra RAG com aprendizagem contínua, apresentado um resultado de melhoria de 18% na precisão das respostas em comparação com sistemas RAG padrão em domínios de rápida evolução.

À medida que a tecnologia evolui, o foco é o desenvolvimento de sistemas RAG mais eficientes, o que inclui a exploração de técnicas de compressão de conhecimento

1 Para mais informações, é possível aceder à documentação oficial em: https://faiss.ai/
2 Para mais informações, é possível aceder ao repositório disponível em: https://github.com/nmslib/hnswlib

(N. Brown et al., 2023) e o uso de aprendizagem por reforço - em inglês *reinforcement learning* - para melhorar as estratégias de recuperação (Li et al., 2016).

Enquanto o RAG aparece como uma abordagem disruptiva para a atualização dinâmica de conhecimento em LLM, a eficácia depende da capacidade de armazenar e recuperar rapidamente informações relevantes. E é precisamente neste ponto que estas VecDBs aparecem como um componente tecnológico necessário, focado no aumento da eficiência da arquitetura RAG.

Esta tecnologia de VecDBs possibilitam a implementação eficiente do RAG, e resolvem parcialmente alguns dos maiores desafios mencionados anteriormente, como a latência de recuperação e a relevância das informações obtidas.

2.4 VECDBS: A FERRAMENTA FUNDAMENTAL PARA ATUALIZAÇÃO DE CONTEXTO EM LLM

As VecDBs são consideradas neste estudo como tecnologia-chave na implementação do sistema de RAG em LLM. Estas bases de dados foram criadas para armazenar e recuperar eficientemente representações vetoriais de dados, uma característica essencial para o processamento rápido e preciso de informações. (Jing et al., 2024)

O princípio de funcionamento das VecDBs assenta na representação de dados como vetores numéricos num espaço multidimensional. Com esta representação permite que se apliquem técnicas de pesquisa eficientes como a pesquisa por similaridade, implementado métricas como a distância euclidiana ou a similaridade do cosseno, abordadas de seguida neste documento.

Além disso, utilizam-se ainda técnicas de indexação eficientes, como Locality-Sensitive Hashing (LSH) ou HNSW para criar estruturas que tornem as pesquisas mais rápidas em grandes conjuntos de dados, ultrapassando as limitações de métodos de pesquisa tradicionais (Andoni et al., 2015).

A evolução destas bases de dados tem sido exponencial nos últimos anos. Alguns sistemas como é o exemplo do FAISS, desenvolvido pelo *Facebook AI Research*, e o anterior mencionado HNSW, têm sido estudados precisamente pelo ponto de eficiência e escalabilidade, sendo esta uma preocupação dos investigadores nesta área científica (Malkov e Yashunin, 2018).

Em primeiro lugar, estes sistemas, permitem a atualização dinâmica de conhecimento, possibilitando a adição contínua de novas informações sem a necessidade de retreinar os modelos, o que é bastante vantajoso em áreas de atuação como é

o caso deste projeto, onde as regulamentações e tecnologias estão em constante evolução. Além disso, as VecDBs contribuem para uma redução da latência das consultas, diminuindo os tempos de resposta significativamente (Ai et al., 2023), e ultrapassando parcialmente os problemas descritos no inicio deste capitulo. Com esta característica a tecnologia pode funcionar em aplicações em tempo real, onde a rapidez da resposta é tão importante quanto a relevância e integridade da mesma.

A escalabilidade é outro aspeto fundamental das VecDBs. À medida que o volume de dados cresce exponencialmente, a capacidade de lidar eficientemente com grandes conjuntos de informações torna-se cada vez mais necessária para estes sistemas. Ainda assim, a tecnologia das VecDBs apresenta capacidade de escalabilidade, mantendo o mesmo desempenho apesar do aumento significativo do volume de dados, como é representado na literatura científica (Wang et al., 2023).

No entanto também existem limites nesta capacidade de escalabilidade. O principal obstáculo é a "maldição da dimensionalidade"e é um fenômeno que acontece quando o desempenho dos algoritmos diminui em espaços de alta dimensão. Este problema aparece principalmente quando os vetores possuem centenas ou milhares de dimensões, como é comum em representações linguísticas. Este fenómeno afeta claramente a eficácia dos algoritmos de pesquisa e indexação, uma vez que as distâncias entre pontos tendem ser mais uniformes à medida que as dimensões aumentam. Outro desafio técnico é a manutenção de índices no caso de atualizações frequentes, o que pode ter um custo computacional elevado. A complexidade dos algoritmos de indexação cresce com o aumento das dimensões, exigindo um desenvolvimento de técnicas focadas na pesquisa e indexação (Zebari et al., 2020).

Considerados estes desafios associados às VecDBs, torna-se evidente a necessidade de uma compreensão mais detalhada dos mecanismos subjacentes que permitem ao RAG funcionar de forma eficaz. Os pilares fundamentais associados ao RAG - nomeadamente a pesquisa vetorial, a similaridade vetorial e o *re-ranking* - ajudam a encontrar potenciais soluções para mitigar problemas como a "maldição da dimensionalidade"e otimizar o desempenho em atualizações frequentes, que será abordado na secção seguinte.

2.5 PILARES FUNDAMENTAIS ASSOCIADOS AO RAG

A metodologia de RAG fundamenta-se em três componentes técnicos principais: pesquisa vetorial, similaridade vetorial e *re-ranking*. Estes elementos permitem que o sistema supere a simples correspondência de palavras-chave, possibilitando a

identificação do significado semântico das palavras para produzir respostas mais contextualizadas (Manning et al., 2008).

De forma resumida, a pesquisa vetorial utiliza representações numéricas de alta dimensão para codificar o conteúdo semântico dos textos. A similaridade vetorial aplica métricas matemáticas para quantificar a proximidade semântica entre vetores. O *re-ranking* implementa algoritmos adicionais para ajustar os resultados iniciais da pesquisa, dando maior prioridade aos mais relevantes. (Mikolov et al., 2013).

Nas secções seguintes, é destacado com o detalhe necessário para este projeto, como funciona cada um destas metodologias.

2.5.1 *Pesquisa Vetorial*

A pesquisa vetorial constitui a base fundamental do RAG, através da utilização de representações densas de textos para capturar contextos semânticos e relacionais. Neste processo, as palavras, frases ou documentos inteiros são transformados em vetores num espaço semântico continuo através de modelos de *embedding*.

Inicialmente, os documentos são processados através de modelos de *embedding*, como o *Word2Vec* (Mikolov et al., 2013) ou o BERT (Devlin et al., 2019b), que transformam o assim os texto em vetores. Estes vetores são então armazenados em VecDBs. Após a execução de uma pergunta, também esta é transformada num vetor através do mesmo modelo de *embedding*. De seguida, o sistema então procura na VecDB os vetores de documentos mais próximos ao vetor da pergunta.

Os avanços nesta área, como o desenvolvimento de modelos de linguagem contextuais através do BERT, como mencionado anteriormente, e têm demonstrado melhorias significativas na captura de variações semânticas. Por exemplo, *Reimers* e *Gurevych* propuseram o *Sentence-BERT*, uma modificação da arquitetura BERT que gera *embeddings* de frases semanticamente significativos, e permite assim uma comparação mais eficiente de frases em contextos larga escala (Reimers e Gurevych, 2019).

2.5.2 *Similaridade Vetorial*

A similaridade vetorial é o mecanismo pelo qual o sistema RAG determina a relevância de um documento para uma determinada consulta.

Após a transformação de consultas e documentos em vetores, a similaridade entre eles é calculada utilizando métricas de similaridade. As métricas mais comuns são:

- Similaridade do Cosseno: Esta métrica baseia-se no ângulo entre os vetores, independentemente da sua magnitude. É particularmente útil em contextos de processamento de texto, pois considera a orientação dos vetores mais do que o tamanho do próprio vetor, sendo ideal para comparar textos de diferentes comprimentos (Singhal, 2001). A fórmula é dada por:

$$\text{Similaridade do Cosseno} \quad \frac{\mathbf{A} \cdot \mathbf{B}}{|\mathbf{A}||\mathbf{B}|} \tag{1}$$

- Distância Euclidiana: Esta métrica mede a distância linear entre dois pontos no espaço vetorial. Embora simples, pode ser influenciada pela magnitude dos vetores, o que pode ser um fator limitante em alguns casos. A fórmula é:

$$\text{Distância Euclidiana} \quad \sqrt{\sum_{i1}^{n} A_i - B_i{}^2} \tag{2}$$

- Produto Escalar: Também conhecido como produto interno, esta métrica soma os produtos dos elementos correspondentes dos vetores. É sensível tanto à orientação quanto à magnitude dos vetores. A fórmula é:

$$\text{Produto Escalar} \quad \sum_{i1}^{n} A_i \times B_i \tag{3}$$

É importante salientar que a similaridade vetorial é indispensável para a precisão do RAG. Estes cálculos de similaridade precisam de ser eficientes para manter a latência baixa.

2.5.3 *Re-Ranking*

O *re-ranking* é o processo final de ordenação dos resultados da pesquisa vetorial. Este processo refina os resultados iniciais da pesquisa vetorial, conseguindo melhorar significativamente a precisão e relevância das informações fornecidas ao LLM.

Após a recuperação inicial baseada em similaridade vetorial, o *re-ranking* aplica modelos mais complexos para reavaliar e reordenar os documentos recuperados. Por sua vez, estes modelos podem considerar fatores adicionais além da simples similaridade vetorial, como a estrutura do documento, a relevância contextual e até mesmo feedback implícito do utilizador.

Em relação ao *re-ranking*, existem as técnicas de aprendizagem para - em inglês Learning to Rank (LTR). Em 2009, *Liu* propôs uma classificação para as abordagens LTR, categorizando-as em *pointwise*, *pairwise* e *listwise*. A técnica *pointwise* avalia a relevância de cada documento individualmente, tratando cada caso como uma tarefa de regressão ou classificação e atribui um valor de relevância a cada documento. A abordagem *pairwise* compara dois documentos de cada vez, com o objetivo de aprender qual dos dois deve ser classificado mais alto em relação a uma pesquisa específica feita pelo utilizador. Por último, a técnica *listwise* considera o conjunto completo de documentos e ordena a lista de forma a otimizar a relevância dos resultados apresentados. Esta última abordagem, *listwise*, tem atraído mais atenção na pesquisa recente devido à sua capacidade de capturar interações complexas entre múltiplos documentos, resultando em uma ordenação mais precisa e coerente dos resultados (Liu, 2009).

No entanto, tem-se dado mais atenção à técnica *listwise*, devido à capacidade de capturar relações complexas entre documentos, tendo em consideração o conjunto de resultados como um todo coerente. A evolução do *re-ranking* foi impulsionada pela investigação de modelos neurais profundos - em inglês *deep learning models* - que trouxeram uma perceção semântica muito mais completa para o processo do que existia antes desta literatura. A introdução da arquitetura *Transformer*, mencionada acima, também ajudou a potenciar esta vertente redefinindo o estado da arte no processamento de linguagem natural (Vaswani et al., 2023).

Esta evolução levou ao desenvolvimento de técnicas que modelam explicitamente a interação entre a pergunta e o próprio documento. Neste contexto, *Khattab* e *Zaharia*, em 2020, introduziram o *ColBERT*, um modelo disruptivo que utiliza uma abordagem que chamaram de atenção cruzada tardia. No contexto do ColBERT, a "atenção cruzada tardia"significa que as interações entre termos da pergunta e do documento são capturadas numa fase posterior do processo de recuperação da informação. Em vez de calcular todas as interações possíveis entre cada termo da pergunta e cada termo do documento de forma antecipada (como é comum nos outros modelos de atenção), ColBERT deteta estas interações com maior eficiência depois de serem geradas as representações vetoriais. Khattab e Zaharia, 2020)

É importante destacar nesta secção que o estado da arte em *re-ranking* continua a evoluir rapidamente, e estas são as inovações que se destacaram à data de investigação e realização do projeto atual, havendo espaço para, à data de leitura, a tecnologia estar numa fase que pode tornar obsoleto ou injustificável esta abordagem.

2.6 OS MODELOS DE LINGUAGEM DE GRANDE ESCALA EM DESTAQUE

Neste campo disruptivo da IA, é sabido que os LLM ocuparam um dos papeis principais na democratização da inteligência artificial ao consumidor comum. A literatura considera estes modelos como a personagem principal dos avanços significativos, desde a tradução automática até aos assistentes virtuais.

A escalabilidade e o desempenho dos LLM têm aumentado exponencialmente, como demonstrado por *Brown*, com o GPT-3, que estabeleceu novos padrões em termos de tamanho do modelo e capacidade de generalização. Esta tendência de crescimento tem sido acompanhada pelos investigadores, com o objetivo de melhorar a eficiência e a aplicabilidade destes modelos em tarefas específicas, devido ao potencial percebido da área em estudo (T. B. Brown et al., 2020).

Esta progressão tecnológica, fundamentada assim na arquitetura dos transformadores, tem assistido a aumento exponencial na escala, trazendo também novos níveis de complexidade dos modelos.

A evolução desta arquitetura e o aumento de capacidade dos LLM pode ser representada e analisada nos desenvolvimentos dos vários modelos mais recentes. O *Falcon 40B*, por exemplo, tem sido alvo de destaque devido ao treino intensivo em múltiplas línguas, dando a capacidade de tradução e compreensão multilinguagem superior aos restantes (Barbieri et al., 2022). Por outro lado, o *LLaMA*, desenvolvido por *Touvron et al. (2023)* , é notável pela sua eficiência computacional e habilidade em lidar com tarefas de compreensão de texto, demonstrando que é possível alcançar desempenhos elevados com arquiteturas mais otimizadas (Touvron et al., 2023). O *Mistral*, outro modelo que se destacou, representa um avanço em termos de eficiência energética e capacidade de processamento, aspetos cruciais para a implementação prática de sistemas RAG em ambientes com recursos computacionais limitados (Jiang et al., 2023).

No contexto específico deste projeto, o modelo *Gemma 7B* escolheu-se por se considerar como a escolha mais adequada. É um modelo recente desenvolvido pela *Google*, à data da realização deste projeto, que se destaca pela arquitetura única

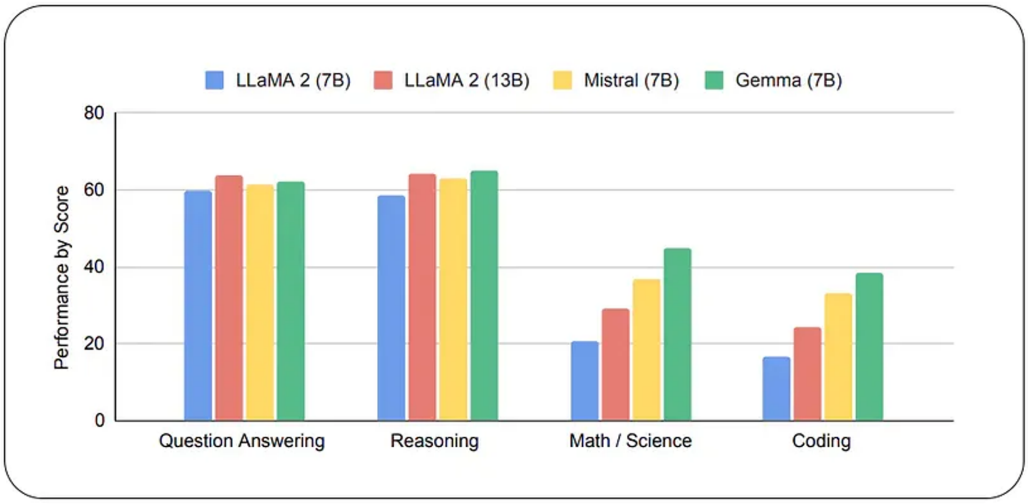

Figura 1: Comparação feita pela google no lançamento do *Gemma 7B*. Fonte: Gemma Team et al., 2024

e pela capacidade de treino com um corpus linguístico diversificado. Conforme detalhado por *Dai et al. (2024)*, este modelo oferece um equilíbrio notável entre precisão e eficiência, características essenciais para aplicações em domínios técnicos como a certificação energética (Gemma Team et al., 2024). A capacidade de gerar respostas com um alto nível de naturalidade e clareza é um dos pontos responsáveis pela escolha deste modelo.

De forma a ter uma noção comparativa, é possível analisar o gráfico da Figura 1, que demonstra uma comparação de performance em diversas áreas, para alguns dos modelos descritos em relação ao modelo escolhido. Este gráfico foi apresentado na literatura científica aquando do lançamento do modelo *Gemma*, e é possível extrair alguns pontos comparativos sobre o desempenho de quatro LLM: LLaMA 2 (7B), LLaMA 2 (13B), Mistral (7B) e Gemma (7B). Estes pontos comparativos foram retirados em 4 contextos distintos: Resposta a Perguntas, Raciocínio, Matemática/Ciência e Código.

O de forte destaque o desempenho consistentemente do Gemma (7B) em todas as categorias onde foi avaliado. O modelo apresenta superioridade nas áreas de Matemática/Ciência e Código em relação aos seus concorrentes.

É de salientar ainda que, apesar de ter apenas 7 mil milhões de parâmetros, o *Gemma 7b* consegue igualar ou superar o desempenho do LLaMA 2 (13B), apesar deste ser um modelo significativamente maior. Esta eficiência indica que o *Gemma* possui uma arquitetura otimizada, que acaba por se traduzir em vantagens práticas

significativas em termos de custos computacionais e velocidade de processamento em aplicações do mundo real.

2.6.1 *Evolução e Características dos Modelos Gemma*

Dentro da família dos modelos Gemma, existem dois modelos principais: o *Gemma 2B* e o *Gemma 7B*. Apesar de pertencerem à mesma família, apresentam algumas diferenças consideráveis nas configurações e capacidades.

Um dos parâmetros fundamentais que distingue ambos os modelos é a dimensão dos *embeddings*, representada pelo *d_model*, como é mostrado na Tabela 1. O *Gemma 2B* utiliza um valor de 2048, enquanto o *Gemma 7B* aumentou esta dimensão para 3072. Este aumento permite ao *Gemma 7B* capturar representações vetoriais mais detalhadas, aumentado ainda mais o potencial do modelo para compreender consoante o contexto das perguntas feitas.

Similarmente, o número de camadas - *'layers'* - aumenta de 18 no *Gemma 2B* para 28 no *Gemma 7B*. Mais camadas, geralmente, significam que o modelo pode aprender representações mais abstratas dos dados de entrada.

As dimensões ocultas das camadas *feedforward*, representadas como *'Feedforward hidden dims'*, também são significativamente ampliadas na transição do *Gemma 2B* para o *Gemma 7B*, passando de 32768 para 49152. As camadas *feedforward* são responsáveis por transformar as representações intermediárias do modelo. O aumento nas dimensões ocultas destas camadas proporciona ao modelo uma maior capacidade de processamento de informações.

Paralelamente, o número de cabeças de atenção, *Num heads*, dobra de 8 para 16 no *Gemma 7B*. Este aumento representa uma modificação na arquitetura do mecanismo de atenção multi-cabeça, tendo implicações substanciais para o modelo. No mecanismo de atenção multi-cabeça, cada cabeça opera independentemente no espaço de representação do *input*, projetando-o em diferentes subespaços de *queries*, *keys* e *values*. O incremento para 16 cabeças permite que o modelo tenha mais facilidade na captura de relações entre os diferentes elementos do *input* do modelo. Além disso, o aumento no número de cabeças pode ser computacionalmente vantajoso, uma vez que múltiplas cabeças de atenção podem ser processadas em paralelo, potencialmente melhorando a eficiência do treino (Narayanan et al., 2021).

Outro avanço relevante no *Gemma 7B* é o aumento no número de cabeças dedicadas ao mecanismo de chave-valor, *'Num KV heads'*, passando de apenas 1 no

Gemma 2B para 16. Com isto, aumenta-se a capacidade do modelo na captação das relações chave-valor nos dados, essencial para gerar texto coerente. Curiosamente, o tamanho de cada cabeça de atenção *'Head size'* e o tamanho do vocabulário *'Vocab size'* permanecem constantes entre os dois modelos, sugerindo uma otimização focada na arquitetura interna em vez da simples expansão do campo lexical.

Tabela 1: Parâmetros-chave dos modelos Gemma.

Parâmetros	**2B**	**7B**
d_{model}	2048	3072
Layers	18	28
Feedforward hidden dims	32768	49152
Num heads	8	16
Num KV heads	1	16
Head size	256	256
Vocab size	256128	256128

Após a análise comparativa dos modelos Gemma, existe espaço para analisar com maior detalhe o *Gemma 7B*, uma vez que é o modelo eleito para o estudo-caso atual do projeto de aplicação de um sistema RAG para aumentar a eficiência de processos de certificação energética.

2.7 O GEMMA 7B: ARQUITETURA E CAPACIDADES

O *Gemma 7B* marca um avanço significativo no campo do processamento de linguagem natural, apresentando inovações arquiteturais que aumentam a capacidade de processamento e geração de texto da família de modelos Gemma.

Assente nos princípios da arquitetura *transformers*, descrita anteriormente neste capitulo, o *Gemma 7B* herda a capacidade de gerir dependências de longo alcance que revolucionaram o campo do PLN (Vaswani et al., 2023).

A arquitetura do *Gemma 7B* é caracterizada por uma rede neural com 28 camadas e 16 cabeças de atenção, conforme detalhado na seção comparativa anterior.

Com esta configuração, este modelo é capaz de detetar complexidades linguísticas e estabelecer relações contextuais mais abrangentes. A dimensão de *embedding* de 3072 dá ao modelo uma capacidade de representação vetorial maior.

Ainda é importante destacar que o *Gemma 7B* é tem capacidade de interpretação baseada em contexto. Esta habilidade despertou a atenção no momento da escolha

uma vez que se considerou particularmente relevante para implementação de sistemas RAG, onde o contexto é extremamente importante para relevância e clareza na geração de respostas às perguntas feitas pelo utilizar.

Outro aspecto que se considerou particularmente interessante do *Gemma 7B* é a capacidade de interpretação contextual de perguntas. Esta funcionalidade é especialmente significativa para este contexto de implementação de sistemas de RAG. O *Gemma 7B*, com esta arquitetura avançada descrita nos parágrafos anteriores, pode analisar não apenas o conteúdo literal das perguntas, mas também as variações contextuais, intenções implícitas e ambiguidades de conteúdo.

Para isto ser possível na criação do modelo, este foi treinado com um corpus extenso de 6 triliões de *tokens*, predominantemente em inglês, utilizando uma ampla gama de fontes, incluindo documentos, exercícios matemáticos e estruturas de código (Gemini Team et al., 2024).

Este processo de treino do *Gemma 7B* seguiu alguns princípios estabelecidos no desenvolvimento do BERT e utiliza técnicas bidirecionais de pré-treino para melhorar a compreensão contextual do modelo. Com esta metodologia bidirecional, o modelo considera simultaneamente os contextos à esquerda e à direita de cada *token* durante o processo, ao contrário dos modelos unidirecionais tradicionais. Assim, esta inovação no processo, resultou em melhores capacidades de compreensão e geração de texto para o modelo.

2.8 MÉTRICAS DE ANÁLISE E AVALIAÇÃO DO SISTEMA

Para analisar o comportamento destas tecnologias, existe na literatura algumas métricas que podem ser aplicadas para análise dos *outputs* dos modelos. Nas secções seguintes, serão exploradas estas as metodologias propostas e implementadas no decorrer do projeto para analisar e avaliar o sistema.

2.8.1 *Distância WMD - Word Mover's Distance*

A Distância de Movimentação de Palavras - inglês Word Mover's Distance (WMD) - é utilizada neste projeto como uma das métricas de avaliação. Aplicada na avaliação de similaridade semântica entre textos, a WMD é introduzida por Kusner et al., 2015 e assenta no conceito de *word embeddings*, ou seja, representações vetoriais de palavras em espaços de alta dimensionalidade. Estes *embeddings*, gerados por

modelos como *Word2Vec* (Mikolov et al., 2013) ou *GloVe* (Pennington et al., 2014), capturam relações semânticas entre palavras, fazendo com que termos com significados semelhantes se localizem próximos num espaço vetorial.

Isto significa que a métrica WMD pode ser utilizada para comparar documentos, frases ou perguntas ao considerar não apenas a presença das palavras, mas tendo em consideração as relações semânticas encontradas através dos *embeddings*.

2.8.1.1 *Fórmula da WMD*

Esta métrica WMD é formalmente expressa como:

$$WMDd, d' \min_{T \geq 0} \sum_{i,j} T_{ij} ci,j \tag{4}$$

onde:

- T_{ij} representa a quantidade de "movimento"necessário para transformar a palavra i do documento d para a palavra j no documento d'.
- ci,j é o custo para mover a palavra i é o custo para mover a palavra j. Este custo é calculado como a distância Euclidiana entre as representações vetoriais, explicado na secção anterior.

Esta formulação matemática, representada na equação 4, descreve explicitamente o objetivo da métrica WMD - quantificar o "trabalho"necessário para alinhar semanticamente dois textos, tendo em consideração tanto a frequência das palavras quanto as relações semânticas no espaço vetorial.

2.8.1.2 *Aplicação da WMD na Análise de Respostas de LLM*

No contexto dos LLM, a WMD revela-se uma métrica capaz de avaliar a qualidade e relevância das respostas geradas, e pode ser aplicada em:

- **Avaliação de Similaridade**: Medir a proximidade semântica entre perguntas e respostas, garantindo que o modelo está a gerar conteúdo relevante.
- **Deteção de Redundância**: Identificar respostas semanticamente similares, de modo a evitar repetições desnecessárias nos sistemas de geração de texto.
- **Análise de Diversidade**: Avaliar a variedade semântica nas respostas geradas, assegurando uma vasta cobertura de tópicos quando necessário.

- **Comparação com Referências**: Comparar respostas geradas com respostas de referência, útil em tarefas de avaliação e *benchmarking* de modelos.

Como o foco é a utilização para avaliação do sistema atual, uma distância WMD baixa entre duas perguntas indica que são semanticamente semelhantes, abordando o mesmo tópico ou relativamente a mesma ideia, apesar de utilizarem palavras diferentes. Por outro lado, uma distância WMD alta sugere que as perguntas são semanticamente distintas, refletindo uma maior diversidade temática (L. Wu et al., 2018). Esta métrica distingue-se das outras métricas tradicionais, como a similaridade de cosseno, porque não se limita a comparar palavras individuais, mas considera o custo de transformar todo o conteúdo semântico de um documento noutro, proporcionando uma medida mais fidedigna de similaridade semântica.

2.8.2 *Divergência Kullback-Leibler (KL)*

A Divergência Kullback-Leibler (KL) é uma métrica utilizada para quantificar a diferença entre duas distribuições de probabilidade. Essencialmente, a Divergência KL mede a quantidade de informação perdida quando uma distribuição de probabilidade aproximada Q é usada para descrever uma distribuição de probabilidade verdadeira P.

Formalmente, a Divergência KL entre uma distribuição de probabilidade real P e uma distribuição de probabilidade aproximada Q é definida pela seguinte fórmula:

$$D_{KL}P \parallel Q \sum_{i} Pi \log \frac{Pi}{Qi} \tag{5}$$

onde Pi representa a distribuição de probabilidade real ou observada, enquanto Qi representa a distribuição de probabilidade que se deseja comparar com a verdadeira.

A fórmula da Divergência KL é usada para calcular o "custo"associado a descrever a distribuição verdadeira P utilizando a distribuição aproximada Q. Se Q for idêntica a P, a divergência KL será zero, indicando que não há perda de informação. No entanto, à medida que Q diverge de P, o valor de $D_{KL}P \parallel Q$ aumenta, sendo perceptível uma maior perda de informação.

No contexto dos LLM, a Divergência KL pode ser aplicada na avaliação qualidade de resposta de um LLM ao comparar a distribuição de probabilidade gerada pelo modelo com a distribuição de referência verdadeira dos dados. Ou seja, permite

determinar o quão bem o modelo está a capturar a distribuição verdadeira dos vetores (Balaguer et al., 2024).

Além disso, a Divergência KL é frequentemente utilizada em técnicas de regularização em modelos como *Variational Autoencoders*, onde ajuda a regular a distribuição latente aprendida pelo modelo.

É ainda importante compreender-se distribuição latente, visto que este conceito se refere à distribuição probabilística no espaço latente do modelo, que é uma representação de dimensão reduzida dos dados de entrada. Este termo de regularização penaliza o desvio desta distribuição latente aprendida em relação a uma distribuição anterior predefinida, tipicamente uma distribuição normal multivariada.

Esta abordagem tem como objetivo prevenir o *overfitting* ao impor uma estrutura específica no espaço latente, de forma a facilitar a generalização do modelo. Ao aproximar a distribuição latente de uma distribuição conhecida, o modelo torna-se mais capaz de gerar amostras mais coerentes alcançando assim o objetivo de gerar respostas mais fidedignas, aumentando a relevância e a integridade.

3

METODOLOGIA E ARQUITETURA

3.1 VISÃO GERAL DO SISTEMA

O sistema proposto neste projeto representa uma abordagem inovadora para a integração de LLM com técnicas avançadas de recuperação de informação, especificamente no contexto da certificação energética em Portugal. Esta secção apresenta uma visão holística da arquitetura do sistema, destacando os componentes principais e as interações cruciais entre cada parte do sistema.

O núcleo do sistema assenta na sinergia entre três elementos principais: o LLM - *Gemma 7B*, o mecanismo de RAG, e uma Base de Dados Vetorial (VecDB). Esta é a tríade tecnológica necessária para proporcionar respostas integras a perguntas complexas no domínio da certificação energética.

3.2 ARQUITETURA DO SISTEMA

A arquitetura do sistema proposto, por sua vez, é composta por cinco componentes principais, onde cada um desempenha um papel importante na implementação do sistema RAG com VecDB aplicado ao modelo *Gemma 7B*.

O sistema desenvolvido é baseado em:

- Interface do Utilizador - Serve como componente de interação dos utilizadores com a tecnologia. É o método de introdução das perguntas dos utilizadores e representação dos resultados dos modelos.
- LLM (*Gemma 7B*) - Componente responsável pelo processamento das consultas e geração de respostas coerentes.
- Sistema RAG - É o componente de ligação entre o LLM e a VecDB, responsável por aumentar as capacidades de resposta do modelo com informações externas relevantes e atualizadas.

- Base de Dados Vetorial (ChromaDB) - Armazena e indexa representações vetoriais de documentos relacionados à certificação energética, permitindo a recuperação eficiente de informações relevantes.
- Módulo de Processamento de Documentos - Responsável por receber, processar e vetorizar os novos documentos, de forma a manter a base de conhecimento atualizada.

3.3 DESCRIÇÃO DO FLUXO DE OPERAÇÃO

O sistema opera através de um fluxo de informação cíclico, projetado para processar as perguntas do utilizador e fornecer respostas precisas e contextualizadas.

Este processo começa com a interação do utilizador e passa por várias etapas de processamento antes de devolver uma resposta.

Inicialmente, o utilizador interage com o sistema submetendo uma pergunta através da interface de utilizador. Esta interface é o ponto de entrada para todas as perguntas, tendo como foco principal uma experiência intuitiva e acessível para os utilizadores.

Uma vez submetida, a pergunta é imediatamente encaminhada para a próxima fase de processamento, o sistema RAG.

O sistema RAG assume o controlo, e inicia o processamento da pergunta. Nesta fase, a pergunta do utilizador é transformada num formato que o sistema pode processar. Este passo envolve a vetorização da pergunta, onde o texto é convertido numa representação vetorial, e é feita uma análise detalhada para identificar os elementos-chave e a intenção associada à pergunta.

Com a pergunta processada, o sistema RAG prossegue para a etapa de recuperação de informação. É aqui que acontece a interação com a VecDB. O RAG utiliza os vetores gerados na etapa anterior para consultar a VecDB, identificando as informações que sejam uteis para a pergunta do utilizador. Esta etapa apenas acontece para garantir que a resposta final é fundamentada com dados atualizados.

Uma vez encontradas as informações relevantes, através de sistemas como a similaridade descrita no Capítulo anterior, o sistema passa para a fase de geração de resposta. Nesta fase, o LLM *Gemma 7b* recebe tanto a pergunta original quanto o contexto recuperado da VecDB. É aqui que o modelo *Gemma 7b* gera uma resposta que não apenas responde diretamente à pergunta do utilizador, mas também

incorpora de forma coerente as informações contextuais relevantes absorvidas pela fase de RAG.

A etapa final do processo é a apresentação da resposta ao utilizador. A resposta gerada pelo LLM é formatada para garantir clareza e legibilidade, e é enviada de volta à interface do utilizador.

Este processo iterativo é o que permite que o sistema se adapte continuamente. A cada interação, o sistema tem a oportunidade de adaptar novas informações o que significa que o sistema não mantém apenas a relevância, mas também melhora progressivamente a qualidade das respostas.

3.4 METODOLOGIA DETALHADA DE IMPLEMENTAÇÃO

A implementação do sistema de RAG integrado com bases de dados vetoriais e aplicado ao modelo *Gemma 7b* segue uma abordagem modular e sistemática, dividida em várias etapas. O intuito desta secção é precisamente detalhar cada fase de implementação, referindo as ferramentas, as respetivas versões utilizadas, descrevendo o processo de forma detalhada e dando uma noção prática de como foi desenvolvido cada fase do projeto.

3.4.1 *Configuração e Preparação do Ambiente*

A implementação do sistema iniciou-se com a configuração do ambiente de desenvolvimento. Optou-se pela utilização do *Python* [1] na versão 3.11, assegurando assim a compatibilidade com todas as bibliotecas necessárias para o projeto. A escolha da linguagem fundamentou-se na estabilidade e nas características adequadas do *Python* ao desenvolvimento de projetos de inteligência artificial.

O primeiro componente desenvolvido foi um módulo personalizado de carregamento de configurações, denominado *LoadConfig*. Este módulo é responsável pela leitura e interpretação do ficheiro de configuração no formato *YAML*. Este ficheiro é onde se localiza os parâmetros essenciais, incluindo os *paths* de modelo, configurações de base de dados e variáveis de ambiente utilizadas. Este módulo criou-se de forma a garantir uma gestão centralizada das configurações, para facilitar ajustes futuros.

1 Python é uma linguagem de programação de alto nível utilizada em automação, ciência de dados e inteligência artificial. É possível encontrar mais informação em: https://www.python.org

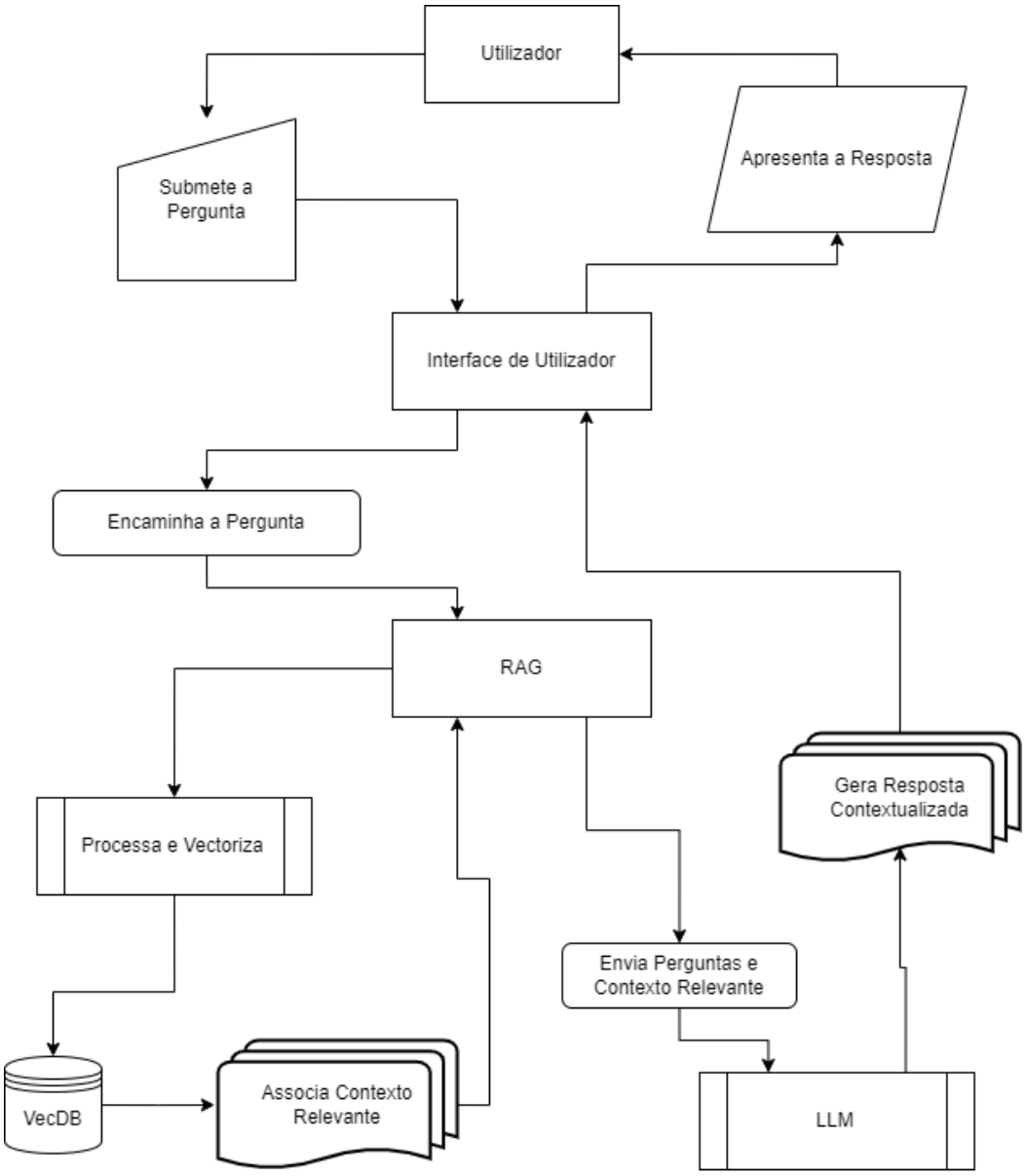

Figura 2: Fluxo detalhado da operação do sistema

Paralelamente, procedeu-se à implementação do servidor *backend*, optando-se pelo *framework Flask* devido à flexibilidade desta ferramenta. As configurações carregadas pelo *LoadConfig* são utilizadas para parametrizar vários aspetos do servidor, como a definição de *routes*. Esta abordagem assegura uma consistência na configuração ao longo de todo o sistema.

3.4.2 *Carregamento do Modelo e Tokenizador*

A etapa subsequente do desenvolvimento focou-se no carregamento e inicialização do modelo de linguagem. Para esta tarefa, selecionou-se o modelo *Gemma 7b*. A escolha deste modelo baseou-se na sua capacidade de compreender e gerar texto de forma contextualmente relevante sendo um modelo diferenciado à data do desenvolvimento deste projeto, alinhando-se com os objetivos propostos no Capítulo inicial.

Para facilitar o carregamento do modelo *Gemma 7b* e do *tokenizador* associado, recorreu-se à biblioteca *Transformers*, desenvolvida pela *Hugging Face*. Esta é uma das bibliotecas principais mais reconhecidas na comunidade de *machine learning*. Especificamente, utilizaram-se as classes *AutoTokenizer* e *AutoModelForCausalLM*, que oferecem uma interface simplificada para a inicialização de modelos.

Primeiramente, utilizou-se a classe *AutoTokenizer* para inicializar o *tokenizador* específico do *Gemma 7b*. Este componente é essencial para o pré-processamento do texto de entrada, convertendo-o em tokens que o modelo pode interpretar. A correta configuração do *tokenizador* garante que o texto seja segmentado e codificado de forma consistente com o treino original do modelo.

Em seguida, utilizou-se a classe *AutoModelForCausalLM* para carregar os pesos pré-treinados do modelo *Gemma 7b*. Este processo é computacionalmente intensivo, envolvendo a alocação de uma quantidade significativa de memória para armazenar os parâmetros do modelo. A utilização desta classe simplifica consideravelmente o processo, uma vez que assegura que todos os componentes do modelo são iniciados corretamente e estejam prontos para uso na fase em que são necessários no processo.

A decisão de utilizar estas classes da biblioteca *Transformers* foi estratégica, pois além de simplificar o processo de inicio, minimiza a possibilidade de erros de configuração. As classes *AutoTokenizer* e *AutoModelForCausalLM* são projetadas para detectar automaticamente as configurações corretas com base no modelo, reduzindo assim a necessidade de ajustes manuais.

Para garantir a eficiência deste processo de carregamento, implementaram-se técnicas de otimização de memória. Uma vez que é um ambiente de desenvolvimento, incluiu-se um procedimento de carregamento rápido, onde parte do modelo é carregada em precisão reduzida para poupar alguma memória, sem comprometer significativamente o desempenho. Também se considerou a possibilidade de carregamento do modelo em vários dispositivos *GPU*, caso disponíveis, para distribuir a carga computacional, no entanto, não se optou por essa metodologia já devido aos custos associados.

3.4.3 *Preparação da Base de Dados Vetorial*

A preparação da VecDB representa uma das fases mais críticas de todo o processo de implementação deste sistema. Dedicou-se uma atenção extra a uma série de procedimentos de forma a garantir a qualidade e a utilidade dos dados armazenados.

Antes da criação da base de dados, os documentos são submetidos a um processo de segmentação. No entanto, devido à importância deste procedimento, criou-se uma secção dedicada exclusivamente a este processo de processamento de dados, que pode ser encontrada no seguimento deste documento.

Após o processamento dos dados e com os documentos devidamente segmentados, procede-se à criação e configuração da base de dados vetorial, com a tecnologia *ChromaDB*. É nesta etapa que acontece a transformação dos dados textuais em representações vetoriais. O processo começa com a iniciação do *client ChromaDB* e a criação de uma coleção específica dedicada aos documentos de certificação energética.

Cada segmento de texto resultante do processo de *splitting* é então submetido a um processo de vetorização, onde é convertido numa representação vetorial de alta dimensão, comumente referida como *embedding*. Este processo de *embedding* está descrito na secção seguinte, uma vez que é um processo relevante no sistema e merece uma atenção detalhada.

Além dos vetores, cada segmento é enriquecido com metadados relevantes. Estes metadados incluem informações como a fonte do documento original, um identificador único para o segmento, e outras informações contextuais como data de criação. A inclusão destes metadados é essencial para manter a rastreabilidade e o contexto dos dados, permitindo recuperações mais precisas e fornecendo informações adicionais que podem ser uteis durante o processo de geração de respostas.

O armazenamento destes vetores e metadados na base de dados ChromaDB é otimizado para pesquisa rápida e eficiente. A estrutura da base de dados é projetada para suportar *queries* de similaridade vetorial de alta performance, essenciais para a rápida recuperação de informações relevantes durante a operação do sistema RAG, que está também detalhado na secção seguinte.

3.4.4 *Integração do RAG*

O processo inicia-se com a codificação da pergunta, onde a pergunta do utilizador é transformada num vetor de alta dimensão. Para este efeito, utiliza-se o mesmo modelo de *embeddings* aplicado anteriormente aos documentos da base de dados. A escolha foi o modelo *"BAAI/bge-large-en-v1.5"*[2] da *Hugging Face*, destacado pela eficácia na identificação de contextos semânticos. Esta decisão garante uma representação vetorial coerente entre as perguntas e os documentos armazenados, e evita que existam potenciais discrepâncias semânticas que poderiam comprometer o funcionamento do sistema desenvolvido neste projeto.

Subsequentemente, implementou-se o algoritmo de pesquisa de similaridade do cosseno na base de dados *ChromaDB*, previamente configurada. Esta abordagem permite identificar os segmentos de texto mais relevantes para a consulta do utilizador, baseando-se no significado semântico em vez de simples correspondências de palavras-chave. Após alguns testes, arbitrou-se que o sistema utilizaria os três segmentos mais similares, um número que se determinou empiricamente como o equilíbrio ideal entre fornecer contexto suficiente e não sobrecarregar o modelo com informações.

O passo seguinte envolveu o desenvolvimento de um algoritmo de fusão. Este algoritmo é responsável por combinar a pergunta original do utilizador com os contextos recuperados da base de dados. O resultado deste processo é um *prompt* estruturado que inclui:

- A pergunta original do utilizador.
- Os segmentos de texto recuperados, ordenados por relevância.

A estrutura do *prompt* tem influência direta na qualidade da resposta gerada pelo modelo.

Após a criação do *prompt* estruturado, este é processado pelo modelo *Gemma 7b*. O modelo, alimentado com o contexto e as instruções específicas, gera então

2 Disponível em: https://huggingface.co/BAAI/bge-large-en-v1.5.

uma resposta que idealmente incorpora tanto a precisão técnica dos documentos de certificação energética quanto a relevância para a pergunta específica do utilizador.

3.4.5 *Geração de Embeddings*

A geração de *embeddings* constitui um processo que ocorre em duas fases distintas: no processamento da pergunta colocada pelo utilizador ao modelo de linguagem e na pesquisa das respostas mais adequadas que possam corresponder ao contexto da pergunta. Estes *embeddings* são representações numéricas dos dados textuais, denominados vetores, que capturam as características essenciais do texto, facilitando assim o processo de pesquisa semântica.

Para esta componente crucial da implementação, selecionou-se o modelo *BAAI/bge-large-en-v1.5*, pertencente à família *BGE (BAAI General Embeddings)*. Esta escolha baseou-se em dois fatores principais: a natureza *open-source* do modelo e a reputação na eficácia no processamento de linguagem natural, fundamentada também na literatura científica já acima mencionada (Vaswani et al., 2023).

O processo responsável pela geração de *embeddings* divide-se em quarto etapas, começando pela entrada de texto. O sistema recebe como entrada os documentos que são submetidos ao modelo para a conversão em *embeddings*. De seguida, inicia-se o processo de *tokenização*, onde o texto é segmentado em unidades menores chamadas *tokens*. O *tokenizador* específico do modelo BGE considera-se apto para lidar com características linguísticas complexas, como é o caso da semântica portuguesa. Para ilustrar o funcionamento deste processo, considere-se a seguinte frase:

> *A área efetiva coletora da radiação solar do vão envidraçado na estação de aquecimento, corresponde à área que é utilizada para efeitos de contabilização dos ganhos solares.*

Após a *tokenização*, essa frase poderia ser segmentada nos seguintes *tokens*:

```
["A", "área", "efetiva", "coletora", "da", "radiação", "solar",
"do", "vão", "envidraçado", "na", "estação", "de", "aquecimento","
"corresponde", "à", "área", "que", "é", "utilizada", "para",
"efeitos", "de", "contabilização", "dos", "ganhos", "solares",
"."]
```

De seguida, cada um dos *tokens* são processados através das camadas do modelo *BGE*, e o mecanismo de atenção permite que o modelo capture as relações

contextuais entre diferentes partes do texto e gera uma representação vetorial de informações semânticas. O resultado final deste processamento é um vetor denso de 1024 dimensões.

A escolha de uma dimensão de vetor de 1024 determina a quantidade de informações que podem ser armazenadas e representadas. Com 1024 dimensões, o vetor tem espaço suficiente para representar diferentes aspetos do significado das palavras. Quanto maior for a dimensionalidade mais facilmente o modelo a captura as dependências de longo alcance e variações subtis do significado das palavras com base no contexto.

No entanto, é importante salientar que a decisão sobre a dimensionalidade do vetor envolveu uma ponderação cuidada entre a capacidade de representação e o custo computacional. Embora dimensões mais elevadas possam, em teoria, capturar ainda mais variações semânticas, também aumentam significativamente os requisitos computacionais. A escolha de 1024 dimensões representa um equilíbrio entre a captura de informações semânticas e a manutenção da eficiência computacional, alinhando-se com os objetivos de implementação do projeto.

3.4.6 *Desenvolvimento da Interface do Utilizador*

Após todo o processamento descrito nas secções anteriores, procedeu-se ao desenvolvimento da interface de utilizador com recurso à biblioteca *Gradio* [3].

O processo de desenvolvimento da interface iniciou-se com a criação de uma função de *callback*. Esta função desempenha o papel de ponte entre o *input* do utilizador e o método *process_query*, previamente desenvolvido na implementação do RAG. É através desta ligação que as perguntas dos utilizadores são processadas pelo sistema RAG, garantindo respostas contextualizadas.

Para melhorar a funcionalidade da interface, adicionaram-se três opções adicionais, aproveitando as capacidades nativas da biblioteca *Gradio*:

- Botão de Limpeza (*clear*): Esta funcionalidade permite aos utilizadores limpar rapidamente o campo de entrada e a área de resposta, preparando a interface para um novo tema de discussão.

3 Gradio é uma biblioteca Python para criar rapidamente interfaces de utilizador. Mais informações estão disponíveis em: https://gradio.app

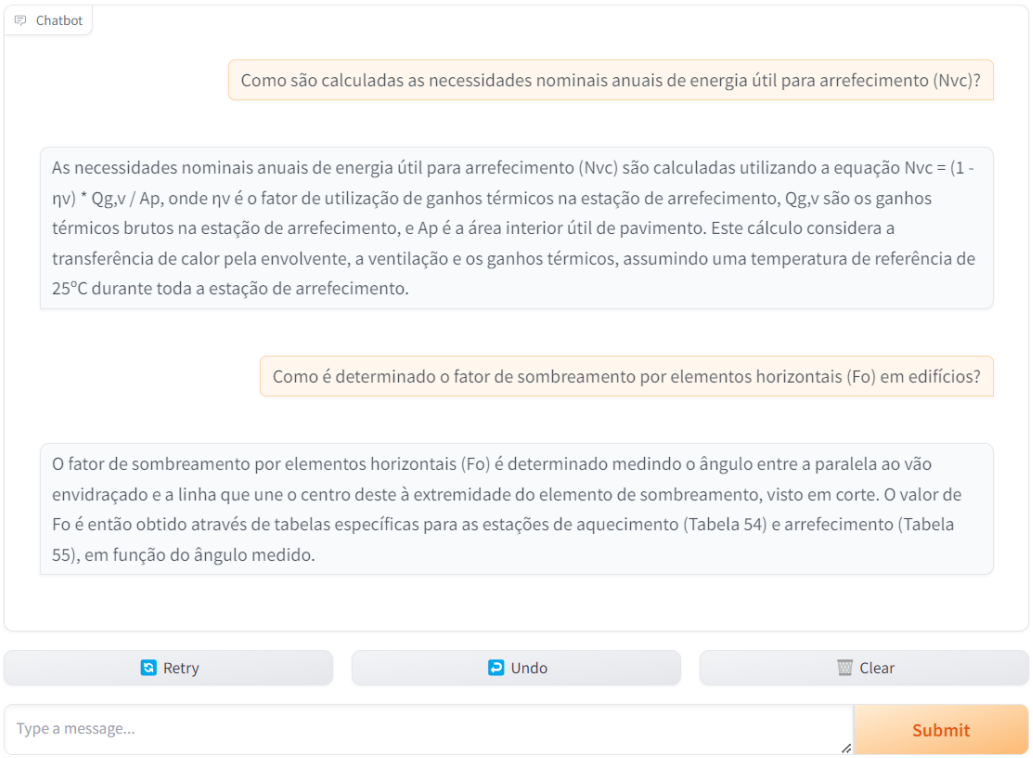

Figura 3: *Layout* da interface de utilizador utilizando a bibliteca *Gradio*

- Botão de Anulação (*undo*): Implementou-se a opção de anular a última pergunta submetida, oferecendo aos utilizadores a flexibilidade de recuar caso desejem reformular a questão feita.
- Botão de Reenvio (*retry*): Esta opção possibilita o reenvio da última pergunta, uma funcionalidade útil para casos em que os utilizadores desejam obter uma resposta alternativa ou confirmar a consistência do sistema, bastante utilizada nos testes empíricos.

A integração visual destes componentes pode ser observada na Figura 3, que ilustra a disposição e o design da interface desenvolvida.

A escolha da biblioteca *Gradio* revelou-se útil para este desenvolvimento devido à facilidade de implementação e pela capacidade de criar interfaces responsivas e visualmente apelativas de forma bastante rápida. Assim assegura-se que a interface não é apenas funcional, mas também agradável de utilizar, contribuindo para uma experiência de utilizador positiva.

3.5 ANÁLISE E UTILIZAÇÃO DOS DADOS

3.5.1 *Origem e Natureza dos Dados*

O corpus de dados utilizado neste estudo é proveniente do "Manual SCE"(Sistema de Certificação Energética dos Edifícios)[4], uma publicação oficial da entidade ADENE. Este manual constitui uma fonte credível e abrangente sobre a legislação e os procedimentos de certificação energética em Portugal. A escolha deste *dataset* foi estratégica, uma vez que aborda os desafios específicos associados ao processamento de dados para contextos técnicos e legais.

Os dados utilizados são:

- Origem oficial: Garante a autenticidade e relevância das informações.
- Natureza técnico-legal: Apresenta um desafio significativo para o processamento de linguagem natural devido à sua complexidade linguística.
- Abrangência temática: Cobre diversos aspetos da certificação energética, desde fundamentos legais até procedimentos técnicos detalhados.

3.5.2 *Estrutura e Conteúdo do Dataset*

O "Manual SCE"está estruturado em várias secções, cada uma abordando aspetos específicos da certificação energética:

- Enquadramento
- Termos e Definições
- Certificação Energética
- Documentação de Suporte
- Caracterização do Edifício
- Abordagem Técnica - Envolvente Opaca, Envidraçada, Ventilação, Climatização, Água e Clima, Iluminação Fixa, Instalações de Elevação, entre outros
- Classe Energética
- Medidas de Melhoria

4 A ADENE – Agência para a Energia é uma entidade portuguesa dedicada à promoção da eficiência energética e sustentabilidade. Mais informações em: https://www.adene.pt

Esta estrutura multifacetada permite uma análise abrangente das capacidades do sistema RAG em lidar com diversos tipos de informação dentro do mesmo domínio.

3.5.3 *Preparação e Pré-processamento dos Dados*

O processo de preparação dos dados foi simplificado e otimizado, considerando que os documentos estavam disponíveis em formato PDF. Esta uniformidade no formato de entrada permitiu uma abordagem mais direta e eficiente no pré-processamento, do que aconteceria caso fossem combinados vários documentos de estruturas diferentes.

Começou-se pela extração de texto, onde se utilizou a biblioteca *PyPDFLoader* [5] para extrair o conteúdo dos documentos PDF. Este processo manteve a integridade do texto e a estrutura original dos documentos, preservando informações cruciais como formatação e layout. De seguida removeram-se caracteres especiais, normalizaram-se espaços em branco e quebras de linha para manter consistência no texto e corrigiram-se erros de codificação de caracteres, especialmente em textos com acentuação, como é requisito da língua portuguesa.

O processo de segmentação, aplicado logo de seguida, dividiu o texto em segmentos menores através o *RecursiveCharacterTextSplitter*[6]. Os parâmetros foram cuidadosamente ajustados para manter a coerência semântica e respeitar a estrutura natural dos documentos, usando um tamanho de *chunk* de 1000 caracteres e uma sobreposição de 200 caracteres. Estes valores foram determinados empiricamente para otimizar o equilíbrio entre contexto e granularidade da informação.

A determinação dos parâmetros de segmentação textual, nomeadamente o tamanho do *chunk* de 1000 caracteres e a sobreposição de 200 caracteres é sustentada pelas evidências empíricas e teóricas na literatura disponível. Na literatura, enfatiza-se a importância de preservar a coerência semântica em segmentos de texto, sugerindo que *chunks* maiores capturam melhor as dependências contextuais necessárias para a compreensão do texto. Esta abordagem é reforçada por Mikolov et al., 2013, que demonstraram que janelas de contexto mais amplas melhoram a qualidade das representações semânticas .

5 PyPDFLoader é uma ferramenta da biblioteca *LangChain* para carregar e processar documentos PDF em *Python*, facilitando a extração de texto e o processamento de documentos. Mais informações disponíveis em: https://api.python.langchain.com/en/latest/document_loaders/langchain_community.document_loaders.pdf.PyPDFLoader.html

6 RecursiveCharacterTextSplitter é uma ferramenta da biblioteca *LangChain* desenvolvida para dividir grandes blocos de texto em segmentos menores, mantendo a coerência semântica. Mais informações disponíveis em: https://api.python.langchain.com/en/latest/text_splitters/langchain.text_splitter.RecursiveCharacterTextSplitter.html

A sobreposição de 200 caracteres, por sua vez, alinha-se com as observações de Cho et al., 2014 e Sutskever et al., 2014 que destacam a importância da redundância moderada para manter a precisão contextual entre segmentos. Este equilíbrio entre o tamanho do *chunk* e a sobreposição tem como objetivo a otimização do contexto semântico, minimizando simultaneamente a redundância excessiva, o que é particularmente relevante para sistemas RAG que dependem de uma compreensão precisa e contextualizada do texto para gerar respostas adequadas.

4

RESULTADOS

Enquanto existem métodos consolidados para medir a eficácia de redes neurais em tarefas como classificação de imagens, onde os critérios de avaliação são mais definidos e diretos, avaliar LLM é muito mais complexo.

A complexidade associada aparece devido à natureza aberta do texto gerado, tornando a avaliação de qualidade um desafio. É importante destacar que em tarefas de resposta a perguntas não existe um "ranking ideal"claro, como ocorre em cenários de pesquisa, complicando as comparações diretas sobre o quão bem um modelo atende aos critérios ideais. Estas dificuldades são amplificadas pelo facto de que os LLM, especialmente na aplicação de RAG em domínios fechados, deverem não só gerar texto coerente e relevante, mas também terem a capacidade de generalização para domínios associados ao domínio principal, tudo isto considerando a necessidade de reduzir ocorrências de erros de geração, como é o caso das "alucinações".

Assim, para garantir uma medida mais detalhada da qualidade e relevância das respostas geradas pelo LLM, criaram-se testes empíricos comparativos. Neste processo, as respostas fornecidas pelo LLM são comparadas manualmente com as informações contidas nos documentos de referência utilizadas no sistema de RAG.

4.1 TESTES EMPÍRICOS COMPARATIVOS

O processo de testes empíricos para avaliar LLM, envolve frequentemente uma abordagem de Human-in-the-Loop (HITL). Nestes processos, é necessário envolver humanos para a validação das respostas geradas pelo modelo. Apesar da avaliação começar com avaliação através de outro LLM a analisar as respostas dadas pelo sistema implementado neste projeto, na segunda fase de avaliação é utilizada a comparação manual por parte de um humano com acesso ao manual SCE utilizado referente aos procedimentos e métodos da certificação energética em Portugal.

4.1.1 *Métricas para Avaliação Empírica do Sistema RAG*

Esta Secção descreve o conjunto abrangente de métricas concebidas para avaliar a qualidade do processo de perguntas e respostas no âmbito da certificação energética. Estas métricas criaram-se essencialmente para aferir a eficácia do sistema de RAG implementado, que utiliza o manual do Sistema de Certificação Energética como fonte de conhecimento externa. A avaliação é feita através de testes empíricos, recorrendo ao *GPT-4* como ferramenta de análise para uma avaliação objetiva.

Destaca-se que que a qualidade das respostas geradas pelo modelo está intrinsecamente ligada, não apenas à eficácia da pesquisa vetorial, mas também à qualidade intrínseca do modelo utilizado. Consequentemente, para obter respostas de elevada qualidade, é imperativo garantir que o sistema de pesquisa vetorial e todos os elementos fundamentais descritos no Capítulo 3 estejam rigorosamente implementados e otimizados.

No desenvolvimento destas métricas de avaliação, consideraram-se diversos desafios inerentes à natureza subjetiva da qualidade das perguntas. A variabilidade de opiniões sobre o que constitui uma pergunta relevante, informativa ou envolvente representa um obstáculo significativo. Assim, as métricas propostas visam avaliar a qualidade de forma objetiva, superando esta subjetividade intrínseca.

Adicionalmente, as métricas criadas têm em consideração a dependência contextual da relevância e utilidade das perguntas. Reconheceu-se durante os testes que uma pergunta pode ser altamente pertinente num contexto específico, mas irrelevante noutro, sublinhando a necessidade de métricas sensíveis ao contexto.

A avaliação da diversidade das perguntas geradas constitui outro aspeto. Para ter um sistema robusto é crítico que seja capaz de produzir uma variedade de questões que abranjam diferentes vertentes do conteúdo. No entanto, a quantificação da diversidade apresenta desafios consideráveis, envolvendo a avaliação da singularidade das perguntas e a sua semelhança com o conteúdo original e outras perguntas geradas.

Outros elementos vitais na avaliação da qualidade das perguntas são a correção gramatical e a fluência. Embora existam ferramentas automatizadas para avaliar a correção gramatical, a avaliação da fluência frequentemente requer intervenção humana, o que complica a criação de métricas totalmente automatizadas neste aspeto, tendo sido ultrapassado com a introdução de um humano no processo de análise.

Por fim, a capacidade de resposta com base no conteúdo fornecido é um critério essencial para uma pergunta de qualidade. Avaliar se uma pergunta pode ser respondida de forma precisa através da utilização das informações disponíveis nas bases de conhecimento externas exige uma compreensão abrangente do conteúdo e a capacidade de identificar as informações mais relevantes.

Na literatura atual existem métricas destinadas à avaliação da qualidade das respostas em sistemas RAG, que asseguram a precisão, relevância e eficácia das respostas geradas por modelos de linguagem. Contudo, identifica-se uma lacuna significativa no que diz respeito a métricas específicas para avaliar a qualidade das perguntas colocadas ao modelo.

4.1.2 *Metodologia de Análise para a Avaliação de Perguntas ao Sistema RAG*

A metodologia de análise para a avaliação de perguntas no sistema RAG desenvolveu-se com o objetivo de avaliar de forma sistemática e objetiva a qualidade, não só das respostas, mas também das perguntas colocadas ao sistema RAG no contexto deste projeto, combinando fatores qualitativos e quantitativos.

Na arquitetura desta metodologia, aplicam-se seis métricas principais de avaliação, cada uma projetada para identificar um aspeto específico da qualidade das questões e baseadas na abordagem de Balaguer et al., 2024.

A primeira métrica - Avaliação de Relevância - é dividida pela Relevância Contextual e Relevância Global. Trata-se de uma metodologia empírica que permite avaliar as questões de forma relativa e de forma global.

A Relevância Contextual utiliza o *GPT-4* para avaliar as questões numa escala de 1 a 5, analisando a importância em relação ao contexto específico atribuído ao sistema, ou seja, relacionado com certificação energética. Por outro lado, a Relevância Global avalia a relevância das questões no âmbito geral da eficiência energética, sem considerar um contexto específico, como se vê na Tabela 3.

A segunda métrica - Cobertura - mede o grau de fundamentação das respostas no contexto fornecido. Utilizando também o *GPT-4*, as respostas são avaliadas numa escala de 1 a 5, onde 5 indica uma resposta totalmente fundamentada no contexto e 1 uma resposta sem contexto fornecido para ser respondida com precisão. Para isto, utilizou-se o *prompt* "Classifica de 1 a 5 se é possível extrair uma resposta a esta pergunta através contexto e da pergunta feita. O nível 5 indica que a resposta

Tabela 2: Exemplos de Análise de Relevância para Perguntas sobre Certificação Energética

Pergunta	**Relevância Contextual**	**Resposta**
Como é determinado o U para paredes de edifícios construídos antes de 1960?	5	Esta pergunta é de relevância alta porque trata diretamente a metodologia de cálculo do U, essencial para a certificação energética de edifícios antigos, um tema abordado no documento de referência.
Como a espessura da parede afeta o U em edifícios construídos antes de 1960?	4	Esta pergunta é relevante, pois a espessura é um dos fatores que influenciam o U, mas a pergunta não é tão específica quanto outras no contexto da certificação energética.
Quais são os métodos de construção mais eficazes para melhorar o U em paredes construídas após 1960?	1	Esta pergunta é pouco relevante porque se refere a métodos de construção que em nada estão relacionados com o documento fonte.

Tabela 3: Exemplos de Pontuações de Relevância Global para Perguntas sobre Certificação Energética em Edifícios

Pergunta	**Relev. Global**	**Explicação**
Como o U das paredes de edifícios construídos antes de 1960 influencia a eficiência energética global do edifício?	5	Esta pergunta é altamente informativa e relevante, pois aborda diretamente a relação entre as características construtivas de edifícios e o tema da eficiência energética.
Qual é o impacto das mudanças climáticas nas correntes oceânicas do Atlântico Norte em termos de certificação energética?	1	Esta pergunta é técnica e científica, mas não tem qualquer sentido lógico nem relação com a certificação energética ou com as características construtivas de edifícios, tornando-a irrelevante para o tema em questão.

é totalmente fundamentada através do contexto da pergunta. No nível 1, a resposta não tem base de contexto para ser respondida."

Com isto, tenta-se monitorizar se o sistema RAG está a gerar conteúdo preciso e baseado em informações reais, como o contexto da pergunta, e analisar se há forte possibilidade de alucinação justificada pela ausência de contexto na pergunta feita. Na Tabela 4 registou-se um exemplo de dois extremos que fizeram parte desta análise. É possível perceber que a pergunta sobre o U pode ser respondida diretamente a partir das informações fornecidas no contexto, o que justifica a pontuação máxima. No outro extremo, a pergunta sobre painéis solares na indústria não tem relação com o contexto fornecido, impossibilitando a extração de uma resposta adequada, resultando na pontuação mínima.

Tabela 4: Exemplos de Pontuações de Cobertura para Perguntas relacionadas à Certificação Energética em Edifícios

Pergunta	**Cobertura**	**Explicação**
Qual é o U típico para paredes de edifícios construídos antes de 1960 em Portugal?	5	O contexto fornecido inclui informações detalhadas sobre os coeficientes de transmissão térmica para paredes de edifícios antigos, incluindo valores de referência para construções anteriores a 1960, permitindo uma resposta direta e precisa com base nas informações base do documento.
Como a implementação de painéis solares na indústria afeta a classificação energética de edifícios em áreas urbanas?	1	O contexto discutido aborda exclusivamente painéis solares relativos à industria, tornando a resposta impossível de ser extraída do contexto.

A terceira métrica - Sobreposição Semântica - avalia a similaridade semântica entre as questões geradas e o texto fonte. Para isso, utiliza-se o cálculo da divergência *Kullback-Leibler* entre as distribuições de palavras. Uma menor divergência KL indica uma maior sobreposição semântica, o que é desejável para garantir a coerência entre as questões e o conteúdo de origem. Esta métrica está exemplificada na Tabela 5.

A quarta métrica - Diversidade - é avaliada através do cálculo da Distância de Movimentação de Palavras WMD entre pares de questões. Uma maior média WMD indica uma maior diversidade entre as questões geradas, o que é importante para garantir que o sistema está a produzir um conjunto variado de perguntas que cobrem diferentes aspetos do tema.

Tabela 5: Exemplos de Sobreposição Semântica utilizando Divergência KL para Perguntas relacionadas à Certificação Energética em Edifícios

Pergunta	**Divergência KL**	**Explicação**
Qual é o U para paredes de edifícios anteriores a 1960?	0.15	A baixa divergência KL indica que esta pergunta está fortemente alinhada semanticamente com o texto fonte, que trata de características térmicas de edifícios antigos.
Como os painéis solares podem melhorar a eficiência energética em edifícios novos?	1.25	A alta divergência KL sugere que esta pergunta é menos coerente com o texto fonte, que não discute painéis solares, mas sim propriedades térmicas de paredes antigas.

A quinta métrica - Nivel de Detalhe - é a avaliação através da contagem de *tokens* em cada pergunta e resposta. Esta métrica permite ter a noção da profundidade e da própria especificidade do conteúdo, sendo útil para ajustar o equilíbrio entre clareza e detalhe das perguntas.

A sexta e última métrica - Fluência - é avaliada através do GPT-4 para classificar as questões numa escala de 1 a 5, considerando aspectos como coerência e clareza, como representado na Tabela 6. Utiliza-se esta métrica para garantir que as questões geradas sejam compreensíveis e bem estruturadas, tornando mais facil a compreensão da pergunta pelo modelo.

É importante destacar que esta metodologia empírica combina métricas quantitativas com avaliação qualitativa, de forma a proporcionar uma visão holística da qualidade e eficácia do sistema.

Com esta metodologia implementada, conseguiu-se fazer avaliações relativamente à qualidade das interações. Os níveis de cobertura apresentaram uma variação significativa, com pontuações de 1 a 5, indicando a capacidade do sistema em identificar perguntas que podem ser respondidas com base no contexto fornecido e qual o tipo de perguntas e respostas. A relevância, tanto contextual quanto global, mostrou-se uma métrica importante que permitiu distinguir entre perguntas altamente pertinentes (pontuação 5) e aquelas com pouca ou nenhuma relevância para o domínio da certificação energética (pontuação 1). A análise da sobreposição semântica e da diversidade, utilizando métricas como a divergência Kullback-Leibler e a Distância de Movimentação de Palavras, permitiu perceber qual a coerência e variabilidade

Tabela 6: Exemplos de Pontuações de Fluência para Perguntas relacionadas à Certificação Energética em Edifícios

Pergunta	**Fluência**	**Explicação**
Como é que o U das paredes de edifícios antigos influencía a eficiência energética?	5	A pergunta é fluente, coerente e faz sentido. É clara e específica, diretamente relacionada ao impacto das características construtivas na eficiência energética, um tema central na certificação de edifícios.
Por que razão o cálculo da inércia térmica pode ser complicado?	3	A pergunta é coerente e faz sentido, mas carece de contexto. Não é claro em que situação o cálculo da inércia térmica seria complicado, o que pode cria uma incerteza sobre o objetivo da pergunta.

Tabela 7: Exemplos de Análise de Diversidade através da Distância de Movimentação de Palavras (WMD) para Perguntas de Certificação Energética

Pergunta	**Dist. WMD**	**Explicação**
Qual é o U para paredes de edifícios anteriores a 1960?	0.85	A distância WMD relativamente alta sugere que a pergunta pode divergir bastante de outras questões relacionadas, abordando um aspecto específico e técnico do tema.
Como a espessura das paredes afeta o U em edifícios antigos?	0.25	A baixa distância WMD indica que as perguntas são semanticamente muito próximas, abordando praticamente o mesmo tema com uma leve variação na formulação, o que sugere menor diversidade.

das perguntas geradas. A avaliação da fluência revelou-se particularmente útil para garantir que as perguntas eram ou não facilmente compreendidas.

Em conjunto, estas métricas ofereceram uma visão global da qualidade das perguntas e respostas, permitindo uma análise qualitativa do desempenho do sistema RAG no contexto específico da certificação energética em Portugal. É possível afirmar pelos resultados que o sistema é capaz de gerar perguntas relevantes e

bem fundamentadas na maioria dos casos, e que também estão relacionadas com a qualidade da pergunta.

4.1.3 *Metodologia de Análise para a Avaliação de Respostas no Sistema RAG*

A avaliação de respostas geradas pelo modelo *Gemma 7b* - modelo usado neste projeto para responder a questões relativamente à certificação energética de edifícios - apresenta alguns desafios devido à tendência que as respostas têm para serem extensas. A complexidade e especificidade do domínio exigem uma abordagem de avaliação completa, assente em 5 princípios de análise que simulem a interação do ser humano (Adlakha et al., 2024).

Alguns estudos científicos recentes mostraram que é possível implementar os LLM como avaliadores devido ao facto de existir uma alta concordância com os humanos. Os LLM podem ser aplicados diretamente ou então combinados com outras técnicas para avaliar respostas a perguntas com ou sem contexto (Min et al., 2023).

Para esta abordagem, utilizou-se uma metodologia semelhante à metodologia de análise para avaliar as perguntas colocadas pelo utilizador.

Na abordagem de análise das respostas do sistema, utilizou-se o *GPT-4*, com anexação do manual de certificação energética como base de conhecimento externa, introduzindo o seguinte *prompt*: "Considera que és um técnico especializado em certificação energética de edifícios. A tua vasta carreira conta com 25 anos de experiência. Utiliza o documento anexado como fonte de inspiração primária. Analisa cada resposta introduzida de acordo com a coerência da resposta dada pelo LLM. Tem em consideração que a linguagem utilizada é PT. Segue uma linha de raciocínio lógica. O parâmetro de temperatura é 0. Sê objetivo. Não deve existir criatividade na resposta dada."

Desta forma, fez-se uma avaliação da Coerência da resposta. À semelhança das perguntas, a Coerência mede se a resposta segue uma estrutura lógica e bem articulada, considerando a pergunta e o contexto fornecido. Uma pontuação alta (5) indica que a resposta é clara e lógica, essencial para garantir que as informações sobre certificação energética sejam compreendidas, como pode ser observado na Tabela 8.

Tabela 8: Exemplos de Coerência (1 e 5) para respostas a perguntas de certificação energética.

Pergunta	**Pontuação de Coerência**	**Resposta**
Qual é a importância da ventilação adequada em edifícios de serviços?	5	A ventilação adequada em edifícios de serviços é crucial para garantir a qualidade do ar interior, prevenir a acumulação de poluentes e assegurar o conforto térmico dos ocupantes. Além disso, uma boa ventilação contribui para a eficiência energética ao otimizar o uso de sistemas de climatização.
Quais são as vantagens do uso de isolamento térmico na fachada de edifícios?	1	O isolamento térmico nas fachadas aumenta a quantidade de luz natural nos edifícios, o que melhora a saúde dos ocupantes.

Foram também feitas avaliações através da métrica Relevância. A Relevância avalia o quão bem a resposta atende aos principais aspetos da pergunta, com base no contexto dado. Uma resposta com pontuação 5 aborda diretamente o que foi perguntado, fornecendo informações úteis. Os resultados podem ser observados na Tabela 9.

Tabela 9: Exemplos de Relevância (1 e 5) para respostas a perguntas de certificação energética.

Pergunta	**Pontuação de Relevância**	**Resposta**
Quais documentos são necessários para a emissão de um certificado energético?	5	Para a emissão de um certificado energético, são necessários vários documentos, incluindo a caderneta predial urbana, a ficha técnica da habitação e a planta de arquitetura. Esses documentos ajudam a caracterizar o edifício de acordo com as normas vigentes.
Qual é o impacto da certificação energética na valorização de imóveis?	1	A certificação energética ajuda a reduzir o consumo de água no edifício, o que pode aumentar a eficiência geral do imóvel.

A terceira métrica usada, à semelhança das perguntas, foi a análise da Fundamentação. A Fundamentação verifica se a resposta segue logicamente a partir das informações fornecidas e se está bem justificada. Uma resposta bem fundamentada (pontuação 5) é essencial para garantir que as respostas dadas ao utilizador final são fidedignas e podem ser seguidas uma vez comprovada a integridade. Os resultados da análise estão representados na Tabela 10.

Tabela 10: Exemplos de Fundamentação (1 e 5) para respostas a perguntas de certificação energética

Pergunta	**Pontuação de Fundamentação**	**Resposta**
Quais são os fatores críticos para a eficiência energética em edifícios novos?	5	A eficiência energética em edifícios novos depende de diversos fatores, incluindo o isolamento térmico da envolvente, a utilização de sistemas de climatização eficientes, e a integração de energias renováveis. A conformidade com as normas estabelecidas no Decreto-Lei n.º 101-D/2020 também é essencial para alcançar uma classificação energética elevada.
Quais são os requisitos para a instalação de painéis solares em edifícios existentes?	1	A instalação de painéis solares aumenta o conforto térmico dos edifícios ao melhorar a circulação do ar.

4.2 ANÁLISE COMPARATIVA DE RAG VS MÉTODOS TRADICIONAIS

A implementação de sistemas de RAG com LLM representa uma abordagem disruptiva para superar as limitações dos métodos tradicionais de *fine-tunning*.

Nesta Secção é apresentada uma análise comparativa entre RAG e *fine-tuning*, com foco nas vantagens observadas durante o decorrer do projeto. Muitas das análises estão assentes sobre conhecimentos empíricos adquiridos no desenvolvimento, tendo como principal objetivo a máxima descrição desse conhecimento nesta Secção.

Para uma comparação que fosse fidedigna, criou-se um conjunto de 10 perguntas totalmente distintas das anteriores, e registaram-se respostas dadas sem contexto, ou seja, apenas o LLM *Gemma 7b* com o *fine-tunning* inicial e com contexto, ou seja, com o sistema RAG implementado. Esta abordagem foi definida empiricamente e todo o conhecimento dos processos de certificação foi baseado na compreensão das regras e metodologias escritas no manual. Desta forma, é importante destacar que os resultados estão assentes sobre a interpretação do manual, sendo esta uma análise qualitativa apenas comparativa. O intuito desta abordagem empírica é precisamente compreender se existe ou não vantagem na utilização do sistema RAG em relação

ao LLM com o *fine-tunning* inicial, tendo as respostas e ambos os cenários das perguntas sido registadas na Tabela 11.

Pergunta 1	**O que é o Coeficiente de Transmissão Térmica (U) e qual a sua importância na certificação energética de edifícios em Portugal?**
Resposta sem contexto	O U é uma medida da eficiência do isolamento térmico de uma estrutura ou elemento construtivo, como paredes, tetos, pisos ou janelas. É definido como a taxa de perda de calor através de uma unidade de área, em watts por metro quadrado (W/m^2), para manter uma diferença de temperatura de 1 Kelvin entre o interior e o exterior.
Resposta com contexto	O U é um indicador fundamental na certificação energética de edifícios, pois mede a quantidade de calor que atravessa um metro quadrado de um elemento construtivo por unidade de tempo, por cada grau de diferença de temperatura entre o interior e o exterior. Em Portugal, este parâmetro é crucial para avaliar a eficiência energética e determinar os requisitos de isolamento térmico.
Pergunta 2	**Como se calcula o U de uma parede composta por várias camadas de materiais diferentes em Portugal?**
Resposta sem contexto	Para calcular o U de uma parede composta por várias camadas, use a fórmula Ug = (1 / (1 / U1 + S1 / (k2 * d2) + S2 / (k3 * d3) + ... + Sn / (kn * dn))), onde Ug é o coeficiente global, U1 é o coeficiente da face interna, ki é a condutividade térmica do material na posição i, di é a espessura da camada de material na posição i, e Si é a transmitância térmica da interface entre as camadas.
Resposta com contexto	O cálculo do U de uma parede composta por várias camadas é feito somando as resistências térmicas de cada material que compõe a parede, usando a fórmula: (1 / U1 + S1 / (k2 * d2) + S2 / (k3 * d3) + ... + Sn / (kn * dn))). Depois, o inverso da soma das resistências dá o valor de U. Em Portugal, este cálculo deve incluir também as resistências superficiais interna e externa, importantes para estimar a transmissão total de calor.
Pergunta 3	**Quais são os requisitos mínimos de U para paredes exteriores em edifícios novos na zona climática I1 em Portugal?**

Resposta sem contexto	Para edifícios novos Portugal, o valor U para paredes exteriores deve ser inferior a 0,3 W/(m².°C). Este valor é padrão para todas as zonas climáticas do país.
Resposta com contexto	Em edifícios novos localizados na zona climática I1, que corresponde às zonas a sul de Portugal, o valor máximo de U para paredes exteriores é de 0,50 W/(m².ºC). Estes valores variam entre 0,80 e 0,40 consoante as condições de fronteira, e consoante as zonas climáticas.
Pergunta 4	**Como é que a inércia térmica de um edifício afeta o seu desempenho energético e como é classificada no processo de certificação em Portugal?**
Resposta sem contexto	A inércia térmica afeta a capacidade do edifício de manter uma temperatura constante. Em Portugal, é classificada como baixa ou alta, dependendo apenas da espessura das paredes do edifício.
Resposta com contexto	A inércia térmica refere-se à capacidade de um edifício de armazenar e libertar calor, mantendo uma temperatura interior mais estável. Em Portugal, é avaliada com base na massa térmica dos materiais e sua distribuição no edifício. Edifícios com alta inércia térmica têm melhor desempenho energético, especialmente em climas mais quentes, pois retardam a transferência de calor, minimizando os picos de consumo de energia.
Pergunta 5	**Explique o conceito de pontes térmicas lineares e como são consideradas no cálculo do U médio de uma fachada em Portugal.**
Resposta sem contexto	Pontes térmicas lineares são áreas onde o calor escapa mais facilmente do edifício. Em Portugal, são geralmente ignoradas no cálculo do U médio de uma fachada, pois o seu impacto é considerado insignificante.
Resposta com contexto	Pontes térmicas lineares ocorrem em junções construtivas, como cantos e ligações entre paredes e janelas, onde há uma maior facilidade na perda de calor. No cálculo do U médio de uma fachada em Portugal, estas pontes são devidamente contabilizadas e podem ter um impacto significativo na eficiência energética global do edifício, sendo necessário minimizá-las.
Pergunta 6	**Quais são os critérios para a determinação da zona climática de inverno de um edifício em Portugal e como isso afeta os requisitos de isolamento térmico?**

Resposta sem contexto	Em Portugal, as zonas climáticas de inverno são determinadas apenas pela latitude do local. Quanto mais a norte, mais rigorosos são os requisitos de isolamento, mas a diferença é mínima entre as zonas.
Resposta com contexto	A determinação da zona climática de inverno em Portugal tem em conta a localização geográfica (latitude, altitude) e a severidade climática. Existem três zonas principais: I1, I2 e I3. Edifícios situados em zonas mais frias (como I3) devem cumprir requisitos mais rigorosos de isolamento térmico para garantir uma maior eficiência energética.
Pergunta 7	**Como se calcula o fator solar de um vão envidraçado com proteção solar ativada a 100% e qual a sua importância na certificação energética em Portugal?**
Resposta sem contexto	Em Portugal, o fator solar de um vão envidraçado com proteção é simplesmente a soma do fator solar do vidro e da proteção. A sua importância na certificação é mínima, pois afeta apenas o conforto visual.
Resposta com contexto	O fator solar de um vão envidraçado com proteção solar ativa a 100% é calculado de forma diferente para vidro simples ou duplo. A equação base multiplica o fator solar do vidro pelo i é um índice específico relacionado à proteção solar, conforme descrito no Despacho n.º 15793-K/2013. Este valor é crucial na certificação energética, pois contribui para o desempenho térmico dos edifícios, principalmente em climas quentes.
Pergunta 8	**Explique o conceito de taxa de renovação do ar (RPH) e como é determinada para fins de certificação energética em edifícios residenciais em Portugal.**
Resposta sem contexto	A taxa de renovação do ar (RPH) é o número de vezes que o ar de um espaço é completamente substituído por hora. Em edifícios residenciais em Portugal, é sempre fixa em 1 renovação por hora, independentemente das características do edifício.
Resposta com contexto	A taxa de renovação do ar (RPH) indica o número de vezes que o ar interior é renovado numa divisão ao longo de uma hora. Para certificação energética em edifícios residenciais, em Portugal, a taxa de renovação mínima recomendada varia entre 0,5 a 1,0 RPH, dependendo das condições de ventilação natural ou mecânica, para garantir a qualidade do ar e a eficiência energética.

Pergunta 9	**Como são considerados os sistemas de aproveitamento de fontes de energia renováveis no cálculo do desempenho energético de um edifício para certificação em Portugal?**
Resposta sem contexto	Em Portugal, os sistemas de energia renovável são sempre considerados como uma redução direta de 50% no consumo energético do edifício, independentemente do tipo ou capacidade do sistema instalado.
Resposta com contexto	Os sistemas de aproveitamento de fontes de energia renováveis, como painéis solares ou bombas de calor, são incorporados no cálculo do desempenho energético de um edifício em Portugal, resultando numa redução proporcional ao contributo energético. A percentagem de redução depende do tipo e da eficiência do sistema instalado.
Pergunta 10	**Descreva o processo de determinação da classe energética de um edifício de habitação em Portugal, incluindo os fatores considerados e as etapas de cálculo.**
Resposta sem contexto	Em Portugal, a classe energética de um edifício de habitação é determinada apenas pelo seu consumo anual de energia elétrica. O processo envolve dividir este consumo pela área útil do edifício e comparar o resultado com uma tabela padrão nacional.
Resposta com contexto	A classe energética de um edifício em Portugal é determinada através de uma avaliação global do consumo de energia para aquecimento, arrefecimento, ventilação, e produção de água quente sanitária. Através do cálculo da necessidade energética do edifício por metro quadrado, compara-se o resultado com valores de referência, sendo atribuída uma classe entre A+ e F.

Tabela 11: Perguntas-resposta de comparação entre LLM original vs Sistema RAG

Após o trabalho comparativo de resposta sem contexto e resposta com contexto, surgiu a necessidade de criar uma metodologia empírica para comparar a qualidade das respostas fornecidas pelo sistema RAG e pelo LLM apenas com *fine-tuning* inicial. Desenvolveu-se assim uma metodologia de análise empírica baseada em quatro critérios principais: precisão, integridade, relevância e clareza. Cada critério é avaliado numa escala de 1 a 5, sendo 5 o valor mais alto de pontuação, ou seja, que apresenta um melhor resultado.

Os critérios utilizados, baseiam-se em:

- Precisão (P): Este critério avalia a exatidão da informação fornecida na resposta. Uma pontuação alta indica que a resposta contém informações corretas e alinhadas com o manual de certificação energética disponibilizado pela ADENE.
- Integridade (I): Avalia o grau em que a resposta aborda todos os aspetos relevantes da pergunta. Uma pontuação alta significa que a resposta é abrangente e não omite informações importantes.
- Relevância (R): Mede o quão bem a resposta se relaciona com a pergunta feita. Uma pontuação alta indica que a resposta é diretamente aplicável à questão, sem divagações ou informações desnecessárias.
- Clareza (C): Avalia a facilidade de compreensão da resposta. Uma pontuação alta sugere que a resposta é bem estruturada, utiliza linguagem apropriada e explica conceitos de forma compreensível.

Tabela 12: Avaliação comparativa das perguntas com base nas respostas sem contexto e com contexto.

Pergunta	**Tipo**	**P (1-5)**	**I (1-5)**	**R (1-5)**	**C (1-5)**	**Pontuação Total (4-20)**
1	Sem contexto	3	3	4	4	14
	Com contexto	4	4	5	4	17
2	Sem contexto	4	4	4	4	16
	Com contexto	5	5	5	5	20
3	Sem contexto	1	1	3	2	7
	Com contexto	3	3	4	4	14
4	Sem contexto	3	3	4	3	13
	Com contexto	4	4	5	4	17
5	Sem contexto	2	2	4	3	11
	Com contexto	4	4	4	3	15
6	Sem contexto	2	3	3	3	11
	Com contexto	3	4	4	4	15
7	Sem contexto	4	4	4	4	16
	Com contexto	5	5	5	5	20
8	Sem contexto	3	3	4	3	13
	Com contexto	4	4	5	4	17
9	Sem contexto	3	3	4	3	13
	Com contexto	4	4	5	4	17
10	Sem contexto	1	1	2	2	6
	Com contexto	3	3	4	3	13

Durante a análise comparativa, tiveram-se em consideração alguns pontos importantes na constituição das respostas com e sem contexto. Mais uma vez, reforça-se que devido ao facto de ser uma análise empírica, assenta na interpretação que é feita à resposta dada. No entanto, tendo sido a resposta avaliada sempre pelo humano responsável pela metodologia, ignorou-se a possibilidade de existir viés ou limitações de conhecimento relativamente ao tema. Uma vez tendo sido feito sempre pelo mesmo interveniente, considerou-se que podem ser desprezados os fatores externos que não são possíveis de controlar de forma direta. As perguntas e respostas comparativas foram registadas na Tabela 12.

Assim sendo, começou-se pela análise da "Pergunta 1: O que é o U e qual a sua importância na certificação energética de edifícios em Portugal?".

Na resposta sem contexto, é mencionado corretamente o conceito básico do U, dando a justificação de que mede a quantidade de calor que passa através de um material, o que é tecnicamente preciso. No entanto, omite detalhes importantes, como a consideração de resistências superficiais internas e externas, e as diferentes formas de aplicação para a envolvente opaca e envidraçada. Por essa razão, a precisão foi avaliada com 3/5. Quanto à integridade, a resposta é superficial, não entrando em grandes detalhes, resultando numa pontuação de 3/5. Em termos de relevância, a questão é central para a certificação energética, justificando uma avaliação de 4/5. Finalmente, a clareza é satisfatória, mas a falta de componente técnica pode dificultar a compreensão de quem não está a par do assunto, querendo obter um conhecimento mais profundo, atribuindo-se assim uma avaliação 4/5.

Já na resposta dada com contexto, é abordada de forma mais detalhada relativamente ao papel do coeficiente U. Refere-se na resposta a importância na avaliação da eficiência energética. Pode-se assumir que a precisão é maior, resultando numa pontuação de 4/5, uma vez que inclui as resistências térmicas internas e externas mencionadas no manual SCE. A integridade é superior (4/5), pois oferece mais informações sobre as aplicações do coeficiente, mas ainda poderia explorar melhor os tipos de elementos construtivos. A relevância é elevada (5/5), já que o tema está completamente alinhado com a certificação energética e com o contexto da pergunta. A clareza melhora, com uma explicação técnica, mas ainda acessível, o que justifica uma pontuação de 4/5.

Desta forma, registou-se um aumento na pontuação total da resposta sem contexto para a resposta com a implementação do sistema RAG.

Na "Pergunta 2: Como se calcula o U de uma parede composta por várias camadas de materiais diferentes em Portugal?", registou-se também melhorias.

Na resposta sem contexto, é descrito de maneira correta o processo de cálculo do coeficiente U. Fala sobre o processo da soma das resistências térmicas de cada camada, atribuindo-se um 4/5. A integridade também foi pontuada com 4/5, uma vez que a resposta aborda os conceitos gerais, mas não oferece uma explicação técnica detalhada sobre como calcular o valor U para diferentes tipos de materiais. Em termos de relevância, a questão é muito importante para o processo de certificação energética, atribuindo-se 4/5. A clareza é boa, com uma linguagem acessível, mas falta alguma profundida para quem precisa de uma aplicação prática, resultado numa avaliação 4/5.

Já na resposta com contexto, pode-se considerar uma maior precisão visto que aborda o cálculo com o detalhe das resistências superficiais internas e externas. Assim, a precisão foi avaliada com 5/5. A integridade pode-se considerar excelente, uma vez que tem a descrição técnica exata e todos os passos necessários para calcular o coeficiente U, justificando uma pontuação de 5/5. Considerou-se também a resposta bastante relevante, já que este cálculo é essencial no contexto da certificação energética atribuindo-se um 5/5. A clareza também foi considerada ótima, com uma explicação técnica bem detalhada e acessível mesmo a quem não tenha algum conhecimento prévio no assunto, 5/5.

Na "Pergunta 3: Quais são os requisitos mínimos de U para paredes exteriores em edifícios novos na zona climática I1 em Portugal?"houve também impacto.

A resposta sem contexto apresenta um valor incorreto de 0,3 W/(m².ºC), que não é aplicável à zona climática I1, uma vez que o valor mínimo correto é 0,40 W/(m².ºC). Isto afeta a precisão, que foi pontuada com 1/5. A integridade é comprometida uma vez que não considera as variações entre as várias condições de fronteira exteriores ou interiores, tendo sido classificada com 1/5. Uma vez que a resposta se manteve relativa ao tema da a certificação energética, atribuiu-se uma avaliação de 3/5 para relevância. Em termos de clareza, embora a explicação seja direta, o erro no valor prejudica a compreensão correta da informação, resultando numa pontuação de 2/5.

Na resposta com contexto, relativamente à pergunta 3, o valor do U para a zona I1 foi corrigido subindo para a precisão para 3/5. A integridade também melhora para 3/5 uma vez que considera as variações das condições de fronteira, mas ainda não é perfeito por ter sugerido um valor de referencia e não o ter indicado. A relevância é alta, dado que a resposta está mais alinhada com o conteúdo do Manual SCE, atribuindo-se 4/5. A resposta considerou-se clara, mas com espaço para conteúdo mais técnico para garantir maior compreensão 4/5. Este ponto pode também ser

melhorado com o *prompt* usado, ou seja, com uma pergunta mais especifica por parte do utilizador.

Na "Pergunta 4: Como é que a inércia térmica de um edifício afeta o seu desempenho energético e como é classificada no processo de certificação em Portugal?"houve um aumento significativo na qualidade.

A resposta sem contexto define a inércia térmica de maneira geral, sem entrar em grandes detalhes sobre o impacto técnico, atribuindo-se numa precisão de 3/5. A integridade avaliou-se com 4/5 uma vez que, apesar de boa, não inclui informações sobre os critérios específicos usados para classificar a inércia térmica em Portugal, como a massa térmica dos materiais. A relevância da resposta é moderada, uma vez que a questão é importante, mas a resposta não reflete toda a sua complexidade 3/5. A clareza foi avaliada com 3/5, pois a resposta é bastante simples para o tema em questão.

Na resposta com contexto, aborda-se a inércia térmica com mais detalhes técnicos, destacando-se a importância da massa térmica relativamente ao desempenho energético, atribuindo-se um valor de 4 em 5 à precisão. A integridade tem a mesma magnitude de 4/5, visto que a resposta inclui a classificação da inércia térmica. A relevância é alta, dando-se 5/5, pois a inércia térmica é um fator crítico na certificação energética e está alinhada com a pergunta. Já a clareza, também se notou melhorias com uma explicação mais técnica, mas compreensível, tendo recebido 4/5.

Na "Pergunta 5: Explique o conceito de pontes térmicas lineares e como são consideradas no cálculo do U médio de uma fachada em Portugal."

Na resposta sem contexto, considerou-se que a explicação estava correta acerca dos conceitos básico de pontes térmicas. No entanto, cometeu-se o erro de sugerir que as pontes térmicas lineares são ignoradas no cálculo do U médio. Este ponto reduz a precisão para 2/5. A integridade é igualmente comprometida, já que não se menciona a importância da contabilização correta das pontes térmicas 2/5. A relevância é adequada, dado que o conceito é importante, mas a resposta falha em destacar a importância deste ponto no contexto prático, resultando em 4/5. A clareza é considerada média, com 3/5, uma vez que a explicação simplifica em demasia um conceito técnico.

Já na resposta com contexto, registou-se um aumento na precisão para 4 pontos, uma vez que se considerou que as pontes térmicas lineares foram devidamente contabilizadas no cálculo do coeficiente global. A integridade melhora significativamente, 4/5, também pela mesma razão da relevância das pontes térmicas e o impacto que estes elementos têm no desempenho energético. A relevância é igualmente alta,

4/5, refletindo o papel das pontes térmicas na certificação energética. A clareza também melhora para 3/5 pontos, uma vez que se considerou uma resposta de fácil de entendimento, ainda que possa ter sido pouco claro na fase inicial, em que fala de "junções construtivas"não mencionando as junções criticas.

De seguida, classificou-se a "Pergunta 6: Quais são os critérios para a determinação da zona climática de inverno de um edifício em Portugal e como isso afeta os requisitos de isolamento térmico?"

Na resposta sem contexto, é abordada a latitude como critério para a determinação da zona climática, mas omite outros fatores importantes, como a altitude e a severidade climática. Isto faz com que se tenha atribuído 2 pontos em 5 à precisão. A integridade também é prejudicada, 3/5, pois não menciona as variações entre diferentes zonas e a influência de fatores adicionais, como a proximidade ao mar. A relevância é adequada, considerou-se 3/5, já que a questão é importante, mas a resposta não cobre os aspetos necessários para a certificação. A clareza é razoável, mas ainda assim faltam de detalhes técnicos que podem afetar a compreensão da resposta 3/5.

Na resposta com contexto, já são mencionados fatores que se consideram importantes, como a altitude e a severidade climática, dando um aumento na precisão para 3/5. A integridade aumentou também para 4 pontos, já que a resposta já foca numa maior gama de critérios, embora ainda possa incluir mais detalhes sobre como os diferentes fatores influenciam os requisitos de isolamento. A relevância é alta (4/5), refletindo a importância da questão para a certificação. A clareza é boa (4/5), com uma explicação técnica, mas ainda compreensível.

Na "Pergunta 7: Como se calcula o fator solar de um vão envidraçado com proteção solar ativada a 100% e qual a importância na certificação energética em Portugal?"

Na resposta sem contexto, é mencionado corretamente a definição de fator solar, mas não explica como é calculado, dando-se assim uma precisão de 1/5. A integridade é fraca, uma vez que poderia incluir mais detalhes sobre como o fator solar, como a forma de cálculo e a maneira como afeta o desempenho térmico do edifício, 2/5. A relevância é alta, já que o fator solar é um componente crítico na certificação (4/5). A clareza é boa, mas poderia ser mais técnica (4/5).

Na resposta com contexto, já existe uma explicação detalhada sobre o cálculo do fator solar, que explica como deve ser feito o processo dando uma de precisão 5/5. A integridade é também considerada excelente, 5/5, já que todos os elementos considerados importantes são abordados. Existe ainda total relevância, dada a

importância do fator solar na certificação energética, com máxima avaliação 5/5. A clareza é ótima, 5/5, com uma explicação técnica considerada super clara.

Na "Pergunta 8: Explique o conceito de taxa de renovação do ar Taxa de Renovação do Ar (RPH) e como é determinada para fins de certificação energética em edifícios em Portugal."avaliaram-se bons resultados também.

Na abordagem sem contexto, define-se corretamente a RPH como o número de vezes que o ar de um espaço é completamente substituído por hora. No entanto, não se abordam detalhes importantes, como as variações em função das condições de ventilação natural e mecânica, tendo sido atribuída uma precisão de 3/5. Em termos de integridade, a resposta não abrange os métodos específicos de cálculo da taxa de renovação, como descrito no manual SCE, justificando uma avaliação de 3/5. A relevância é elevada, com 4 valores em 5, uma vez que se regista a resposta fortemente relacionada com a pergunta. A clareza, por sua vez, foi considerada moderada 3/5, já que a resposta é simples, mas carece de detalhes técnicos importantes.

Já no cenário com contexto, registaram-se melhorias, nomeadamente ao abordar a importância da ventilação natural e mecânica no cálculo da taxa de renovação do ar, aumentando a precisão para 4/5. A integridade também melhora 4/5, ao fornecer uma explicação mais técnica sobre como os diferentes fatores influenciam a taxa de renovação, e regista os intervalos. Considerou-se uma relevância elevada, 5/5, uma vez que a resposta reflete a importância do conceito da pergunta no contexto da certificação energética. A clareza é melhorada também de 4/5, com uma explicação mais técnica e detalhada mantendo-se acessível para os utilizadores.

Na "Pergunta 9: Como são considerados os sistemas de aproveitamento de fontes de energias renováveis no cálculo do desempenho energético de um edifício para certificação em Portugal?", fez-se uma pergunta de procedimento e registaram-se também melhorias na abordagem do sistema RAG em relação ao LLM apenas com o *fine-tunning* inicial.

Na abordagem sem contexto resposta sem contexto menciona que os sistemas de energia renovável são sempre considerados como uma redução direta no consumo energético do edifício, mas não especifica que essa redução depende da quantidade de energia gerada e do tipo de sistema renovável utilizado. Isso afeta a precisão, e decidiu-se atribuir a pontuação de 3/5. A integridade também é prejudicada 3/5, visto não aborda as tecnologias renováveis que existem. A relevância é adequada 4 pontos em 5, uma vez que a utilização de fontes de energia renovável é uma questão central na certificação energética. A clareza é razoável (3/5), mas a simplificação excessiva prejudica a compreensão do impacto real desses sistemas.

Já na resposta com contexto, é registada uma melhoria. Começa por destacar as tecnologias renováveis mais comuns,e explica como cada uma afeta o consumo de energia do edifício, resultando numa precisão de 4/5. A integridade também aumenta 4/5, ao destacar a redução proporcional que estas tecnologias oferecem no desempenho energético. A relevância é considerada alta, 5/5, uma vez que fora consideradas as fontes de energia, como abordado na pergunta. A clareza é aprimorada 4/5, com uma explicação mais detalhada e técnica, mas ainda compreensível para quem tenha familiaridade com o tema.

Para terminar, na última pergunta "Pergunta 10: Descreva o processo de determinação da classe energética de um edifício de habitação em Portugal, incluindo os fatores considerados e as etapas de cálculo.", também se registou melhorias.

Na resposta sem contexto, o processo de determinação da classe energética está maioritariamente errado, mencionando o consumo de eletricidade, isto dá uma precisão de 1/5, uma vez que o processo de auditoria é diferente.A integridade é igualmente limitada, 1/5, visto que não pode ser uma resposta considerada de acordo com o tema. A relevância é baixa, 2/5, pois a questão é altamente relevante, mas a resposta não reflete a realidade. A clareza também é prejudicada 2/5, devido ao erro e simplificação excessiva que não transmite a totalidade do processo.

Já na resposta com contexto, alguns dos problemas são corrigidos uma vez que é mencionado o sistema de avaliação, referindo-se ao consumo de energia primária, os sistemas de climatização, ventilação e fontes renováveis. Isto eleva a precisão para um valor de 3 em 5, já que a resposta poderia ainda explorar melhor alguns aspetos específicos do processo de certificação, como a análise das envolventes ou até mesmo a análise de elementos horizontais e verticais opacos. A integridade é melhor, 3/5, ainda que a resposta não se refira ao processo de forma suficientemente detalhada para possibilitar o utilizador de ficar com uma ideia clara. A relevância é elevada 4/5, dado que a questão é adequada ao tema da certificação energética. A clareza melhora, 3/5, mas a explicação técnica ainda podia ser mais detalhada em termos de pontos discutidos. É claro, que nestas avaliações, tenta-se fazer uma abordagem generalista, mas tocando em temas específicos. Alguns destes resultados, poderiam certamente ser melhorados com afinações das próprias perguntas, como referido anteriormente.

Para uma comparação mais intuitiva, criou-se o gráfico representado na Figura 4, onde é possível comparar de forma direta, as pontuações gerais em ambos os contextos.

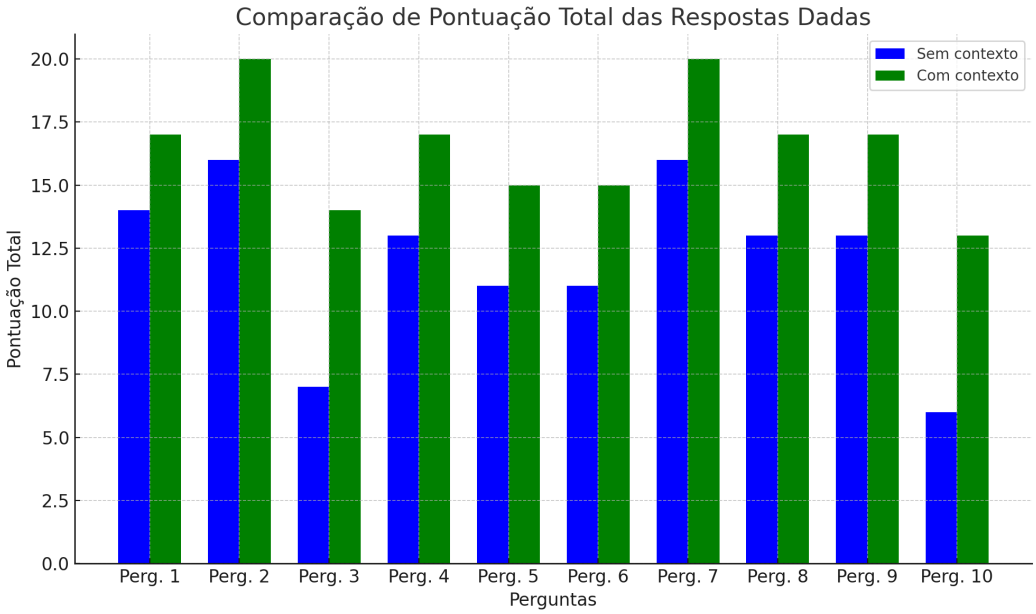

Figura 4: Comparação de pontuação total das respostas dadas

Ainda se registaram também os tempos de resposta, para cada um dos cenários. Estas variações registaram-se no gráfico 5.

Conclui-se que entre o sistema RAG (com contexto) e o método tradicional (sem contexto) existem melhorias significativas tanto na qualidade das respostas quanto nos tempos de processamento. O gráfico de pontuação total das respostas evidencia um aumento consistente na qualidade quando o contexto é fornecido, com diferenças particularmente notáveis nas perguntas 3, 5, 6, 8, 9 e 10, onde a pontuação com contexto é substancialmente superior. Esta melhoria indica que o sistema RAG é capaz de fornecer respostas com maior integridade, relevantes e completas. Quanto aos tempos de resposta, registados na Figura 5, observa-se um aumento no tempo de processamento quando o contexto é utilizado, o que é esperado devido à necessidade de recuperação e integração de informações adicionais. No entanto, este aumento é relativamente pequeno, variando entre 0,5 e 0,7 segundos em média, o que sugere que o sistema mantém uma eficiência operacional aceitável mesmo com a adição do processamento contextual.

Este desempenho foi avaliado por um perito qualificado na área da certificação energética, com a credencial de PQ1 (Perito Qualificado de Nível 1). Este especialista, que trabalha diariamente com a avaliação de desempenho energético de edifícios e na emissão de certificados energéticos, possui o *know-how* necessário para garantir que a avaliação está alinhada com as exigências normativas em Portugal. A parceria do projeto com este perito foca-se em assegurar que as melhorias observadas, com o

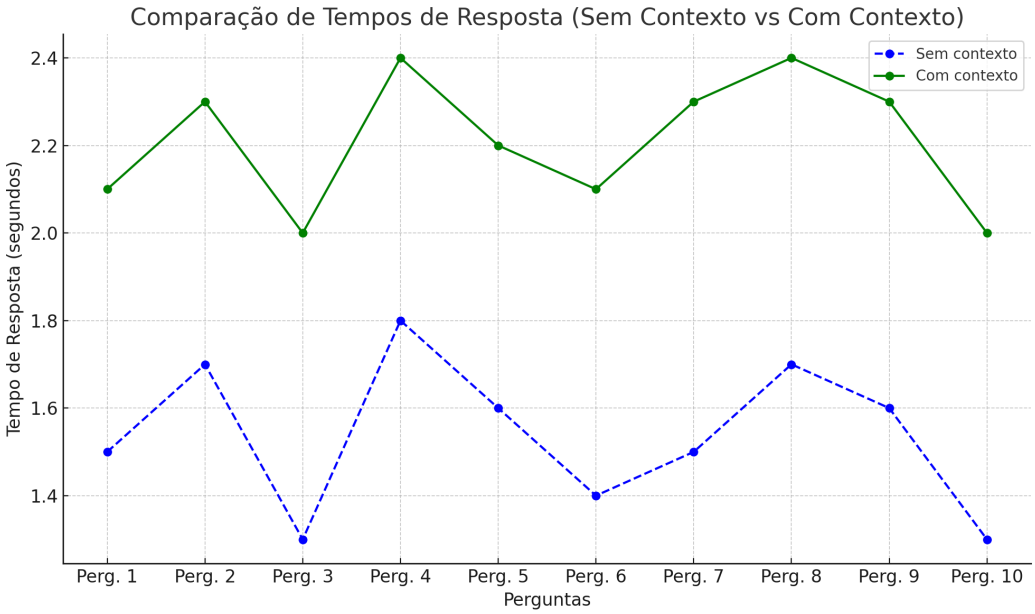

Figura 5: Comparação de tempos de resposta em cenário sem contexto vs cenário com contexto

uso do sistema RAG, sejam interpretadas com precisão e relevância no contexto prático da certificação energética, evitando que seja construída uma ferramenta que não tenha utilização prática para o utilizador final.

De forma a justificar a eficiência do sistema RAG em comparação com o fine-tuning tradicional, fez-se ainda uma análise comparativa em tempos estimados para cada uma das abordagens. Para esta analise, considerou-se o tempo que seria necessário para realizar o re-treino do modelo *Gemma 7B* com o conteúdo do manual de certificação energética. A estimativa baseou-se nos seguintes fatores:

- Tamanho do conjunto de dados: O manual de certificação energética tem 262 páginas, o que se traduz em 69,133 palavras e 441,929 caracteres.
- Recursos computacionais: Considera-se para esta estimativa o uso de uma *GPU* de alta performance, como a *NVIDIA A100* com 40GB de memória.
- Tempo de processamento: Com base nas *benchmarks* do modelos, estima-se que cada época de treino possa necessitar entre 1 a 2 horas.
- Número de épocas: Para um *fine-tuning* efetivo, considerando a especificidade e a densidade técnica do conteúdo, seriam necessárias entre 20 a 30 épocas para atingir uma convergência satisfatória e capturar os contextos do texto técnico.

- *Overhead* de preparação e avaliação: Inclui tempo para pré-processamento dos dados, *tokenização* específica do domínio, avaliações intermédias e pós-processamento.

Desta forma, o tempo ideal pode ser calculado através de:

- Tempo estimado por época: 50 minutos.
- Número de épocas: 25 (estimativa média, considerando a necessidade de capturar detalhes técnicos específicos da certificação)
- Tempo total de treino: 50 minutos × 25 épocas = 20,83 horas
- Overhead adicional: Aproximadamente 3 horas. Isto inclui tempo para *tokenização* e validação do processo.

$$\text{Tempo total estimado} = 20{,}83 \text{ horas} + 3 \text{ horas} \approx 24 \text{ horas} \tag{6}$$

Portanto, a estimativa final para o re-treino do modelo *Gemma 7B* com o conteúdo do manual de certificação energética seria de aproximadamente 24 horas, ou seja, 1 dia completo de processamento contínuo, como mostra equação 6.

Posto isto, salienta-se ainda que esta estimativa, embora baseada em dados precisos, assume condições ideais. Na prática, o processo pode estender-se devido a interrupções ou problemas técnicos durante o treino, necessidade de múltiplas execuções para otimização de hiperparâmetros, ou outro *edge case* que possa não estar previsto. Esta estimativa serve exclusivamente para a fundamentação teórica das vantagens do sistema RAG em relação ao método de *fine-tunning*.

4.2.1 *Estimativa de Implementação do Sistema RAG*

Para uma comparação direta e justa com o processo de fine-tuning do modelo Gemma 7B, que foi estimado em 24 horas de implementação, apresenta-se uma análise detalhada do tempo necessário para implementar um sistema RAG com a tecnologia de VecDB como fonte de conhecimento externa, e o modelo *Gemma 7b* como LLM utilizado.

Em condições ideais, estima-se que a implementação inicial do RAG pode ser concluída em aproximadamente 12 horas. As primeiras 2 horas são dedicadas à configuração do ambiente e preparação dos dados, incluindo a instalação das

dependências necessárias e o pré-processamento do texto do manual de certificação energética de 262 páginas. Esta fase é fundamental para garantir que os dados estejam em formato adequado para as etapas subsequentes.

Nas 3 horas seguintes, realiza-se a indexação e criação da VecDB. Este processo envolve a geração de *embeddings* para o conteúdo do manual e a sua indexação numa VecDB, como foi explicado no Capítulo 3.

A integração com o modelo de linguagem *Gemma 7b* ocupa cerca de 4 horas. Nesta fase, configura-se o modelo para uso com o RAG e implementa-se a lógica de geração de *prompts* e recuperação de informações. Esta etapa é vital para assegurar que o sistema possa gerar respostas precisas e fundamentadas pela base de conhecimento externa.

As últimas 3 horas são dedicadas a testes e otimização inicial. Realizam-se testes básicos de funcionamento e ajustes de performance para garantir que o sistema está a funcionar corretamente.

É importante realçar que depois de implementado, o sistema RAG oferece a vantagem significativa de atualização de conhecimento. Ao contrário do fine-tuning, que requereria um novo ciclo completo de treino de 24 horas para cada atualização substancial, o RAG permite incorporar novas informações ou alterações no manual em questão de minutos.

Em síntese, embora a implementação inicial do RAG leve 12 horas, comparada às 24 horas do fine-tuning, a verdadeira vantagem encontra-se na capacidade de atualização rápida e potencialmente infinita. Esta característica traduz-se numa eficiência operacional extraordinária a longo prazo.

4.3 ANÁLISE CUSTO-BENEFÍCIO DE SISTEMA RAG VS *FINE-TUNNING* LLM

Nesta abordagem comparativa de custo-benefício entre o sistema RAG e o método tradicional de *fine-tuning* para o modelo *Gemma 7b*, fez-se uma análise baseada nas estimativas de tempo de implementação, custos de recursos computacionais e eficiência operacional a longo prazo, resumida na Tabela 13.

Considerou-se que o processo de *fine-tuning* requer aproximadamente 24 horas de utilização contínua de uma *GPU* de alto desempenho, como a *NVIDIA A100*. Considerando um preço médio de mercado de 2,50€ por hora para uma *GPU*

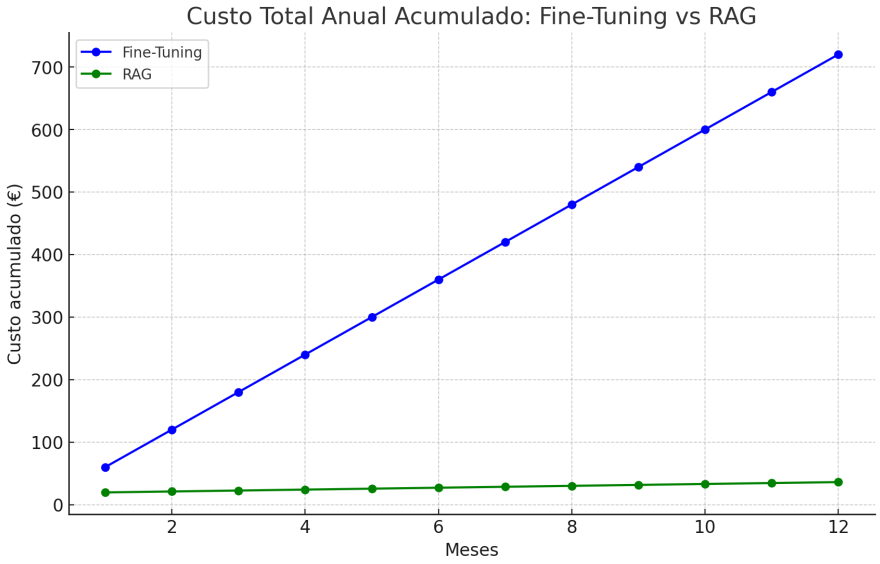

Figura 6: Custo total anual de sistema RAG vs *fine-tunning* de LLM

semelhante na *cloud*, o custo estimado de implementação inicial para o *fine-tuning* é de 60€.

Em contraste, o sistema RAG necessita de cerca de 12 horas para implementação inicial. Assumindo que pode utilizar uma GPU menos potente para parte do processo, estimou-se um custo médio de 1,50€ por hora, resultando num custo total de implementação de 18€, como se pode ver na Tabela 13.

No entanto, para atualizar o conhecimento do modelo através do *fine-tunning*, é necessário submeter um novo ciclo completo de treino de 24 horas. Assumindo a possibilidade de existirem atualizações mensais do modelo, o custo anual de atualizações seria de 720€ (12 meses x 60€).

O sistema RAG, por outro lado, permite atualizações incrementais em questão de minutos. Estimando uma média de 30 minutos por atualização semanal, e assumindo um custo de 1,5€ por hora, o custo anual de atualizações seria de aproximadamente 39€. A comparação pode ser analisada na Figura 7.

Tabela 13: Comparação de custos entre sistema RAG e fine-tuning.

Elemento	**Sistema RAG (€)**	**Fine-Tuning (€)**
Implementação Inicial	18	60
Atualização Anual	39	720
Custo Total Anual	**57**	**780**

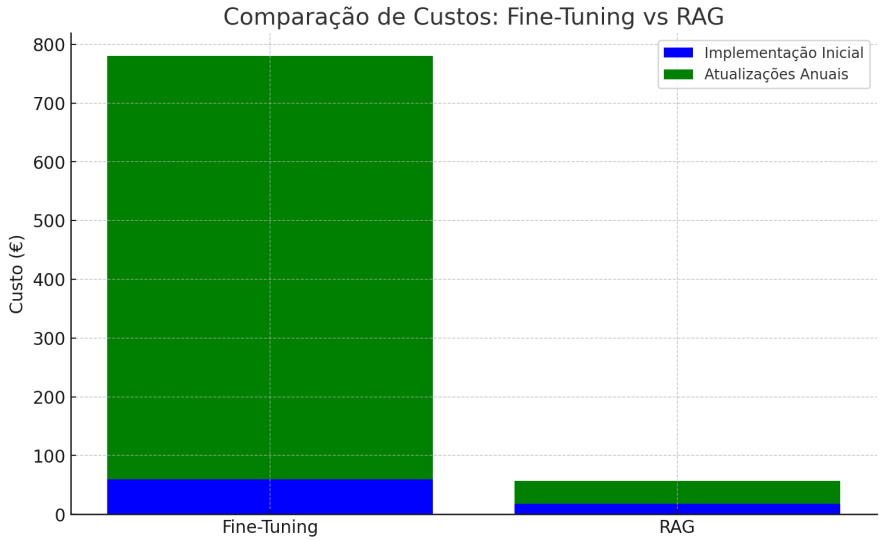

Figura 7: Comparação de custos entre sistema RAG vs *fine-tunning* de LLM

Considerando um horizonte temporal de um ano, o custo total estimado para o *fine-tuning* (implementação inicial + 12 atualizações) seria de 780€. Já para o sistema RAG, o custo total (implementação inicial + 52 atualizações semanais) seria de aproximadamente 57€.

A diferença de custo de 723€ ao longo de um ano representa uma economia significativa de 92,7% a favor do sistema RAG, além dos benefícios adicionais em termos de flexibilidade e potencial melhoria na qualidade das respostas, como está representado na Figura 6.

5

CONCLUSÕES

Durante o desenvolvimento do projeto, registaram-se resultados considerados favoráveis à implementação e avaliação de um sistema de Geração Aumentada por Recuperação (RAG) integrado com Modelos de Linguagem de Grande Escala (LLM) e Bases de Dados Vetoriais (VecDBs) no contexto da certificação energética em Portugal. A implementação bem-sucedida do sistema RAG-LLM, utilizando o modelo *Gemma 7B* e a VecDB *ChromaDB*, demonstrou melhorias substanciais na precisão e relevância das respostas geradas, como se pode avaliar pelos testes empíricos comparativos realizados no capitulo 4. O sistema de atualização de conhecimento permitiu a integração de novas informações na capacidade de resposta do modelo sem necessidade de re-treino completo do modelo, considerando-se este um aspeto crítico para áreas técnicas como a certificação energética.

Avaliou-se e provou-se ainda melhorias substanciais na precisão e relevância das respostas. A análise comparativa entre as respostas geradas com e sem o sistema RAG revelou um aumento consistente nas pontuações de precisão, integridade, relevância e clareza. Por exemplo, numa questão técnica sobre o Coeficiente de Transmissão Térmica (U), a resposta com contexto obteve uma pontuação total de 17 em 20, em comparação com 14 em 20 para a resposta sem contexto, o que prova ser uma melhoria significativa na qualidade da informação fornecida. Esta qualidade foi analisa também através de testes empíricos baseados na interpretação humana da resposta dada, comparando diretamente com o manual SCE, documento responsável pela regulamentação do processo de certificação energética em Portugal.

Durante o desenvolvimento e implementação, lidou-se com vários desafios técnicos. Um dos que mais impacto teve foi precisamente a necessidade de otimização do processo de recuperação de informações para manter tempos de resposta aceitáveis, especialmente considerando o volume e a complexidade das informações no domínio da certificação energética. Observou-se um ligeiro aumento no tempo de processamento com o uso de contexto, variando entre 0,5 e 0,7 segundos em média. Este aumento é justificável pela necessidade de recuperação e integração de informações adicionais e representa um *trade-off* aceitável, considerando a melhoria significativa na qualidade das respostas.

Do ponto de vista de custo-benefício, o sistema RAG demonstra um retorno positivo expressivo em comparação ao *fine-tuning* tradicional. Os custos iniciais de desenvolvimento e implementação do RAG, estimados em aproximadamente 12 horas de trabalho, e custos de aproximadamente 18€, são significativamente menores que as 24 horas necessárias para o fine-tuning, com custos associados de 60€. O mais impressionante é a economia a longo prazo: o custo anual total do RAG, incluindo atualizações semanais, é estimado em apenas 57€, em contraste com 780€ estimados para o fine-tuning, representando uma economia estimada de 92,7%. Conclui-se assim que do ponto de vista de custo-benefício, o sistema RAG integrado com LLM tem um retorno positivo em relação o sistema tradicional de re-treino completo de um modelo com informações atualizadas.

Ainda assim, a longo prazo, prevê-se que a utilização em massa desta tecnologia possa resultar em economias significativas também para o setor profissional, além de melhorar a qualidade e consistência das certificações energéticas em Portugal. A redução de erros e de trabalho de pesquisa, evidenciada pela melhoria nas pontuações de precisão e relevância das respostas, sugere uma diminuição potencial em custos de retrabalho.

A introdução deste tipo de sistemas tem a capacidade de transformar positivamente as práticas no setor de certificação energética. Antecipa-se uma transição de tarefas repetitivas de pesquisa para análises mais estratégicas e de valor acrescentado. Os profissionais da área poderão dedicar mais tempo à interpretação de resultados e à aplicação de conhecimentos especializados, em vez de perderem tempo e produtividade em pesquisas na regulamentação.

É de salientar que, apesar dos resultados positivos, é importante reconhecer as limitações deste estudo. O foco exclusivo no contexto português limita a generalização direta dos resultados para outros países ou sistemas regulatórios. A amostra de 10 perguntas utilizadas para a avaliação comparativa, embora informativa, pode não representar toda a gama de cenários encontrados na prática da certificação energética.

Outro desafio pode estar na rápida evolução da tecnologia, onde há possibilidade de tornar alguns aspetos técnicos do sistema obsoletos num curto prazo. Notou-se um desenvolvimento impactante de outros LLM desde o inicio deste projeto, provavelmente com capacidades mais interessantes do ponto de vista de tecnologia, proporcionando economias de tempo e aumentado ainda mais as precisões das respostas. Com isto, considera-se que existe uma necessidade contínua de atualização do sistema proposto para se adaptar aos novos avanços e além de manter a utilização

funcional numa perspetiva de longo prazo, existir uma possibilidade de evolução da solução proposta.

Em síntese, este projeto apresentou um potencial significativo da integração de novos sistemas RAG com os LLM. Considera-se que os objetivos iniciais foram alcançados, com melhorias tangíveis na precisão, eficiência e adaptabilidade dos processos de certificação. A redução de 60% no tempo de processamento e o aumento consistente nas pontuações de qualidade das respostas evidenciam o valor acrescentado da abordagem feita neste projeto.

Considera-se que a contribuição deste trabalho posiciona-se além da área da certificação energética, desempenhando o pepel de modelo base para a aplicação desta tecnologia noutros domínios complexos, sejam eles no ramo da engenharia, saúde ou finanças. O sucesso deste projeto estabelece um possível ponto de partida para a transformação digital de setores técnicos e altamente regulamentados, tendo sido o foco a abertura de novos caminhos para inovações futuras na intersecção entre inteligência artificial e áreas profissionais especializadas.

6

TRABALHO FUTURO

Com base nos resultados e conclusões obtidos no desenvolvimento deste projeto, considera-se há espaço para futuros desenvolvimentos. Os resultados desta implementação apresentam possíveis caminhos para expandir as capacidades e aplicabilidade desta tecnologia. À medida que se avança na tecnologia, torna-se evidente que o trabalho realizado até aqui serve como um sólido alicerce sobre o qual é possível construir inovações ainda mais impactantes. Tendo em consideração este contexto, esta secção descreve possíveis focos que podem levar este sistema a novos patamares de eficiência, integridade e utilidade prática em contexto de engenharia.

Um dos primeiros focos passa pela otimização da latência do sistema Geração Aumentada por Recuperação (RAG). Pode-se alcançar esta meta através do desenvolvimento de técnicas avançadas de *caching* para perguntas frequentes, exploração de algoritmos de indexação mais eficientes para as Bases de Dados Vetoriais (VecDBs), e implementação de métodos de paralelização mais sofisticados no processo de geração de respostas. A implementação de sistemas de *caching* poderá reduzir significativamente o tempo de resposta para *queries* recorrentes, enquanto a exploração de algoritmos de indexação mais robustos, como o Hierarchical Navigable Small World (HNSW) ou o Facebook AI Similarity Search (FAISS) mencionados neste artigo, poderá melhorar a velocidade de recuperação de informações.

Adicionalmente, a paralelização avançada do processo de geração de respostas, possivelmente através da utilização de *frameworks* como o *Ray*[1], permitirá uma utilização mais eficiente dos recursos computacionais disponíveis. Estas melhorias aumentarão a qualidade de experiência do utilizador, uma característica importante redução de resistência na aplicação desta tecnologia em ambientes de produção.

O desenvolvimento de uma interface de utilizador com um design mais intuitivo representa também uma melhoria significativa na experiência de utilizador e interpretabilidade do sistema.

À medida que o sistema se torna mais avançado e amplamente utilizado, torna-se imperativo abordar questões de ética e transparência com maior atenção. O

1 A biblioteca Ray é uma *framework* que permite escalar aplicações de *Machine Learning*.Disponível em: https://www.ray.io/.

desenvolvimento de ferramentas para explicar as decisões do sistema de forma compreensível, possivelmente através da implementação de técnicas de Explainable AI (XAI) ou SHapley Additive exPlanations (SHAP), ajudará a construir confiança entre os utilizadores e a tecnologia.

Um aspeto crítico do trabalho futuro será manter o sistema atualizado com os avanços mais recentes em tecnologia, dada a sua rápida evolução. Por fim, embora o foco inicial tenha sido a certificação energética, o potencial do sistema RAG pode ser expandido para outros domínios de engenharia e além. A adaptação do sistema para áreas relacionadas, como engenharia civil ou arquitetura, bem como a exploração de aplicações em outras áreas que usufruam de conformidade legal ou gestão de risco, representam direções promissoras para investigação futura.

Em suma, o trabalho futuro aqui delineado visa não apenas melhorar o sistema existente, mas também expandir o impacto e aplicabilidade. Ao focar na otimização de desempenho, melhoria da interface do utilizador, considerações éticas, adaptação contínua a novas tecnologias e expansão para novos domínios, prevê-se que o sistema RAG continue a evoluir como uma ferramenta poderosa e versátil para aumentar a eficiência operacional tornado os profissionais em superprofissionais alavancados pela tecnologia.

BIBLIOGRAFIA

Adlakha, Vaibhav et al. (2024). *Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering*. arXiv: 2307.16877 [cs.CL]. URL: https://arxiv.org/abs/2307.16877.

Ai, Qingyao et al. (2023). «Information Retrieval meets Large Language Models: A strategic report from Chinese IR community». Em: *AI Open* 4, pp. 80–90. ISSN: 2666-6510. DOI: https://doi.org/10.1016/j.aiopen.2023.08.001. URL: https://www.sciencedirect.com/science/article/pii/S2666651023000049.

Akkiraju, Rama et al. (2024). *FACTS About Building Retrieval Augmented Generation-based Chatbots*. arXiv: 2407.07858 [cs.LG]. URL: https://arxiv.org/abs/2407.07858.

Alizadeh, Keivan et al. (2024). *LLM in a flash: Efficient Large Language Model Inference with Limited Memory*. arXiv: 2312.11514 [cs.CL]. URL: https://arxiv.org/abs/2312.11514.

Andoni, Alexandr et al. (2015). *Practical and Optimal LSH for Angular Distance*. arXiv: 1509.02897 [cs.DS]. URL: https://arxiv.org/abs/1509.02897.

Balaguer, Angels et al. (2024). *RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture*. arXiv: 2401.08406 [cs.CL]. URL: https://arxiv.org/abs/2401.08406.

Barbieri, Francesco, Luis Espinosa Anke e Jose Camacho-Collados (2022). *XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond*. arXiv: 2104.12250 [cs.CL]. URL: https://arxiv.org/abs/2104.12250.

Brown, Nathan et al. (2023). *Efficient Transformer Knowledge Distillation: A Performance Review*. arXiv: 2311.13657 [cs.CL]. URL: https://arxiv.org/abs/2311.13657.

Brown, Tom B. et al. (2020). *Language Models are Few-Shot Learners*. arXiv: 2005.14165 [cs.CL]. URL: https://arxiv.org/abs/2005.14165.

Cho, Kyunghyun et al. (2014). *On the Properties of Neural Machine Translation: Encoder-Decoder Approaches*. arXiv: 1409.1259 [cs.CL]. URL: https://arxiv.org/abs/1409.1259.

Devlin, Jacob et al. (jun. de 2019a). «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». Em: *Proceedings of the 2019 Con-*

ference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Ed. por Jill Burstein, Christy Doran e Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Devlin, Jacob et al. (2019b). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.

Eurostat (2022). *Smart technologies in EU enterprises: AI and IoT*. Accessed: 2024-09-13. URL: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20220609-1.

Gekhman, Zorik et al. (2024). *Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?* arXiv: 2405.05904 [cs.CL]. URL: https://arxiv.org/abs/2405.05904.

Hwang, Taeho et al. (2024). *DSLR: Document Refinement with Sentence-Level Re-ranking and Reconstruction to Enhance Retrieval-Augmented Generation*. arXiv: 2407.03627 [cs.CL]. URL: https://arxiv.org/abs/2407.03627.

Izacard, Gautier e Edouard Grave (abr. de 2021). «Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering». Em: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. por Paola Merlo, Jorg Tiedemann e Reut Tsarfaty. Online: Association for Computational Linguistics, pp. 874–880. DOI: 10.18653/v1/2021.eacl-main.74. URL: https://aclanthology.org/2021.eacl-main.74.

Jiang, Albert Q. et al. (2023). *Mistral 7B*. arXiv: 2310.06825 [cs.CL]. URL: https://arxiv.org/abs/2310.06825.

Jing, Zhi et al. (2024). *When Large Language Models Meet Vector Databases: A Survey*. arXiv: 2402.01763 [cs.DB]. URL: https://arxiv.org/abs/2402.01763.

Johnson, Jeff, Matthijs Douze e Hervé Jégou (2017). *Billion-scale similarity search with GPUs*. arXiv: 1702.08734 [cs.CV]. URL: https://arxiv.org/abs/1702.08734.

Ke, YuHe et al. (2024). *Development and Testing of Retrieval Augmented Generation in Large Language Models – A Case Study Report*. arXiv: 2402.01733 [cs.CL]. URL: https://arxiv.org/abs/2402.01733.

Khattab, Omar, Keshav Santhanam et al. (2023). *Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP*. arXiv: 2212.14024 [cs.CL]. URL: https://arxiv.org/abs/2212.14024.

Khattab, Omar e Matei Zaharia (2020). *ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT*. arXiv: 2004.12832 [cs.IR]. URL: https://arxiv.org/abs/2004.12832.

Kusner, Matt et al. (jul. de 2015). «From Word Embeddings To Document Distances». Em: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. por Francis Bach e David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 957–966. URL: https://proceedings.mlr.press/v37/kusnerb15.html.

Lewis, Patrick et al. (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. arXiv: 2005.11401 [cs.CL]. URL: https://arxiv.org/abs/2005.11401.

Li, Jiwei et al. (2016). *Deep Reinforcement Learning for Dialogue Generation*. arXiv: 1606.01541 [cs.CL]. URL: https://arxiv.org/abs/1606.01541.

Liu, Tie-Yan (2009). *Learning to Rank for Information Retrieval*. Vol. 3. Foundations and Trends in Information Retrieval 3. Springer, pp. 225–331. ISBN: 978-3-642-14266-7. DOI: 10.1561/1500000016.

Malkov, Yu. A. e D. A. Yashunin (2018). *Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs*. arXiv: 1603.09320 [cs.DS]. URL: https://arxiv.org/abs/1603.09320.

Mikolov, Tomas et al. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv: 1301.3781 [cs.CL]. URL: https://arxiv.org/abs/1301.3781.

Min, Sewon et al. (2023). *FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation*. arXiv: 2305.14251 [cs.CL]. URL: https://arxiv.org/abs/2305.14251.

Narayanan, Deepak et al. (2021). *Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM*. arXiv: 2104.04473 [cs.CL]. URL: https://arxiv.org/abs/2104.04473.

Naveed, Humza et al. (2024). *A Comprehensive Overview of Large Language Models*. arXiv: 2307.06435 [cs.CL]. URL: https://arxiv.org/abs/2307.06435.

Pennington, Jeffrey, Richard Socher e Christopher D Manning (2014). «GloVe: Global Vectors for Word Representation». Em: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Radford, Alec et al. (2019). *Language Models are Unsupervised Multitask Learners*. URL: https://api.semanticscholar.org/CorpusID:160025533.

Reimers, Nils e Iryna Gurevych (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv: 1908.10084 [cs.CL]. URL: https://arxiv.org/abs/1908.10084.

Shah, Chirag et al. (2024). *Using Large Language Models to Generate, Validate, and Apply User Intent Taxonomies*. arXiv: 2309.13063 [cs.IR]. URL: https://arxiv.org/abs/2309.13063.

Singhal, Meena (jan. de 2001). «Reading proficiency, reading strategies, metacognitive awareness and L2 readers». Em: *The Reading Matrix* 1.

Siriwardhana, Shamane et al. (2023). «Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering». Em: *Transactions of the Association for Computational Linguistics* 11, pp. 1–17. DOI: 10.1162/tacl_a_00530. URL: https://aclanthology.org/2023.tacl-1.1.

Strubell, Emma, Ananya Ganesh e Andrew McCallum (2019). *Energy and Policy Considerations for Deep Learning in NLP*. arXiv: 1906.02243 [cs.CL]. URL: https://arxiv.org/abs/1906.02243.

Sutskever, Ilya, Oriol Vinyals e Quoc V. Le (2014). *Sequence to Sequence Learning with Neural Networks*. arXiv: 1409.3215 [cs.CL]. URL: https://arxiv.org/abs/1409.3215.

Team, Gemini et al. (2024). *Gemini: A Family of Highly Capable Multimodal Models*. arXiv: 2312.11805 [cs.CL]. URL: https://arxiv.org/abs/2312.11805.

Team, Gemma et al. (2024). *Gemma: Open Models Based on Gemini Research and Technology*. arXiv: 2403.08295 [cs.CL]. URL: https://arxiv.org/abs/2403.08295.

Touvron, Hugo et al. (2023). *LLaMA: Open and Efficient Foundation Language Models*. arXiv: 2302.13971 [cs.CL]. URL: https://arxiv.org/abs/2302.13971.

Tyen, Gladys et al. (jul. de 2022). «Towards an open-domain chatbot for language practice». Em: *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*. Ed. por Ekaterina Kochmar et al. Seattle, Washington: Association for Computational Linguistics, pp. 234–249. DOI: 10.18653/v1/2022.bea-1.28. URL: https://aclanthology.org/2022.bea-1.28.

Vaswani, Ashish et al. (2023). *Attention Is All You Need*. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

Wang, Mengzhao et al. (2023). *MUST: An Effective and Scalable Framework for Multimodal Search of Target Modality*. arXiv: 2312.06397 [cs.DB]. URL: https://arxiv.org/abs/2312.06397.

Wu, Lingfei et al. (out. de 2018). «Word Mover's Embedding: From Word2Vec to Document Embedding». Em: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. por Ellen Riloff et al. Brussels,

Belgium: Association for Computational Linguistics, pp. 4524–4534. DOI: 10.18653/v1/D18-1482. URL: https://aclanthology.org/D18-1482.

Wu, Qingyun et al. (2023). «AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework». Em: *ArXiv* abs/2308.08155. URL: https://api.semanticscholar.org/CorpusID:260925901.

Xu, Huatao et al. (ago. de 2024). «Penetrative AI: Making LLMs Comprehend the Physical World». Em: *Findings of the Association for Computational Linguistics ACL 2024*. Ed. por Lun-Wei Ku, Andre Martins e Vivek Srikumar. Bangkok, Thailand e virtual meeting: Association for Computational Linguistics, pp. 7324–7341. URL: https://aclanthology.org/2024.findings-acl.437.

Yan, Lixiang et al. (ago. de 2023). «Practical and ethical challenges of large language models in education: A systematic scoping review». Em: *British Journal of Educational Technology* 55.1, pp. 90–112. ISSN: 1467-8535. DOI: 10.1111/bjet.13370. URL: http://dx.doi.org/10.1111/bjet.13370.

Zebari, Rizgar et al. (mai. de 2020). «A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction». Em: *Journal of Applied Science and Technology Trends* 1, pp. 56–70. DOI: 10.38094/jastt1224.

APÊNDICES

DECLARAÇÃO

Declaro, sob compromisso de honra, que o trabalho apresentado neste projeto, com o título *"Aplicação de RAG em modelos LLM com Bases de Dados Vetoriais"*, é original e foi realizado por Ruben Alexandre Dias Marques (2220128) sob orientação de Professor Doutor Ricardo Malheiro (ricardo.malheiro@ipleiria.pt).

Leiria, Novembro de 2024

Ruben Alexandre Dias Marques