UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Aplicando LLMs para Question Answering

Arthur Sakayan Vieira de Melo

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP
SÃO CARLOS

Arthur Sakayan Vieira de Melo

Aplicando LLMs para Question Answering

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Dr. Brucce Neves dos Santos

Versão original

São Carlos

2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

S856m

Melo, Arthur Sakayan Vieira de
Aplicando LLMs para Question Answering / Arthur Sakayan Vieira de Melo ; orientador Brucce Neves dos Santos. – São Carlos, 2025.
36 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. LLM 2. Question Answering 3. RAG, 4. Onboarding 5. Chatbot 6. Treinamento Corporativo. I. Santos, Brucce Neves dos, orient. II. Título.

Arthur Sakayan Vieira de Melo

Applying LLMs for Question Answering

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Dr. Brucce Neves dos Santos

Original version

São Carlos

2025

AGRADECIMENTOS

Gostaria de agradecer a minha família por sempre ter me apoiado.

Gostaria de agradecer os professores do MBA em Inteligência Artificial e Big Data do ICMC, por divulgarem o conhecimento destas áreas.

Gostaria de agradecer o meu orientador, Brucce Neves dos Santos, por todo o *feedback* que me deu.

Por fim, gostaria de destacar que as ferramentas baseadas em *Large Language Models*, foram utilizadas exclusivamente para refinar a linguagem e aprimorar a clareza ao longo deste estudo, sem alterar seu significado original, propósito acadêmico ou autoria.

"A Vitória pertence ao mais perseverante"
Napoleão Bonaparte

RESUMO

MELO, A. **Aplicando LLMs para Question Answering**. 2025. 36 p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

Neste trabalho, é investigado a aplicação de Grandes Modelos de Linguagem (LLM) para a construção de um sistema de *Question Answering* para auxiliar no processo de *onboarding* e integração de funcionários. Para poder cumprir tal objetivo, foi utilizada a técnica de *Retrieval Augmented Generation*. A metodologia baseou-se em quatro etapas: recuperação de documentos relevantes a partir de uma pergunta do usuário, contextualização do LLM com esses documentos, geração da resposta e retorno ao usuário. Para avaliação, foi construída uma base de 36 documentos corporativos reais, e foram avaliados cinco modelos de LLMs (Llama 3.1, DeepSeek-R1, Gemma3, Qwen3 e Phi-4), que foram testados usando oito critérios, como factualidade, fidelidade, clareza, ausência de alucinações dentre outros. Os resultados demonstraram que os modelos Llama 3.1 8B e Gemma3 12B obtiveram o melhor desempenho global, produzindo respostas precisas, concisas e alinhadas ao conteúdo original. Em contraste, o DeepSeek-R1 apresentou desempenho insatisfatório na maioria dos critérios. Conclui-se que a integração de LLMs com RAG é uma abordagem viável e promissora para melhorar a comunicação interna, otimizar a gestão do conhecimento e acelerar a adaptação de novos colaboradores, embora desafios como custo computacional e escalabilidade em grandes volumes de documentos ainda persistam.

Palavras-chave:LLM, *Question Answering*, RAG, *Onboarding*, *Chatbot*, Treinamento Corporativo.

ABSTRACT

MELO, A. **Applying LLMs for Question Answering**. 2025. 36 p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

This work investigates the application of Large Language Models (LLM) to build a Question Answering system to assist in the onboarding and employee integration process. To achieve this objective, the Retrieval Augmented Generation technique was used. The methodology consisted of four steps: retrieving relevant documents from a user's question, contextualizing the LLM with these documents, generating the answer, and providing feedback to the user. For evaluation, a database of 36 real corporate documents was constructed, and five LLM models (Llama 3.1, DeepSeek-R1, Gemma3, Qwen3, and Phi-4) were evaluated, which were tested using eight criteria, such as factuality, fidelity, clarity, absence of hallucinations, and others. The results showed that the Llama 3.1 8B and Gemma3 12B models achieved the best overall performance, producing accurate, concise responses that aligned with the original content. In contrast, DeepSeek-R1 performed poorly on most criteria. The conclusion is that integrating LLMs with RAG is a viable and promising approach to improving internal communication, optimizing knowledge management, and accelerating the adaptation of new employees, although challenges such as computational cost and scalability with large document volumes still persist.

Keywords: LLM, Question Answering, RAG, Onboarding, Chatbot, Corporate Training.

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

Dentro de uma empresa, é fundamental que todos os funcionários, desde os novatos até os mais experientes, tenham pleno conhecimento sobre as normas de funcionamento e os processos internos. Esse entendimento é essencial para garantir a eficiência organizacional, uma vez que cada empresa possui regras e métodos específicos, além de uma cultura interna que influencia diretamente o desempenho e a produtividade. A integração eficaz dos novos funcionários, por meio de um bom processo de *onboarding* e treinamento, é crucial para a adaptação deles ao ambiente corporativo, evitando impactos negativos no desempenho e nos resultados da empresa (KLEIN; WEAVER, 2000).

Além disso, com a constante evolução da tecnologia e das práticas de mercado, é vital que os funcionários já estabelecidos na empresa recebam atualizações contínuas sobre novas ferramentas e processos. Caso contrário, a empresa corre o risco de perder competitividade frente aos seus concorrentes, já que a falta de adaptação pode levar à redução da performance individual e, consequentemente, ao impacto negativo nos resultados da organização (Salas *et al.*, 2012).

A ausência de treinamento adequado pode acarretar uma série de problemas, como a diminuição da produtividade, aumento na sobrecarga de trabalho para outros empregados e queda na moral da equipe. Em situações mais graves, falhas operacionais evitáveis podem ocorrer, resultando em prejuízos financeiros e perda de clientes, o que comprometeria a reputação da empresa no mercado (Clardy, 2005).

Atualmente, muitas empresas enfrentam dificuldades na transmissão do conhecimento necessário para o bom desempenho dos seus funcionários (Sun; Scott, 2005). Em muitos casos, esse conhecimento é compartilhado por meio de documentos, treinamentos presenciais ou conversas informais com supervisores. No entanto, esses métodos apresentam limitações significativas. A documentação pode estar desatualizada, o acesso aos materiais nem sempre é claro e os superiores podem não estar disponíveis para fornecer suporte contínuo. Além disso, a formação tradicional nem sempre abrange todas as dúvidas ou necessidades específicas dos funcionários, o que pode resultar em lacunas no aprendizado e na execução dos processos (Szulanski, 1996).

Para resolver esse problema, algumas empresas têm adotado métodos tradicionais da área de Processamento de Linguagem Natural (PLN), que utilizam sistemas para responder às dúvidas dos usuários. Embora esses sistemas já representem um avanço, ainda há espaço para melhorias em termos de eficácia e adaptação às necessidades específicas de cada organização (Jiang *et al.*, 2024).

Com o avanço tecnológico, surgiram novas ferramentas, como os Grandes Modelos

de Linguagem (LLMs), que têm revolucionado o campo da interação com sistemas de inteligência artificial. Esses modelos possuem a capacidade de gerar, traduzir e até resumir textos, oferecendo uma possibilidade inovadora para a criação de sistemas que atendem às necessidades de comunicação dentro das empresas (Xu, 2025). Utilizando essas tecnologias, é possível desenvolver um *chatbot* mais humanizado e inteligente, capaz de responder de maneira precisa e personalizada às dúvidas dos funcionários, melhorando a integração e o aprendizado contínuo dentro da organização (Afzal *et al.*, 2024).

A implementação de um sistema como esse poderia não apenas melhorar a comunicação interna, mas também facilitar a adaptação dos novos funcionários e reduzir a sobrecarga de trabalho, garantindo que todos os colaboradores, independentemente do seu tempo de empresa, tenham acesso rápido e eficiente às informações essenciais para o bom desempenho de suas funções. (Brachten *et al.*, 2020).

1.1 Objetivos

O objetivo geral deste trabalho é explorar e aplicar modelos de LLMs para melhorar o desempenho de sistemas de *Question Answering* (QA) no contexto corporativo, com foco no treinamento e integração de novos funcionários. Para isso, serão avaliados modelos *open-source* capazes de rodar localmente, de modo que os dados não sejam enviados a terceiros, garantindo maior segurança e confidencialidade das informações organizacionais. O intuito é desenvolver uma solução eficiente e inteligente que facilite a disseminação do conhecimento interno e otimize a comunicação dentro das empresas.

Objetivos Específicos:

- Investigar o estado da arte em sistemas de *Question Answering* utilizando LLMs: analisar as abordagens atuais para a aplicação de LLMs em sistemas de *Question Answering*, destacando suas capacidades, limitações e melhores práticas.
- Desenvolver um protótipo de *chatbot* utilizando LLMs: criar um protótipo funcional de um *chatbot* que use modelos avançados de linguagem para responder às dúvidas frequentes dos funcionários, fornecendo informações relevantes sobre normas de funcionamento e processos internos da empresa.

1.2 Organização do texto

Este documento está dividido em 5 capítulos: No Capítulo 1 foi abordado a introdução, onde foram apresentados os problemas que serão resolvidos. No Capítulo 2 é apresentado os fundamentos e trabalhos relacionados necessários para compreender a solução do problema. No Capítulo 3 é abordado a proposta. No Capítulo 4 são apresentados os resultados obtidos. Por fim, no Capítulo 5 é feita a conclusão.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo tem como objetivo fornecer os fundamentos teóricos necessários para o entendimento do trabalho. Na Seção 2.1 será abordada a tarefa de *Question Answering*. Enquanto que na Seção 2.2 serão abordados os LLMs, que são a base da solução proposta. Por fim na Seção 2.3 serão abordados os trabalhos relacionados.

2.1 Question Answering

QA é uma tarefa dentro da área de PLN, cujo objetivo é fornecer respostas precisas a perguntas feitas por usuários. Essa área de pesquisa possui uma longa trajetória histórica, remontando à década de 1960, quando os primeiros sistemas de QA foram desenvolvidos. Inicialmente, tais sistemas eram capazes de responder a perguntas relacionadas a fatos científicos e estatísticas de esportes, como, por exemplo, informações sobre beisebol (Jurafsky; Martin, 2025).

No entanto, os avanços mais significativos na área de QA foram impulsionados pelo desenvolvimento de sistemas como o IBM Watson e sua aplicação no tratamento de grandes volumes de dados textuais. O Watson, em particular, se destacou pela capacidade de processar perguntas complexas e gerar respostas precisas a partir de uma vasta base de dados, empregando técnicas avançadas de PLN e *Machine Learning*. Esse tipo de sistema evidenciou a eficácia dos modelos de aprendizado de máquina em tarefas de compreensão e resposta a perguntas em larga escala, o que gerou um interesse crescente na utilização desses métodos em outras áreas, como saúde, finanças e ciência (Jurafsky; Martin, 2025).

Embora os sistemas de QA tenham demonstrado um desempenho satisfatório ao longo dos anos, os recentes avanços em LLMs têm representado uma evolução significativa, superando os modelos tradicionais em termos de capacidade de generalização e precisão. A literatura recente sugere que os LLMs, ao contrário dos modelos especializados em QA, são capazes de lidar com uma gama mais ampla de questões, independentemente do domínio, oferecendo respostas mais flexíveis e adaptativas. Esse aprimoramento está relacionado à natureza não especializada desses modelos, que são capazes de integrar e gerar respostas com base em uma vasta quantidade de dados de treinamento (Fischer *et al.*, 2024).

2.2 Large Language Models

LLMs representam um avanço significativo no campo do PLN, caracterizando-se por sua elevada capacidade paramétrica e versatilidade em diversas tarefas, como tradução automática, predição de sequência, sumarização textual, resposta a perguntas e geração de código. A base arquitetural predominante desses modelos é o *Transformer*, proposto

por Vaswani *et al.* (2017), cuja principal inovação reside na substituição das estruturas recorrentes tradicionais por mecanismos de atenção completamente paralelizáveis.

A arquitetura *Transformer* é composta por duas partes fundamentais: o *encoder* e o *decoder*. O *encoder* é responsável por converter sequências textuais em representações vetoriais contínuas, conhecidas como *embeddings*, enriquecidas por vetores posicionais que informam a ordem das palavras na sentença. Cada camada do *encoder* contém duas subcamadas principais: (i) uma camada de *multi-head self-attention*, que permite ao modelo considerar o contexto de todas as palavras simultaneamente, e (ii) uma rede neural *feed-forward* totalmente conectada. Ambas as subcamadas são seguidas por conexões residuais (He *et al.*, 2016) e normalização de camada (Ba; Kiros; Hinton, 2016), elementos que facilitam o treinamento estável de redes profundas (Vaswani *et al.*, 2017).

O *decoder*, por sua vez, tem como principal função a geração autorregressiva de texto, ou seja, a predição sequencial de *tokens* com base nos *tokens* previamente gerados. Cada camada do *decoder* também inclui subcamadas de *self-attention*, mas com máscara causal, que impede o acesso a *tokens* futuros, garantindo a coerência temporal da geração. Adicionalmente, há uma subcamada de atenção cruzada (*cross-attention*), que permite ao *decoder* utilizar as *embeddings* oriundos do *encoder* para guiar a geração textual (Vaswani *et al.*, 2017).

A saída final do *decoder* é processada por uma camada linear seguida de uma função *softmax*, que transforma os *logits* gerados em uma distribuição de probabilidade sobre o vocabulário. A palavra com maior probabilidade é então selecionada como saída. Este processo é iterado até que um *token* especial de finalização seja produzido (Vaswani *et al.*, 2017).

Com o avanço da capacidade computacional e a disponibilidade de grandes volumes de dados, os LLMs têm evoluído substancialmente em termos de escala e desempenho. Modelos com dezenas ou centenas de bilhões de parâmetros, como o GPT-4 (OpenAI *et al.*, 2024), demonstram capacidades emergentes que ultrapassam tarefas básicas de linguagem, como raciocínio lógico, interpretação semântica e até resolução de problemas matemáticos.

Um modelo bastante conhecido é o LLaMA, um acrônimo para *Large Language Model* Meta AI, que foi desenvolvido pela Meta. Parte do seu código fonte está disponível para o público, servindo para fomentar a pesquisa e o desenvolvimento da área. Mesmo não possuindo o mesmo tamanho de modelos como GPT-4, ele ainda possui uma eficiência considerável. Nas versões mais recentes, foi lançado um modelo com muito mais parâmetros que obteve bons resultados em vários *benchmarks*.

Outro modelo conhecido é o R1, desenvolvido pela empresa chinesa DeepSeek, destacou-se na área de Processamento de Linguagem Natural por alcançar desempenho superior a diversos modelos existentes em *benchmarks* como MMLU-Redux, LiveCodeBench,

AIME 2024 e CLUEWSC (DeepSeek-AI *et al.*, 2025). Seu diferencial reside na combinação de otimizações arquiteturais e no uso de um extenso volume de dados de treinamento, o que permitiu resultados expressivos com um custo de desenvolvimento significativamente inferior. Além disso, a DeepSeek disponibilizou o código-fonte do modelo, promovendo maior transparência e incentivando a pesquisa aberta na área.

2.3 Trabalhos Relacionados

Com a crescente adoção de LLMs em ambientes corporativos, surgem oportunidades significativas para o desenvolvimento de agentes conversacionais inteligentes capazes de auxiliar novos colaboradores em processos de integração organizacional (*onboarding*). Nesse contexto, uma abordagem promissora é o uso de RAG, que combina a capacidade generativa dos LLMs com a recuperação de informações precisas a partir de bases documentais, mitigando o risco de alucinações e promovendo maior consistência factual nas respostas. Diversos trabalhos recentes têm explorado variações dessa abordagem em tarefas de QA, com aplicações que, embora distintas do domínio organizacional, oferecem *insights* metodológicos valiosos para a construção de sistemas de suporte ao *onboarding*. A seguir, são discutidos três estudos representativos. He *et al.* (2024) propuseram um *framework* para tarefas de QA e compreensão sobre grafos textuais. Para tanto, foi desenvolvido o *benchmark GraphQA*, baseado na combinação de três *benchmarks* prévios: ExplaGraphs, SceneGraphs e WebQSP. Visando mitigar as limitações contextuais das LLMs, os autores empregam uma estratégia de RAG sobre grafos, modelando o problema como uma instância da otimização da árvore de Steiner. Esse *framework* demonstrou bons resultados em métricas de performance e eficiência, além de estudos de ablação que validam suas contribuições.

Siriwardhana *et al.* (2022) propuseram o RAG-*end2end*, uma extensão do modelo RAG tradicional, que permite o treinamento conjunto de todos os seus componentes, tanto o módulo de recuperação quanto o gerador. Essa abordagem oferece maior integração entre as fases de busca e geração, aumentando a robustez geral do sistema. O modelo foi avaliado por meio de métricas como *Exact Match, F1 Score e Top-k Retrieval Accuracy*, obtendo resultados superiores a abordagens desacopladas. No entanto, foram identificadas limitações como alto custo computacional, persistência de inconsistências factuais e dificuldades de escalabilidade em bases documentais muito extensas. Desafios pertinentes ao escopo do presente trabalho, que visa lidar com grandes volumes de informações organizacionais.

Desta *et al.* (2024) investigou a adaptação de LLMs para línguas de poucos recursos, tomando como estudo de caso a língua etíope Amhárico. Utilizando *Low-Rank Adaptation* (LoRA), foi realizado o *fine-tuning* do modelo llama-2-7b-reboot-amharic e implementado um sistema de QA baseado em RAG, capacitado a consultar documentos externos e aumentar a base de conhecimento do modelo. A avaliação foi conduzida com métricas como BLEU e análise de 50 conjuntos de teste. Apesar de seu escopo linguístico específico, o

estudo evidencia o potencial do RAG para suprir lacunas informacionais em contextos com dados especializados, característica análoga à que se observa em ambientes corporativos, nos quais a linguagem é frequentemente dominada por jargões internos e documentação institucional. Entre as limitações, destacam-se o viés nos dados, alucinações e escassez de recursos computacionais e linguísticos.

Com base nesses trabalhos, observa-se que o uso de LLMs em conjunto com mecanismos de recuperação externa oferece um caminho viável e promissor para aplicações voltadas ao suporte em ambientes empresariais. No caso específico do *onboarding*, onde o acesso rápido e confiável às políticas, fluxos e normas internas é essencial para uma integração eficiente do colaborador, a combinação de RAG com bases documentais internas pode representar uma solução eficaz, escalável e personalizada.

3 METODOLOGIA

A metodologia adotada neste trabalho tem como base o paradigma de RAG, uma abordagem amplamente utilizada para aprimorar a geração de respostas em LLMs. O objetivo principal do RAG é combinar a capacidade de geração de linguagem natural dos modelos com a recuperação de informações relevantes de uma base de conhecimento, garantindo respostas mais precisas, contextualizadas e alinhadas com os dados da organização. Essa abordagem é especialmente adequada ao contexto corporativo, em que informações como regras de negócio, procedimentos internos e políticas organizacionais devem ser transmitidas de maneira clara e confiável aos funcionários.

Dessa forma, o *chatbot* desenvolvido terá a capacidade de responder às dúvidas mais recorrentes de novos funcionários durante o processo de *onboarding*, oferecendo suporte imediato e reduzindo o tempo necessário para sua adaptação. Além disso, também servirá como auxílio contínuo para colaboradores já estabelecidos na organização, facilitando a compreensão e o uso de novas ferramentas ou processos introduzidos no ambiente corporativo. Por fim, o sistema garantirá que as respostas fornecidas estejam sempre alinhadas à documentação oficial da empresa, assegurando consistência, confiabilidade e padronização na transmissão do conhecimento organizacional. O processo de RAG pode ser dividido em quatro etapas principais, conforme ilustrado na Figura 1:

1. Pergunta: o processo inicia-se quando o usuário insere uma pergunta em linguagem natural.
2. Recuperação de Documentos (*Retrieval*): a pergunta é utilizada como consulta em uma base de dados (ex.: repositório de documentos, manuais internos ou base vetorial), recuperando os trechos de texto mais relevantes.
3. Geração com Contexto (LLM): os trechos recuperados são fornecidos ao LLM como contexto adicional, juntamente com a pergunta do usuário.
4. Resposta: o modelo gera uma resposta em linguagem natural, combinando seu conhecimento prévio com as informações recuperadas, resultando em uma saída mais precisa e adaptada ao domínio corporativo.

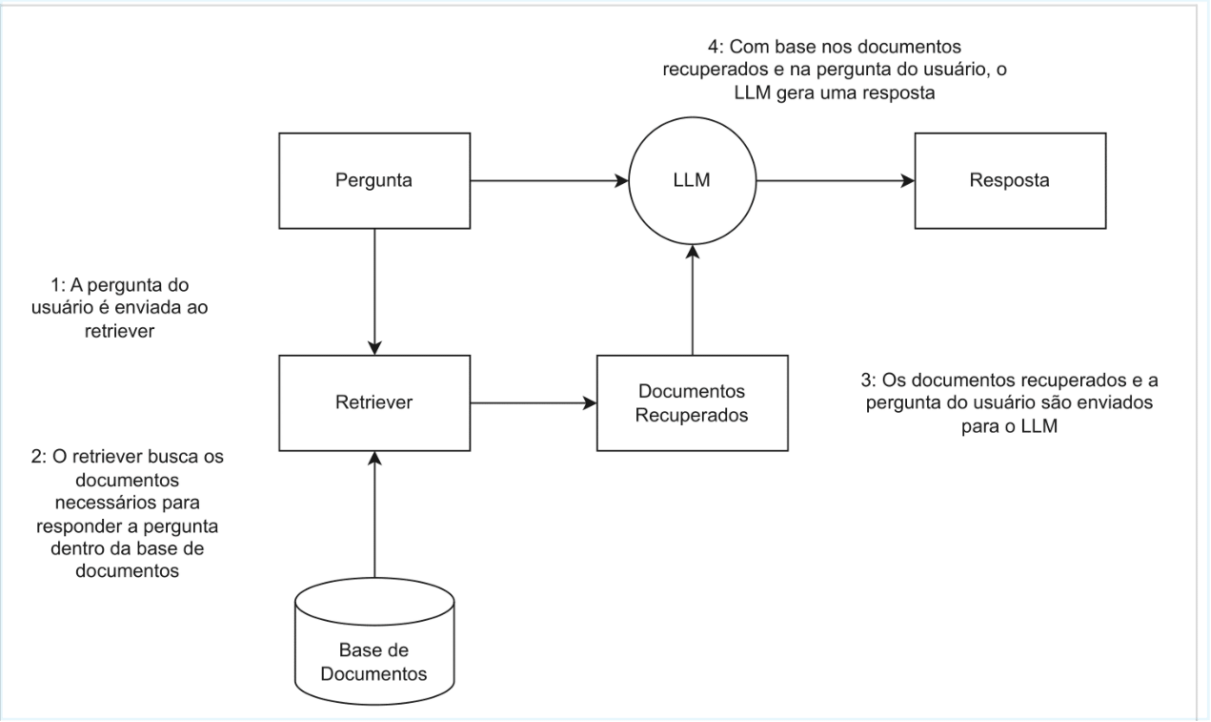

Figura 1 – Uma representação visual do processo

4 AVALIAÇÃO EXPERIMENTAL

Este capítulo tem como objetivo apresentar e discutir os resultados obtidos a partir da aplicação da metodologia proposta. Para isso, o texto está organizado em quatro seções complementares. A Seção 4.1 descreve os dados utilizados no desenvolvimento do estudo, detalhando suas origens e características. Em seguida, na Seção 4.2 são definidos os critérios de avaliação, que permitem mensurar de forma objetiva o desempenho do sistema. Na Seção 4.3, são apresentados os modelos empregados no experimento, com ênfase em suas particularidades e justificativas de escolha. Por fim, na Seção 4.4 é feita à análise dos resultados, na qual os dados obtidos são interpretados de forma crítica.

4.1 Conjunto de Dados para o RAG

A qualidade e a representatividade dos dados utilizados exercem papel central no desempenho de sistemas baseados em *Question Answering*. Neste trabalho, foi construída uma base documental contendo informações relevantes para o contexto organizacional em estudo. Essa base contempla diferentes áreas da gestão de serviços de tecnologia da informação, estruturadas de forma a refletir tanto processos operacionais quanto aspectos estratégicos da organização.

Devido a ausência de uma base propriamente estruturada contendo informações relevantes ao *onboarding* do funcionário dentro da companhia, foi utilizada uma abordagem pragmática dentro deste trabalho, onde apenas as informações relevantes foram extraídas de documentos específicos. Os documentos foram extraídos da base de dados de *Information Technology Service Management* (ITSM) da empresa.

Os dados foram organizados em quatro categorias principais. A primeira corresponde ao Departamento de Gestão Operacional de Serviços, composta por 2 documentos que descrevem os objetivos e deveres deste departamento. Tais documentos foram coletados para explicar mais sobre o contexto geral da empresa, pois os funcionários estarão dentro das divisões que compõem o departamento ao entrar na empresa. A segunda refere-se à Divisão de Transição de Serviços de Tecnologia de Informação e Comunicação (TIC), também representada por 2 documentos, que também descrevem os seus objetivos e deveres. Tais documentos foram coletados para explicar para novos funcionários o que é esperado ao entrar nesta área.

A terceira categoria diz respeito às descrições sobre o processo de gerenciamento de mudanças, formada por 6 documentos que reúnem conceitos, políticas e práticas associadas à gestão de alterações no ambiente corporativo de Tecnologia da Informação (TI). Tais documentos foram reunidos para fornecer uma explicação teórica sobre as

tarefas que o funcionário irá fazer no dia a dia. Por fim, a quarta categoria, de maior abrangência, contempla um total de 22 documentos dedicados às "Tarefas de Gerenciamento de Mudanças", contendo instruções detalhadas sobre os vários tipos de tarefas que serão desempenhadas pelos empregados. O tamanho de tal categoria é justificado por conter descrições das várias tarefas cotidianas dos funcionários, o que será consultado com maior frequência do que os objetivos e deveres da divisão de alocação e do departamento onde a divisão se encontra.

No total, a base é composta por 32 documentos, cobrindo diferentes níveis de granularidade: desde descrições conceituais e normativas até orientações práticas de execução. Para viabilizar sua utilização em um sistema de RAG, os textos foram previamente processados, padronizados e convertidos em formato adequado para indexação em uma base vetorial. Esse pré-processamento envolveu etapas de normalização e segmentação em unidades textuais menores (*chunks*).

A diversidade das categorias selecionadas garante que o conjunto de dados represente de forma equilibrada tanto aspectos de alto nível, relacionados à gestão e políticas organizacionais, quanto instruções de nível operacional, fundamentais para a execução das tarefas diárias. Esse equilíbrio é crucial para avaliar a capacidade do modelo em lidar com diferentes tipos de perguntas, desde as conceituais até as altamente específicas, refletindo situações reais do ambiente corporativo.

4.2 Critérios de Avaliação

Para assegurar uma análise sistemática e comparável do desempenho dos modelos empregados, foi estabelecido um conjunto de critérios de avaliação padronizados. O objetivo desta etapa é minimizar a subjetividade inerente à análise qualitativa das respostas geradas e, consequentemente, permitir uma comparação justa entre os diferentes LLMs.

O processo de avaliação foi conduzido a partir de um *pipeline* de RAG. Nesse *pipeline*, foram submetidas perguntas diretamente relacionadas aos tópicos presentes na base de dados descrita na Seção 4.1, e as respostas retornadas pelos modelos foram avaliadas com base em um conjunto de parâmetros definidos. Cada critério foi formulado de forma a captar diferentes dimensões da qualidade das respostas, com escalas discretas de pontuação que variam de 0 a 2. Os critérios adotados são apresentados a seguir:

- Alucinações de modelos: avalia a ocorrência de informações inventadas ou inconsistentes com a base de dados. Este critério está dividido em (0) Houve alucinações e (2) Não houve alucinações.

- Fidelidade ao conteúdo original: mede o grau de aderência da resposta ao material presente nos documentos de referência. Este critério está dividido em três categorias

sendo elas: (0) Distância significativa em relação ao conteúdo original; (1) Fidelidade parcial, com alguns desvios; e (2) Resposta totalmente fiel ao conteúdo original.

- Completude da resposta: avalia se a resposta contempla integralmente a questão formulada. Este critério está dividido em três categorias sendo elas: (0) A pergunta não foi respondida; (1) Resposta parcial, abordando apenas alguns aspectos; e (2) Resposta completa e abrangente.
- Factualidade da resposta: analisa a correção dos fatos apresentados. Este critério está dividido em três categorias sendo elas: (0) Nenhum fato correto; (1) Apenas alguns fatos corretos; e (2) Todos os fatos corretos.
- Concisão da resposta: examina a adequação do tamanho da resposta em relação ao necessário. Este critério está dividido em três categorias sendo elas: (0) Resposta excessivamente prolixa; (1) Leve redundância, mas ainda razoável; e (2) Resposta concisa e adequada.
- Clareza da resposta: mede a coerência e inteligibilidade da resposta. Este critério está dividido em três categorias sendo elas: (0) Resposta incompreensível; (1) Alguns lapsos de compreensão; e (2) Resposta clara e logicamente estruturada.
- Relevância das informações: avalia a pertinência das informações fornecidas em relação à pergunta. Este critério está dividido em três categorias sendo elas: (0) Nenhuma informação relevante; (1) Informações parcialmente relevantes; e (2) Todas as informações relevantes.
- Proximidade à pergunta: mede o grau de foco da resposta em relação à questão formulada. Este critério está dividido em três categorias sendo elas: (0) Grande perda de foco; (1) Pequena perda de foco; e (2) Nenhuma perda de foco.

A adoção desses critérios possibilita uma avaliação multidimensional das respostas, contemplando desde aspectos objetivos, como factualidade e fidelidade ao conteúdo, até dimensões mais subjetivas, como clareza e relevância. Essa abordagem garante maior robustez à análise e permite comparações mais consistentes entre os modelos avaliados.

4.3 Configuração Experimental

Os experimentos foram conduzidos utilizando um conjunto diversificado de LLMs, selecionados por suas diferentes arquiteturas, tamanhos e características funcionais. O objetivo dessa escolha foi garantir uma análise comparativa abrangente, contemplando tanto modelos generalistas quanto especializados. Foram avaliados os seguintes modelos: LLaMa 3.1 (8B), DeepSeek-R1 (14B), Gemma3 (12B), Qwen3 (14B) e Phi4 (14B). Todos os modelos foram executados com parâmetros padrão, de modo a manter a comparabilidade

dos resultados e evitar vieses decorrentes de ajustes excessivos de hiperparâmetros. A seguir, é apresentado uma breve caracterização de cada modelo avaliado:

- LLaMa 3.1 (8B): modelo desenvolvido pela Meta, lançado em julho de 2024. Seu tamanho relativamente reduzido torna-o adequado para ambientes com restrições computacionais. Possui suporte multilíngue, uma janela de contexto de até 128 mil *tokens* (um avanço significativo em relação às versões anteriores) e demonstrou forte capacidade em tarefas de raciocínio lógico.
- DeepSeek-R1 (14B): modelo voltado para raciocínio, desenvolvido pela companhia chinesa DeepSeek. Tornou-se conhecido internacionalmente por superar o modelo o1 da OpenAI em diferentes *benchmarks* de raciocínio matemático e lógico. Seu foco principal está na capacidade de inferência passo a passo e na resolução de problemas complexos.
- Gemma3 (12B): modelo multimodal e multilíngue capaz de processar tanto texto quanto imagens em seus *prompts*. Suporta mais de 140 idiomas e conta com versões especializadas em Medicina e Geração de Código, ampliando seu potencial de aplicação em domínios específicos. Sua arquitetura busca equilibrar a eficiência de recursos com desempenho em tarefas diversas.
- Qwen3 (14B): desenvolvido pela Alibaba, é projetado para alternar rapidamente entre um modo de respostas rápidas e um modo de raciocínio mais aprofundado, adaptando-se ao tipo de tarefa. Oferece suporte a 119 idiomas e se destaca pela versatilidade em tarefas de compreensão e geração de linguagem natural em múltiplos domínios.
- Phi4 (14B): modelo da Microsoft, projetado com ênfase em raciocínio e aderência a instruções. Apesar de possuir 14 bilhões de parâmetros, apresenta desempenho competitivo, superando diversos modelos de porte equivalente em *benchmarks* padrão. Seu design prioriza clareza e consistência nas respostas, características fundamentais em contextos corporativos.

Em conjunto, esses modelos representam um espectro variado de capacidades e arquiteturas, permitindo uma avaliação robusta sobre os desafios e potencialidades da integração de LLMs em soluções baseadas em RAG.

4.4 Análise dos resultados

Para a avaliação experimental, foi elaborado um conjunto de perguntas especificamente construídas a partir da base de dados descrita na Seção 4.1. Optou-se por não utilizar questões pré-existentes, uma vez que o objetivo principal era garantir que as perguntas

refletissem de forma fiel o conteúdo dos documentos disponíveis. Dessa forma, cada questão foi derivada diretamente das informações presentes na base textual, assegurando uma correspondência direta entre pergunta e conteúdo de referência.

Como consequência desse processo de elaboração, a quantidade de perguntas formuladas aproxima-se do número de documentos disponíveis, o que favorece uma avaliação equilibrada e representativa do conjunto de dados. No total, foram criadas 34 perguntas, distribuídas entre as diferentes categorias temáticas da base, de modo a contemplar tanto aspectos descritivos quanto procedimentais relacionados à gestão de serviços e mudanças. A partir dessas questões, foram conduzidas as avaliações experimentais, cujos resultados, organizados conforme os critérios estabelecidos na Seção 4.2, estão ilustrados na tabela abaixo.

Na Tabela 1 estão ilustrados os resultados obtidos a partir da avaliação experimental dos modelos. As colunas correspondem aos critérios de avaliação previamente definidos na Seção 4.2, enquanto as linhas representam os diferentes LLMs analisados. Cada célula da tabela contém três valores, organizados no formato x/y/z. O primeiro valor (x) indica a quantidade de perguntas classificadas na categoria 0; o segundo valor (y), a quantidade de perguntas classificadas na categoria 1; e o terceiro valor (z), a quantidade de perguntas classificadas na categoria 2. Esse formato de representação permite observar a distribuição das classificações, oferecendo uma visão mais detalhada e comparativa entre os critérios e os sistemas avaliados.

Tabela 1 – Resultado da avaliação dos modelos nos critérios

Modelo	Alucinação	Fidelidade ao conteúdo original	Completude da resposta	Factualidade da resposta	Concisão da resposta	Clareza da resposta	Relevância das informações	Proximidade à pergunta
Deepseek	2/-/33	12/18/5	0/0/34	0/3/32	13/16/6	1/0/34	1/5/29	0/0/35
Gemma3	0/-/35	0/2/33	0/0/35	0/0/35	0/4/31	0/0/35	0/0/35	0/0/35
Llama3.1	0/-/35	0/1/34	0/0/35	0/0/35	0/5/30	0/0/35	0/0/35	0/0/35
Phi4	0/-/34	5/22/8	0/0/35	0/0/35	12/17/6	0/0/35	1/3/31	0/0/35
Qwen3	0/-/35	1/15/19	0/3/31	0/1/34	3/16/16	0/3/35	0/3/32	0/0/35

No que diz respeito ao desempenho dos modelos, os resultados indicaram diferenças significativas entre as arquiteturas analisadas. Os modelos LLaMa 3.1 e Gemma3 obtiveram o melhor desempenho em todos os critérios estabelecidos, apresentando respostas consistentes, concisas e com elevada fidelidade ao conteúdo original.

O modelo Qwen3 demonstrou resultados satisfatórios nos critérios de factualidade, clareza e proximidade à pergunta. Contudo, apresentou fragilidades nos critérios de concisão e fidelidade, o que evidencia certa dificuldade em alinhar as respostas de forma direta e precisa ao conteúdo documental.

Por sua vez, o modelo Phi4 apresentou uma performance destacada nos critérios de factualidade, clareza e completude, o que indica sua robustez na compreensão global das perguntas. Entretanto, sua tendência à verbosidade e a baixa fidelidade ao conteúdo

original comprometeram a qualidade final das respostas, tornando-o menos eficiente em contextos que demandam precisão estrita em relação à fonte.

Por fim, o modelo Deepseek-R1 registrou o pior desempenho geral entre os avaliados, apresentando fragilidades em praticamente todos os critérios, o que limita sua aplicabilidade no cenário proposto.

De forma geral, os resultados confirmam que, embora todos os modelos apresentam pontos fortes em aspectos específicos, apenas alguns, como o LLaMa 3.1 e o Gemma3, mostraram-se adequados para a tarefa de *Question Answering* corporativo quando avaliados sob múltiplas dimensões de qualidade.

5 CONCLUSÕES

Este trabalho teve como objetivo investigar a aplicação de LLMs na tarefa de QA em contexto corporativo, utilizando a técnica de RAG. A partir da realização de experimentos controlados, foi possível avaliar o desempenho de diferentes modelos, bem como identificar suas potencialidades e limitações no suporte à integração de novos funcionários e à disseminação do conhecimento organizacional.

Os resultados obtidos evidenciam que os modelos LLaMa 3.1 e Gemma3 apresentaram o melhor desempenho geral, alcançando altos índices nos critérios de fidelidade, clareza, factualidade e relevância das respostas. Em contrapartida, o modelo Deepseek-R1 mostrou desempenho insatisfatório em praticamente todas as dimensões avaliadas, confirmando sua inadequação para o cenário estudado. Os demais modelos analisados, embora tenham apresentado resultados positivos em aspectos específicos, revelaram limitações importantes em termos de consistência e aderência ao conteúdo original.

O principal achado deste estudo é a demonstração de que a combinação de LLMs com a técnica de RAG constitui uma abordagem promissora para apoiar processos de onboarding de novos funcionários, facilitar a comunicação interna e assegurar maior eficiência na gestão do conhecimento organizacional. A possibilidade de acessar informações relevantes de maneira rápida, consistente e escalável representa um diferencial competitivo para empresas que buscam aumentar sua produtividade e reduzir o tempo necessário para a adaptação de colaboradores.

Apesar dos benefícios constatados, alguns desafios permanecem. O uso de LLMs em conjunto com RAG demanda elevado custo computacional, o que pode inviabilizar sua aplicação em pequenas e médias empresas. Além disso, em grandes organizações, o volume massivo de documentos pode dificultar a indexação e a recuperação eficiente da informação, exigindo estratégias mais robustas de otimização e gerenciamento de dados.

Como perspectiva para trabalhos futuros, sugere-se investigar soluções que reduzam a demanda computacional sem comprometer a qualidade das respostas, como técnicas de compressão de modelos, uso de arquiteturas mais eficientes ou mecanismos híbridos de recuperação. Outra linha de pesquisa promissora consiste no desenvolvimento de estratégias de curadoria e atualização automática da base de conhecimento, de forma a garantir que o sistema se mantenha relevante em contextos de constante mudança.

Em síntese, os resultados alcançados neste trabalho indicam que a aplicação de LLMs combinados ao RAG não deve ser entendida apenas como uma solução pontual, mas como uma mudança estrutural no modo como as empresas podem gerir a comunicação interna e a disseminação do conhecimento. A consolidação dessa abordagem pode não

apenas otimizar processos já existentes, mas também abrir caminho para a inovação contínua em práticas corporativas de gestão da informação.

REFERÊNCIAS

AFZAL, A. *et al.* **Towards Optimizing and Evaluating a Retrieval Augmented QA Chatbot using LLMs with Human in the Loop**. 2024. Disponível em: https://arxiv.org/abs/2407.05925.

BA, J. L.; KIROS, J. R.; HINTON, G. E. Layer normalization. **arXiv preprint arXiv:1607.06450**, 2016.

BRACHTEN, F. *et al.* On the ability of virtual agents to decrease cognitive load: an experimental study. **Information Systems and e-Business Management**, v. 18, 06 2020.

CLARDY, A. Reputation, goodwill, and loss: Entering the employee training audit equation. **Human Resource Development Review**, v. 4, n. 3, p. 279–304, 2005. Disponível em: https://doi.org/10.1177/1534484305278243.

DEEPSEEK-AI *et al.* Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. **arXiv preprint arXiv:2501.12948**, 2025. Disponível em: https://arxiv.org/abs/2501.12948.

DESTA, B. *et al.* **RAG Based QA for Low Resource Languages RAG Based QA for Low Resource Languages**. 2024.

FISCHER, K. *et al.* Question: How do large language models perform on the question answering tasks? answer:. **arXiv preprint arXiv:2412.12893**, 2024. Disponível em: https://arxiv.org/abs/2412.12893.

HE, K. *et al.* Deep residual learning for image recognition. *In*: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [*S.l.: s.n.*], 2016. p. 770–778.

HE, X. *et al.* G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. **arXiv preprint arXiv:2402.07630**, 2024. Disponível em: https://arxiv.org/abs/2402.07630.

JIANG, F. *et al.* **Enhancing Question Answering for Enterprise Knowledge Bases using Large Language Models**. 2024. Disponível em: https://arxiv.org/abs/2404.08695.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models**. 3rd. ed. Independent, 2025. Online manuscript released January 12, 2025. Disponível em: https://web.stanford.edu/~jurafsky/slp3/.

KLEIN, H. J.; WEAVER, N. A. The effectiveness of an organizational-level orientation training program in the socialization of new hires. **Personnel Psychology**, v. 53, n. 1, p. 47–66, 2000. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1744-6570.2000.tb00193.x.

OPENAI *et al.* Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2024. Disponível em: https://arxiv.org/abs/2303.08774.

SALAS, E. *et al.* The science of training and development in organizations: What matters in practice. **Psychological Science in the Public Interest**, v. 13, n. 2, p. 74–101, 2012. PMID: 26173283. Disponível em: https://doi.org/10.1177/1529100612436661.

SIRIWARDHANA, S. *et al.* Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. **arXiv preprint arXiv:2210.02627**, 2022. Disponível em: https://arxiv.org/abs/2210.02627.

SUN, P. Y.; SCOTT, J. L. An investigation of barriers to knowledge transfer. **Journal of Knowledge Management**, v. 9, n. 2, p. 75–90, 04 2005. ISSN 1367-3270. Disponível em: https://doi.org/10.1108/13673270510590236.

SZULANSKI, G. Exploring internal stickiness: Impediments to the transfer of best practice within the firm. **Strategic Management Journal**, v. 17, n. S2, p. 27–43, 1996. Disponível em: https://sms.onlinelibrary.wiley.com/doi/abs/10.1002/smj.4250171105.

VASWANI, A. *et al.* Attention is all you need. *In*: GUYON, I. *et al.* (ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

XU, Q. Practical applications of large language models in enterprise-level applications. **Journal of Computer Science and Artificial Intelligence**, v. 2, p. 17–21, 02 2025.