Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Avaliação de Bases de Dados Técnicas com Modelos de Linguagem de Grande Escala: Um Estudo Aplicado ao Wi-Fi 7

Osiel do Couto Rosa

JUIZ DE FORA
AGOSTO, 2025

Avaliação de Bases de Dados Técnicas com Modelos de Linguagem de Grande Escala: Um Estudo Aplicado ao Wi-Fi 7

Osiel do Couto Rosa

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Edelberto Franco Silva

JUIZ DE FORA

AGOSTO, 2025

Avaliação de Bases de Dados Técnicas com Modelos de Linguagem de Grande Escala: Um Estudo Aplicado ao Wi-Fi 7

Osiel do Couto Rosa

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Edelberto Franco Silva
Doutorado em Computação

Alex Borges Vieira
Doutor em Ciência da Computação

Luciano Jerez Chaves
Doutor em Ciência da Computação

JUIZ DE FORA
14 DE AGOSTO, 2025

Aos meus amigos e familiares.

Aos pais, pelo apoio e sustento.

Resumo

A crescente demanda por tecnologias de Inteligência Artificial impulsiona o uso de Modelos de Linguagem de Grande Escala (LLMs) em diferentes áreas do conhecimento. No entanto, o uso desses modelos em contextos com bases fechadas, como documentos técnicos em PDF, ainda apresenta desafios quanto à precisão das respostas e uso efetivo da base de dados. Essa lacuna se evidencia especialmente quando o objetivo é aplicar LLMs em tarefas que exigem rigor técnico. Diante disso, este trabalho aborda o desenvolvimento de um sistema que utiliza 3 modelos de LLMs, sendo eles, Mistral, Llama 2 e TinyLlama, integrados via a biblioteca Ollama, para analisar documentos de uma base de dados sobre Wi-Fi 7, buscando avaliar sua capacidade de fornecer respostas assertivas e claras. Pesquisas atuais demonstram avanços significativos na integração de IA com bases locais, mas carecem de avaliações práticas em contextos fechados e altamente técnicos. Os testes realizados com 100 perguntas técnicas revelaram que os modelos com RAG apresentaram desempenho superior em assertividade e clareza, sendo o Mistral e o Llama 2 os mais eficazes entre os modelos propostos. Os resultados evidenciam o potencial de uso de LLMs aliados a bases internas para consultas técnicas, reforçando a importância de ajustes contextuais e infraestrutura adequada para sua adoção prática.

Palavras-chave: Inteligência Artificial, Análise de documentação, Modelos de linguagem de Grande Escala, Wi-Fi 7

Abstract

The growing demand for Artificial Intelligence technologies is driving the use of Large Language Models (LLMs) across different fields of knowledge. However, the application of these models in closed environments, such as technical documents in PDF format, still presents challenges regarding the accuracy of the answers and the effective use of the data source. This gap is especially evident when LLMs are applied to tasks that require technical rigor. In this context, this work presents the development of a system that uses three LLMs — Mistral, Llama 2, and TinyLlama — integrated via the Ollama library, to analyze documents from a Wi-Fi 7 knowledge base, aiming to assess their ability to deliver assertive and clear responses. Current research indicates significant progress in integrating AI with local data sources; however, it lacks practical evaluations in closed and highly technical scenarios. Tests conducted with 100 technical questions revealed that models using RAG achieved superior performance in both assertiveness and clarity, with Mistral and Llama 2 being the most effective among those evaluated. The results highlight the potential of LLMs combined with internal knowledge bases for technical queries, reinforcing the importance of contextual adaptation and appropriate infrastructure for their practical adoption.

Keywords: Artificial Intelligence, Documentation Analysis, Language Models, Wi-Fi 7

Agradecimentos

Aos meus pais, pelo encorajamento e apoio durante toda essa jornada, sem o amor e incentivo deles não seria possível alcançar todos os objetivos.

Aos professores dos Cursos de Ciências Exatas e Ciência da Computação, e ao meu orientador Edelberto Franco Silva, pelos ensinamentos e experiências proporcionadas durante o desenvolvimento acadêmico, sem o qual este trabalho não seria desenvolvido.

A Code Empresa Júnior, Atlética de Ciências Exatas e ao Diretório Acadêmico de Ciências Exatas, aos quais pude aprender e amadurecer conhecimentos extracurriculares e que foram cruciais para o meu desenvolvimento pessoal e acadêmico.

A minha noiva e amiga Alessandra, que me acompanhou e incentivou durante toda a carreira acadêmica.

"A vida não é um problema a ser resolvido, mas uma realidade a ser experimentada.".

Soren Kierkegaard

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

802.11be Padrão de Wi-Fi 7.

API Interface de Programação de Aplicações.

CSV Valores Separados por Vírgula.

FAISS Pesquisa de Similaridade de IA do Facebook.

GB Gigabyte.

Gbps Gigabits por segundo.

GHz Gigahertz.

IA Inteligência Artificial.

IEEE 802.11 Padrões das Redes Locais Sem Fio (WLAN).

LGPD Lei Geral de Proteção de Dados.

LLM Modelos de Linguagem de Grande Escala.

MHz Megahertz.

MIMO *Multiplas-entrada, Multiplas-saída*.

MLO Operação Multi-Link.

OFDMA Acesso Múltiplo por Divisão de Frequência Ortogonal.

PDF Formato de Documento Portátil.

PLN Processamento de Linguagem Natural.

RAG Geração Aumentada de Recuperação.

RAM Memória de Acesso Aleatório.

TWT Tempo para Despertar.

Wi-Fi *Wireless Fidelity*.

1 Introdução

1.1 Apresentação do tema

A Inteligência Artificial (IA) tem se consolidado como uma das tecnologias mais influentes da atualidade, permeando diversos setores da sociedade e promovendo transformações significativas em áreas como saúde, indústria, educação e segurança digital. De forma geral, IA pode ser definida como o campo da ciência da computação voltado ao desenvolvimento de sistemas capazes de simular comportamentos inteligentes, aprendendo com dados, interpretando padrões e tomando decisões de maneira autônoma ou semiautônoma (GOODFELLOW; BENGIO; COURVILLE, 2016).

Dentre os ramos mais relevantes da IA moderna, destaca-se o Processamento de Linguagem Natural (PLN), que visa permitir que máquinas compreendam, interpretem e gerem linguagem humana. Essa tecnologia tem modificado de forma profunda o modo como indivíduos e organizações acessam, processam e gerenciam informações. Entre os avanços mais expressivos na área, destacam-se os Modelos de Linguagem de Grande Escala (LLM), (MAKRIDAKIS, 2017). Tais modelos têm sido utilizados em aplicações que vão desde assistentes virtuais e sistemas de busca até soluções voltadas para análise técnica de documentos.

O avanço da IA tem permitido o desenvolvimento de sistemas cada vez mais capazes de lidar com grandes volumes de informação e gerar respostas contextualizadas com alto grau de coerência. Entretanto, a maioria dessas soluções está atrelada a plataformas comerciais amplamente disponíveis, que impõem restrições quanto ao uso de arquivos locais, seja por limites de tamanho, formato ou quantidade, além de, muitas vezes, dependerem de conhecimento treinado em dados amplos da internet, o que pode ser um entrave em ambientes que exigem privacidade(KAN et al., 2024; JÚNIOR et al., 2025).

Diferentemente de sistemas que operam com acesso irrestrito à internet, modelos integrados a bases internas oferecem maior controle sobre a origem das informações, promovendo segurança, consistência e confiabilidade nas respostas geradas(CHEN et al.,

2025). O uso de LLM em bases de dados privadas representa um salto significativo no campo da IA aplicada à gestão do conhecimento. Neste contexto, tecnologias como RAG vêm ganhando destaque ao combinar a geração de linguagem com a busca vetorial de trechos relevantes em bases específicas(KLESEL; WITTMANN, 2025).

Enquanto iniciativas recentes têm explorado o uso de LLM em domínios especializados como redes 5G (KAN et al., 2024) e análise de normas técnicas da 3GPP (JÚNIOR et al., 2025), este trabalho investiga a aplicação de três modelos de linguagem de código aberto, sendo eles LLaMA 2[1], Mistral[2] e TinyLlama[3], na análise de uma base documental técnica composta por materiais especializados sobre o padrão *Wireless Fidelity* (Wi-Fi) 7. Devido a complexidade geral das tecnologias Wi-fi, optou-se por delimitar a base aos conteúdos do Wi-fi 7, tecnologias como operação em múltiplas bandas simultâneas, modulação 4K-QAM, e o uso de canais de até 320 Megahertz (MHz) exigem análise de documentos altamente técnicos e constantemente atualizados.

Ao configurar os modelos para atuarem exclusivamente sobre a base fornecida, excluindo qualquer possibilidade de consulta externa, foi possível avaliar seu desempenho em um ambiente controlado. Essa estratégia permite observar como os modelos se comportam quando limitados a fontes previamente definidas, cenário comum em empresas e instituições que trabalham com dados sensíveis ou proprietários.

A ideia é que o sistema seja escalável e replicável, permitindo futuramente a integração com modelos pagos e mais robustos, mas mantendo como premissa o uso restrito às bases fornecidas. Dessa forma, busca-se viabilizar respostas contextualizadas, com base em informações internas, atendendo a uma demanda real e recorrente de ambientes fechados.

1.2 Motivação

A motivação deste trabalho é sustentada por interesses que envolvem tanto o campo acadêmico quanto aplicações práticas no mundo real. Do ponto de vista acadêmico, o projeto contribui para a análise comparativa do desempenho de diferentes LLMs quando

[1] ⟨https://ollama.com/library/llama2⟩
[2] ⟨https://ollama.com/library/mistral⟩
[3] ⟨https://ollama.com/library/tinyllama⟩

aplicados a uma base técnica específica. Essa abordagem permite investigar como esses modelos se comportam em cenários controlados, alimentados exclusivamente por documentação interna, promovendo o avanço do conhecimento na área de IA aplicada à interpretação de textos técnicos.

Sob a perspectiva prática, busca-se desenvolver uma solução funcional, de fácil adaptação e replicação, que possa ser utilizada por profissionais, estudantes e pesquisadores em diferentes contextos. A proposta atende à necessidade crescente de acessar informações relevantes a partir de grandes volumes de dados não estruturados, como é comum em ambientes corporativos e acadêmicos. Ao permitir que a IA opere diretamente sobre a base de dados privada, o sistema contribui para a agilidade na obtenção de respostas e na redução de esforços manuais para análise e busca de documentação.

1.3 Objetivos

1.3.1 Objetivo Geral

Desenvolver e avaliar um sistema capaz de utilizar LLMs aplicados a uma base documental interna, com foco na extração de informações técnicas e na geração de respostas contextualizadas, baseadas exclusivamente no conteúdo disponível localmente.

1.3.2 Objetivos Específicos

- Investigar o uso de LLMs de código aberto no contexto de análise de documentação interna.
- Implementar uma aplicação em Python, utilizando a biblioteca Ollama, capaz de integrar e comparar diferentes modelos de linguagem.
- Construir uma base de dados local utilizando documentos relacionados ao padrão Wi-Fi 7, garantindo um ambiente controlado para testes.
- Avaliar os modelos segundo critérios de desempenho, incluindo tempo de resposta, consumo de Memória de Acesso Aleatório (RAM), assertividade e clareza das respostas geradas.

- Propor uma solução replicável e adaptável, com potencial de aplicação em contextos acadêmicos, corporativos e técnicos.

1.4 Desafio

O presente trabalho enfrenta desafios relevantes ao propor o uso de LLM de código aberto para consulta a bases técnicas privadas. Primeiramente, destaca-se a necessidade de restringir o conhecimento dos LLMs, garantindo que as respostas sejam baseadas exclusivamente em documentos internos, sem recorrer a informações externas ou à internet. Além disso, há o desafio de gerenciar eficientemente os recursos computacionais, visto que modelos *open source* costumam exigir elevado consumo de memória e processamento, o que pode limitar sua utilização em ambientes com infraestrutura ou orçamento restritos. Outro aspecto crítico é manter a qualidade técnica das respostas, especialmente ao lidar com um padrão emergente e complexo como o Wi-Fi 7, que exige precisão terminológica e coerência conceitual, demandando atenção ao pré-processamento dos dados e à estratégia de consulta. Soma-se a isso a necessidade de definir métricas claras e eficazes para avaliar a qualidade das respostas geradas, considerando aspectos como precisão técnica, fluidez e aplicabilidade prática. Por fim, é fundamental que a arquitetura proposta seja adaptável e escalável, permitindo que outros modelos e conjuntos de documentos possam ser facilmente incorporados, ampliando o potencial de uso da solução desenvolvida.

1.5 Organização

Para alcançar os objetivos propostos, o presente trabalho está constituído por seis capítulos, estruturado para abordar diferentes aspectos desta pesquisa. No Capítulo 1 foi apresentado os conceitos do tema, destacando a importância do estudo, a motivação, objetivos gerais e específicos da pesquisa, além dos principais desafios. Além disso, é oferecida uma visão geral do trabalho. No Capítulo 2 é abordada a fundamentação teórica, com a descrição dos principais termos, conceitos e técnicas utilizados para o desenvolvimento da pesquisa, proporcionando a base necessária para a compreensão do tema. No Capítulo 3 são revisados os trabalhos relacionados, apresentando pesquisas anteriores que contribuíram

para o desenvolvimento deste estudo. Essa revisão de literatura discute métodos e tecnologias relevantes para o desafio proposto. No Capítulo 4 é detalhada a metodologia aplicada, incluindo os procedimentos, técnicas e etapas de implementação seguidos para a realização do trabalho. No Capítulo 5 são apresentados os experimentos realizados, os resultados obtidos, bem como a avaliação e comparação entre os modelos propostos. Nesta seção, também são discutidas as vantagens e limitações de cada modelo. No Capítulo 6, por fim, são expostas as conclusões finais do trabalho, discutindo os principais resultados, contribuições do estudo, limitações e sugestões para trabalhos futuros relacionados ao tema.

2 Fundamentação Teórica

Esta seção apresenta os principais conceitos e tecnologias que sustentam o desenvolvimento da solução proposta, com foco na consulta a bases técnicas privadas por meio de LLM. Inicialmente, abordam-se os avanços no padrão de comunicação sem fio Wi-Fi 7, contexto técnico central utilizado para os testes e avaliações do projeto. Em seguida, discute-se o funcionamento e a arquitetura dos LLMs, destacando suas capacidades, limitações e implicações no uso local. Por fim, são detalhadas as ferramentas e bibliotecas empregadas na construção da solução, incluindo tecnologias como Ollama, LangChain, PyPDFLoader, Pandas, vetorização via embeddings, Pesquisa de Similaridade de IA do Facebook (FAISS) e a técnica RAG.

2.1 Wi-Fi e Avanços do Padrão Wi-Fi 7

A tecnologia Wi-Fi consiste em um conjunto de normas para redes locais sem fio, definidos e mantidos pela Wi-Fi Alliance, uma organização global responsável por garantir a interoperabilidade entre dispositivos que seguem o Padrões das Redes Locais Sem Fio (WLAN) (IEEE 802.11), garantindo interoperabilidade entre dispositivos compatíveis. Essa tecnologia possibilita a troca de dados via ondas de rádio, eliminando a necessidade de cabeamento físico e permitindo maior mobilidade aos dispositivos conectados, como smartphones, notebooks, sensores e equipamentos industriais (REVIEW, 2023).

Desde sua introdução em 1997 com o padrão IEEE 802.11, o Wi-Fi passou por sucessivas evoluções, conhecidas pelas letras que identificam as emendas ao padrão original (Associação de Padrões IEEE (IEEE SA), 2023):

- O 802.11a, lanãdo em 1999, operava na banda de 5 GHz com taxas de até 54 Mbps, oferecendo maior velocidade porém com alcance reduzido.
- O 802.11b, lançado em 1999, popularizou o Wi-Fi residencial com taxas de até 11 Mbps;

- O 802.11g (2003) e o 802.11n (Wi-Fi 4, 2009) elevaram gradualmente a taxa de dados e a estabilidade do sinal;

- O 802.11ac (Wi-Fi 5, 2013) trouxe suporte à banda de 5 Gigahertz (GHz) e *Multiplas-entrada, Multiplas-saída* (MIMO);

- O 802.11ax (Wi-Fi 6, 2019) introduziu Acesso Múltiplo por Divisão de Frequência Ortogonal (OFDMA) e Tempo para Despertar (TWT), otimizando a comunicação com múltiplos dispositivos e reduzindo consumo de energia;

Em comparação com o Wi-Fi 6 (802.11ax) e Wi-Fi 6E (introdução da faixa de 6 GHz), o padrão IEEE 802.11be (Wi-Fi 7) traz avanços notáveis: velocidades teóricas de até 46 Gigabits por segundo (Gbps), redução expressiva de latência e mecanismos para operação simultânea em múltiplas bandas, ideal para ambientes densos com exigência de alta performance (LIU et al., 2023). Entre os recursos fundamentais do Wi-Fi 7 estão:

- **Operação Multi-Link (MLO)**, que permite transmissão simultânea em bandas de 2.4GHz, 5GHz e 6GHz;

- **Canais de até 320MHz**, dobrando a largura de banda do Wi-Fi6;

- **Modulação 4K-QAM**, aumentando a densidade de bits por símbolo e eficiência espectral;

- **Redução de latência e jitter**, adequada a aplicações críticas como realidade aumentada e automação industrial;

Essas inovações tornam o Wi-Fi 7 uma escolha promissora para ambientes de conectividade exigente, como os apresentados por (KAN et al., 2024) e (JÚNIOR et al., 2025).

A Tabela comparativa 2.1, ilustra as principais características e avanços tecnológicos do Wi-Fi 6, Wi-Fi 6E e Wi-Fi 7.

Tabela 2.1: Comparativo técnico entre os padrões Wi-Fi 6, 6E e 7

Característica	**Wi-Fi 6**	**Wi-Fi 6E**	**Wi-Fi 7**
Ano de Lançamento	2019	2020	2024
Bandas de Operação	2.4 GHz e 5 GHz	2.4, 5 e 6 GHz	2.4, 5 e 6 GHz
Largura de Canal Máxima	160 MHz	160 MHz	Até 320 MHz
Modulação	1024-QAM	1024-QAM	4096-QAM (4K-QAM)
Velocidade Teórica Máxima (1 link)	Até 9,6 Gbps	Até 9,6 Gbps	Até 46 Gbps
OFDMA	Sim	Sim	Sim
MU-MIMO (uplink e downlink)	Sim	Sim	Melhorado (até 16 streams)
Tempo de Latência	< 20 ms	< 20 ms	< 5 ms
Multi-Link Operation (MLO)	Não	Não	Sim
Target Wake Time (TWT)	Sim	Sim	Sim

2.2 Modelos de Linguagem de Grande Escala

Modelos de Linguagem de Grande Escala são sistemas de IA treinados para compreender, interpretar e gerar linguagem natural. Baseados em arquiteturas de redes neurais profundas, esses modelos são alimentados com grandes volumes de texto, o que lhes permite identificar padrões linguísticos, prever palavras e gerar respostas coerentes a partir de comandos fornecidos pelo usuário (VASWANI et al., 2017; BROWN et al., 2020).

Os LLMs representam um avanço significativo no campo do PLN, possibilitando aplicações como tradução automática, redação assistida, atendimento virtual, análise de sentimentos e sumarização de documentos. Sua capacidade técnica é viabilizada por redes com milhões ou bilhões de parâmetros treinados via aprendizagem supervisionada e reforço

(BROWN et al., 2020).

A arquitetura dos LLMs combina camadas transformadoras com mecanismos de atenção que alocam peso diferente a cada *token* de entrada. Os parâmetros são valores ajustáveis armazenados em matrizes de grande dimensão, que representam as conexões em diferentes camadas do LLM, a quantidade de parâmetros impacta diretamente o nível de conhecimento do modelo, por exemplo, um modelo de "7B"possui aproximadamente 7 bilhões de parâmetros. Durante a geração, o modelo converte texto em tokens, utiliza *embeddings* e aplica múltiplas camadas de atenção para prever o próximo *token* com base no contexto, repetindo o processo até formar a resposta completa (VASWANI et al., 2017).

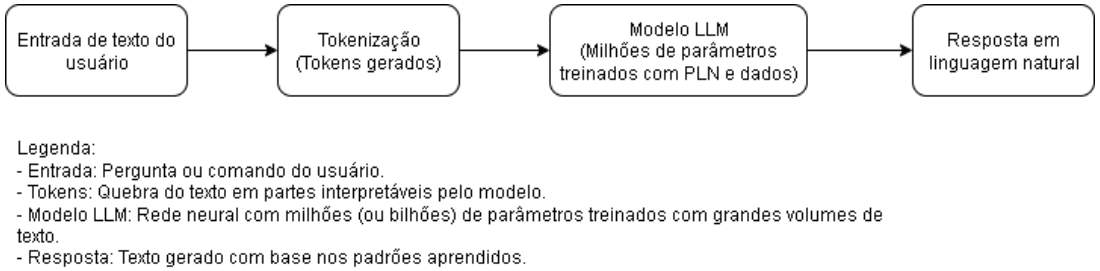

Figura 2.1: Fluxo de funcionamento do LLM

Neste trabalho, utilizamos três LLMs *open-source* com características distintas:

- **LLaMA 2**: desenvolvido pela Meta AI, disponível em tamanhos de 7B, 13B e 70B. É valorizado pelo equilíbrio entre qualidade de geração e eficiência computacional, possibilitando uso local com desempenho excelente (TOUVRON et al., 2023).

- **Mistral**: desenvolvido pela Mistral AI, destaca-se por janelas de contexto ampliadas e otimização do mecanismo de atenção, entregando alta capacidade de raciocínio lógico com uso reduzido de memória (WOLF et al., 2023; SANTOS, 2024).

- **TinyLlama**: modelo compacto com cerca de 1.1B (*open-source*), ideal para ambientes com hardware limitado. Permite avaliar limites mínimos de desempenho mantendo funcionalidade essencial (TOUVRON et al., 2024).

A escolha desses modelos permite analisar diferenças em arquitetura, volume de parâmetros, contexto máximo, qualidade da geração e uso de memória. Modelos maiores expressam profundidade conceitual nas respostas, mas exigem mais RAM e poder

computacional; modelos menores operam com agilidade e menor consumo, embora com limitações de precisão.

Alguns parâmetros são cruciais e influenciam o comportamento dos modelos:

- **Temperatura**: define o nível de aleatoriedade na geração de texto. Valores baixos (por exemplo, 0.1) geram saídas mais previsíveis e precisas; valores mais altos (por exemplo, 0.8) incentivam criatividade linguística e diversidade.
- **System prompt**: instrução inicial que orienta o modelo sobre seu papel, tom e conteúdo. Crucial para que o retorno seja compatível com o resultado esperado.

2.3 Tecnologias para Tratamento de Base de Dados Interna

A utilização de LLM com foco em bases documentais privadas requer uma série de tecnologias auxiliares. Estas ferramentas permitem a extração, transformação, indexação e contextualização dos dados, tornando-os acessíveis e interpretáveis por modelos de IA. A seguir, são descritas as principais soluções utilizadas no desenvolvimento deste projeto.

2.3.1 Ollama

Ollama é uma ferramenta desenvolvida para facilitar a execução local de modelos de linguagem, permitindo carregamento, troca e controle de LLMs com simples comandos, sem necessidade de infraestrutura complexa em nuvem. Essa interface provê integração com diversas bibliotecas de PLN, como LangChain, e suporta modelos *open source* populares como LLaMA, Mistral e outros. Seu diferencial está na leveza e facilidade de uso, especialmente em ambientes offline(CARDOSO; ZAMBERLAN; COMPUTAÇÃO,).

2.3.2 LangChain

LangChain é uma biblioteca em Python voltada para o desenvolvimento de aplicações baseadas em linguagem natural, com suporte nativo à integração entre LLMs e fontes

externas de informação. Ela permite a construção de *pipelines* onde modelos de IA interagem com arquivos, bancos de dados e Interface de Programação de Aplicações (API) de forma modular (LANGCHAIN, 2023). No contexto deste trabalho, foi utilizada para orquestrar a comunicação entre os modelos, os documentos e os dados extraídos.

Leitura de Documentos com PyPDFLoader

A extração do conteúdo textual dos documentos técnicos em Formato de Documento Portátil (PDF) foi realizada com a ferramenta PyPDFLoader, disponibilizada pelo módulo langchain_community. Essa classe é uma das interfaces de carregamento de documentos oferecidas pelo ecossistema LangChain e permite importar arquivos PDF diretamente como objetos já estruturados para integração com *pipelines* de vetorização e recuperação de contexto.

O uso do PyPDFLoader proporciona segmentação por página, facilita a análise individualizada de trechos e é compatível com diversos formatos de entrada, sendo uma ferramenta amplamente adotada no contexto de aplicações com LLMs e RAG (LANGCHAIN, 2023).

2.3.3 Pandas

A biblioteca pandas é utilizada para a manipulação de dados estruturados. No escopo deste projeto, foi essencial para a leitura dos arquivos de perguntas (em Valores Separados por Vírgula (CSV)), bem como para a exportação de métricas extraídas durante os testes. (MCKINNEY, 2010).

2.3.4 Vetorização e *Embeddings*

Para que o modelo de linguagem compreenda e relacione trechos da documentação à pergunta feita pelo usuário, é necessário converter esses textos em representações numéricas vetoriais, os chamados *embeddings*. Essa representação preserva a semântica do conteúdo, permitindo a comparação entre significados. Os *embeddings* utilizados neste trabalho foram gerados por modelos especializados, com vetores armazenados em estruturas otimizadas para busca (MIKOLOV et al., 2013).

2.3.5 FAISS

FAISS (*Facebook AI Similarity Search*) é uma biblioteca desenvolvida pela Meta para realizar buscas rápidas em grandes volumes de vetores. Ao organizar os *embeddings* dos documentos em um índice eficiente, ela permite localizar os trechos mais relevantes para uma pergunta de forma quase instantânea. Sua aplicação é comum em sistemas baseados em RAG (JOHNSON; DOUZE; JÉGOU, 2019).

2.3.6 RAG

O RAG é uma técnica que integra LLM com mecanismos de recuperação de informação. Ao invés de depender unicamente do conhecimento pré-treinado do modelo, o RAG realiza uma etapa de busca em uma base textual específica antes de gerar a resposta. Essa busca retorna trechos relevantes, que são incorporados como parte do contexto apresentado ao modelo gerador. Com isso, o sistema consegue produzir respostas mais precisas, atualizadas e coerentes com a base de conhecimento fornecida, mesmo que o conteúdo em questão não tenha sido incluído durante o treinamento do modelo. Essa abordagem é especialmente útil em cenários que exigem respostas baseadas em fontes internas, como documentação técnica ou bases privadas de dados (LEWIS et al., 2020).

3 Trabalhos Relacionados

3.1 Complexidade das tecnologias do Wi-Fi

Em (VENKATESWARAN, 2025) é apresentado uma revisão abrangente do mecanismo TWT, introduzido no Wi-Fi 6 e aprimorado para uso em redes mais densas e de maior desempenho, como as baseadas no Wi-Fi 7. A relevância desse trabalho para a presente pesquisa está na demonstração da complexidade técnica envolvida na evolução dos padrões Wi-Fi e na necessidade de respostas precisas e atualizadas sobre essas tecnologias. Nesse contexto, a utilização de LLMs com acesso a documentos técnicos via RAG mostra-se promissora para auxiliar na extração de informações detalhadas e de difícil acesso, desde que avaliada cuidadosamente quanto à sua assertividade e clareza, como proposto neste estudo.

3.2 Inteligência Artificial

O trabalho de (MAKRIDAKIS, 2017)) discutiu a chamada "revolução da inteligência artificial", analisando paralelos entre as tecnologias de computadores e comunicações e os impactos das revoluções industriais precedentes. Ele argumenta que tais transformações moldam substancialmente a forma como indivíduos e empresas consomem serviços, consomem cultura e operam no mercado globalizado, destacando efeitos profundos sobre o emprego, a estrutura organizacional e os modelos de negócios.

3.3 LLM aplicado a dados internos

O estudo realizado em (GUARIZI; OLIVEIRA, 2014) destaca a importância de se utilizar bases de dados internas e protegidas na construção de sistemas inteligentes, enfatizando que a personalização e a conformidade com normas como a Lei Geral de Proteção de Dados (LGPD) são essenciais para garantir decisões clínicas mais seguras e eficazes. Esse

princípio reforça a aplicabilidade de arquiteturas como o RAG em ambientes sensíveis, nos quais modelos como os LLMs podem operar de forma eficiente sem acesso à internet ou a bases externas, promovendo respostas mais confiáveis e alinhadas ao contexto organizacional do usuário. Tal abordagem, embora discutida no contexto da saúde, oferece uma base conceitual relevante para sua adoção em outras áreas que também lidam com conhecimento técnico privado, como é o caso do presente trabalho.

Outro trabalho relevante é o (KAN et al., 2024), onde é proposto o uso de um modelo de linguagem de código aberto, baseado na arquitetura Llama, com ajuste fino por instrução para tarefas de análise em redes móveis 5G. O artigo detalha um pipeline de ajuste supervisionado e um conjunto de dados voltado para redes, demonstrando que o modelo ajustado supera abordagens genéricas em precisão e relevância técnica.

4 Metodologia

A metodologia proposta neste trabalho visa avaliar a eficácia de diferentes LLMs aplicados sobre uma base de dados interna, utilizando a arquitetura RAG para geração de respostas técnicas no contexto do Padrões das Redes Locais Sem Fio (WLAN) (IEEE 802.11). O objetivo é explorar como diferentes modelos de linguagem, operando sob restrições de segurança e privacidade, respondem a questões especializadas sem recorrer a dados externos, mantendo coerência, precisão e contextualização.

Para isso, foi desenvolvida uma plataforma experimental composta por:

- Um sistema de carregamento e pré-processamento de documentos técnicos em formato PDF, utilizando bibliotecas como *PyPDFLoader* e *LangChain* para segmentação e vetorização textual.
- Um banco vetorial construído com FAISS, responsável por armazenar os embeddings semânticos dos documentos e facilitar a recuperação dos trechos mais relevantes a cada consulta.
- Três modelos LLM selecionados para os testes (**LLaMA 2**, **Mistral** e **TinyLlama**) executados localmente via *Ollama* com os mesmos parâmetros de geração (como temperatura, contexto e *system prompt*).

A Figura 4.1 ilustra a arquitetura geral adotada. Inicialmente, o usuário realiza uma pergunta técnica. Essa entrada é usada para recuperar trechos semânticos e contextualmente relevantes no banco vetorial, que por sua vez são repassados como contexto ao modelo de linguagem. O modelo então gera a resposta, respeitando as instruções previamente definidas no *system prompt*, garantindo que apenas o conteúdo da base interna seja utilizado.

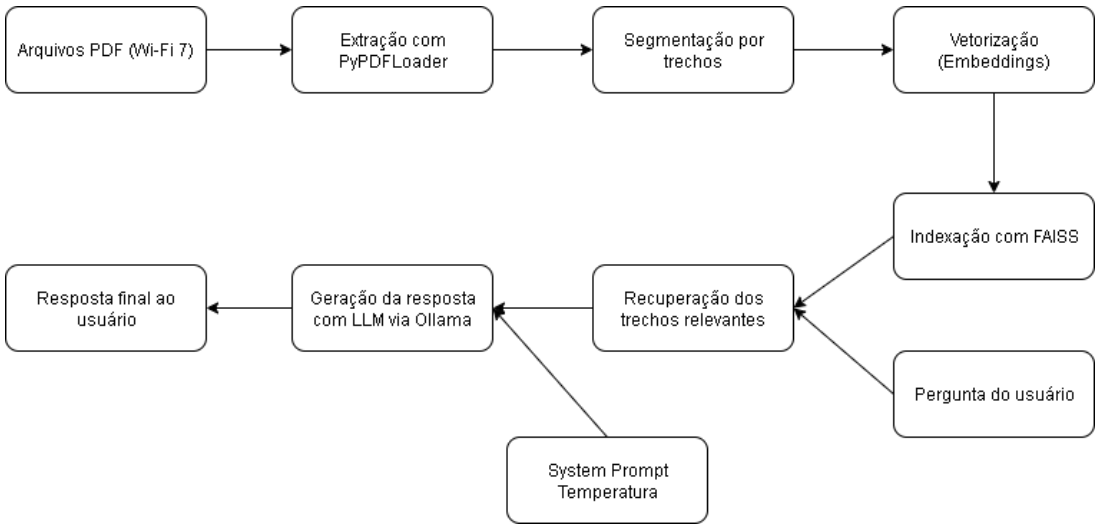

Figura 4.1: Arquitetura geral do sistema baseado em RAG com LLMs locais

Os testes foram estruturados em uma bateria de 100 perguntas objetivas elaboradas sobre conteúdos relativos ao Wi-Fi 7. Cada modelo foi executado com as mesmas instruções e base de dados, sendo avaliado segundo critérios como:

- **Tempo de resposta**: medido em segundos por execução.
- **Uso de memória RAM**: monitorado durante a execução das 100 perguntas.
- **Clareza e Assertividade**: avaliado manualmente a partir de uma classificação qualitativa..

Todos os modelos foram executados exclusivamente com quantização 4-bit para viabilizar a execução em hardware limitado, essa técnica otimiza a precisão númerica dos parâmetros de um LLM para 4 bits por valor, diminuindo drasticamente o uso de memória e acelerando a inferência, com uma perda mínima de qualidade.

Essa metodologia permite, portanto, mensurar não apenas o desempenho técnico dos modelos, mas também sua capacidade de adaptação a cenários restritos, característica essencial em aplicações reais que demandam segurança, como saúde, direito, engenharia e redes privadas corporativas.

4.1 Etapas do Trabalho

O trabalho foi dividido em 6 etapas conforme diagrama da Figura 4.2:

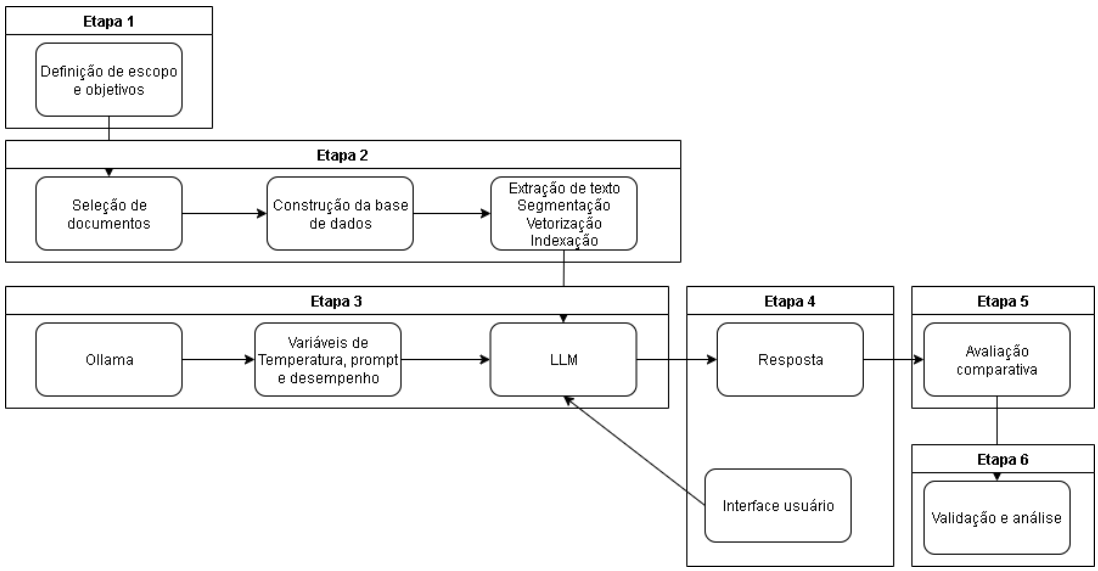

Figura 4.2: Etapas de desenvolvimento

1. Definição do Escopo e Objetivos

   A primeira etapa consistiu na delimitação do tema e na definição dos objetivos gerais e específicos da pesquisa. Este processo incluiu a seleção do Padrão de Wi-Fi 7 (802.11be) como foco de estudo e a escolha de três LLMs para análise e comparação, considerando o contexto de privacidade e eficiência no processamento de dados técnicos.

2. Construção da Base de Dados

   Para a criação de uma base de dados privada, foram realizadas as seguintes atividades:

   - Seleção de Documentos: Coletaram-se artigos acadêmicos, padrões técnicos e documentos relacionados ao Wi-Fi 7.
   - Organização dos Dados: Os documentos foram convertidos para um formato textual uniforme e indexados para possibilitar buscas rápidas e precisas.

3. Configuração dos Modelos de Linguagem

   Os três LLMs escolhidos foram configurados localmente para operar exclusivamente com os dados da base privada.

4. Desenvolvimento do Sistema

O sistema foi projetado em python utilizando a plataforma Ollama para executar modelos de linguagem e tecnologias de recuperação de informação e vetorização para injetar os trechos de documentos internos no prompt do LLM.

5. Avaliação Comparativa

   Para verificar a eficiência dos LLMs, foi aplicada uma bateria de perguntas relacionadas ao Wi-Fi 7. As respostas foram avaliadas para realizar uma comparação dos resultados de cada modelo e contabilizado o tempo de resposta para cada pergunta, além do consumo de memória RAM a cada minuto da execução do sistema.

6. Validação e Análise dos Resultados

   Os resultados foram analisados com base nas métricas definidas, permitindo a comparação entre os modelos e a identificação daquele mais adequado ao contexto proposto.

5 Experimentos e Resultados

Esta seção apresenta de forma detalhada os recursos empregados, as etapas de implementação e os principais desafios enfrentados durante a execução deste trabalho.

Inicialmente, são descritos os componentes de software e hardware utilizados, assim como a estrutura adotada para organização dos dados, vetorização de conteúdo e recuperação semântica de trechos relevantes. Em seguida, detalha-se a configuração dos modelos utilizados, incluindo os parâmetros ajustados, os prompts e as estratégias de controle de contexto, é exibido os modos de interação do usuário com o sistema e por fim, é demonstrado os resultados da pesquisa.

5.1 Preparação do Ambiente Experimental

Os experimentos foram conduzidos em um ambiente computacional isolado, com as seguintes configurações:

- **Hardware**:

Tabela 5.1: Componentes de Hardware do Experimento

Componente	**Especificação**
Sistema Operacional	Windows 11 Home Single Language, versão 10.0.26100
Fabricante do Sistema	Dell Inc.
Modelo do Sistema	Inspiron 3501
Processador	Intel Core i5-1035G1, 4 núcleos, 8 threads, 1.0 GHz (base)
Memória RAM	16 Gigabyte (GB) (15,8 GB utilizável)
Tipo do Sistema	Arquitetura x64

  Inicialmente foi utilizado uma memória RAM de 8GB, contudo, o hardware não suportou a execução dos LLM sendo necessário expansão para 16GB.

- **Softwares**: Conjunto de bibliotecas python especializadas, incluindo ferramen-

tas para integração com modelos de linguagem local (como langchain-community e langchain-ollama), leitura e extração de conteúdo de documentos PDF (pypdf), criação e consulta de índices vetoriais (FAISS) e manipulação e exibição de dados tabulares (pandas e tabulate).

Tabela 5.2: Tecnologias utilizadas na construção e operação do sistema

Ferramenta	**Função no Projeto**	**Categoria**
Ollama	Execução local de LLMs e gerenciamento de modelos	Infraestrutura / Backend
LangChain	Orquestração de fluxos entre LLMs, documentos e ferramentas externas	Pipeline de IA
PyPDFLoader	Extração de texto dos arquivos PDF em formato compatível com LLMs	Pré-processamento
Pandas	Leitura e escrita de arquivos Excel com perguntas e métricas de teste	Manipulação de dados
FAISS	Indexação e busca semântica de vetores de texto	Vetorização e Recuperação
Embeddings	Conversão de trechos textuais em vetores numéricos	Representação Semântica
RAG	Técnica para combinar recuperação de dados com geração textual	Estratégia de Resposta

- **Base de Dados**: Arquivos técnicos e artigos relacionados ao 802.11be, organizados e indexados para facilitar o acesso pelos modelos.

5.2 Desenvolvimento

O processo de desenvolvimento iniciou-se com a preparação da base de dados técnica utilizada nos testes. A leitura dos arquivos PDF foi realizada com a biblioteca PyPDFLoader, segmentando em trechos menores, com o objetivo de preservar o contexto durante a vetorização. Estes trechos foram codificados em vetores por meio de *embeddings* e armazenados com a biblioteca FAISS, permitindo um pipeline completo com RAG para ser utilizados pelos LLM a serem desenvolvidos.

O desenvolvimento da pesquisa enfrentou restrições impostas pelas limitações de hardware disponíveis inicialmente. A primeira abordagem consistiu na utilização de APIs baseadas em nuvem, como *OpenAI API* e *Fireworks API*, visando executar LLM de maneira eficiente, sem exigir processamento local intensivo. No entanto, apesar da disponibilidade de créditos gratuitos, essas plataformas requeriam o cadastramento de um método de pagamento para liberar o uso além do plano básico, inviabilizando a implementação de uma solução inteiramente gratuita.

Em uma segunda etapa, optou-se pela execução local dos modelos por meio da plataforma Ollama, com tentativa de uso dos modelos definidos na pesquisa. Contudo, o ambiente inicial com 8 GB de memória RAM revelou-se insuficiente, apresentando falhas recorrentes de alocação de memória, mesmo após tentativas de otimização, como a finalização de processos em segundo plano e o ajuste de parâmetros de execução.

Diante dessas dificuldades, uma nova abordagem foi explorada utilizando o software LM Studio, que permite a execução local de modelos LLM mais leves. Ainda assim, os recursos computacionais disponíveis mostraram-se insuficientes para uma operação estável e contínua dos modelos, interrompendo o progresso da pesquisa.

A solução definitiva consistiu na ampliação da capacidade de memória RAM para 16 GB, possibilitando, a partir desse ponto, a execução local dos modelos com estabilidade e viabilizando o desenvolvimento completo da solução.

O sistema final foi desenvolvido em uma arquitetura modular, composta pelas seguintes camadas funcionais, conforme ilustrado na figura 5.1:

- Módulo de Processamento de Documentos: Responsável pela leitura e segmentação dos arquivos PDF que compõem a base técnica, utilizando a classe PyPDFLoader

da biblioteca *LangChain* para extração do conteúdo textual.

- Módulo de Vetorização e Recuperação: Implementa a técnica conhecida como RAG, com vetorização dos documentos por meio do modelo all-MiniLM-L6-v2 da *Hugging Face*, selecionado por sua leveza e boa acurácia sem exigir altos recursos computacionais. Os vetores são armazenados e indexados com a biblioteca FAISS, especializada em busca por similaridade em grandes espaços vetoriais.

- Módulo de Conexão com LLMs: Realiza a interface com os modelos de linguagem por meio da biblioteca Ollama, garantindo controle sobre parâmetros como temperatura (nível de criatividade) e *system prompt* (contextualização e escopo das respostas).

- Módulo de Interface com o Usuário: Controla os diferentes modos de operação e a interação do sistema com o usuário final e fornece a resposta final.

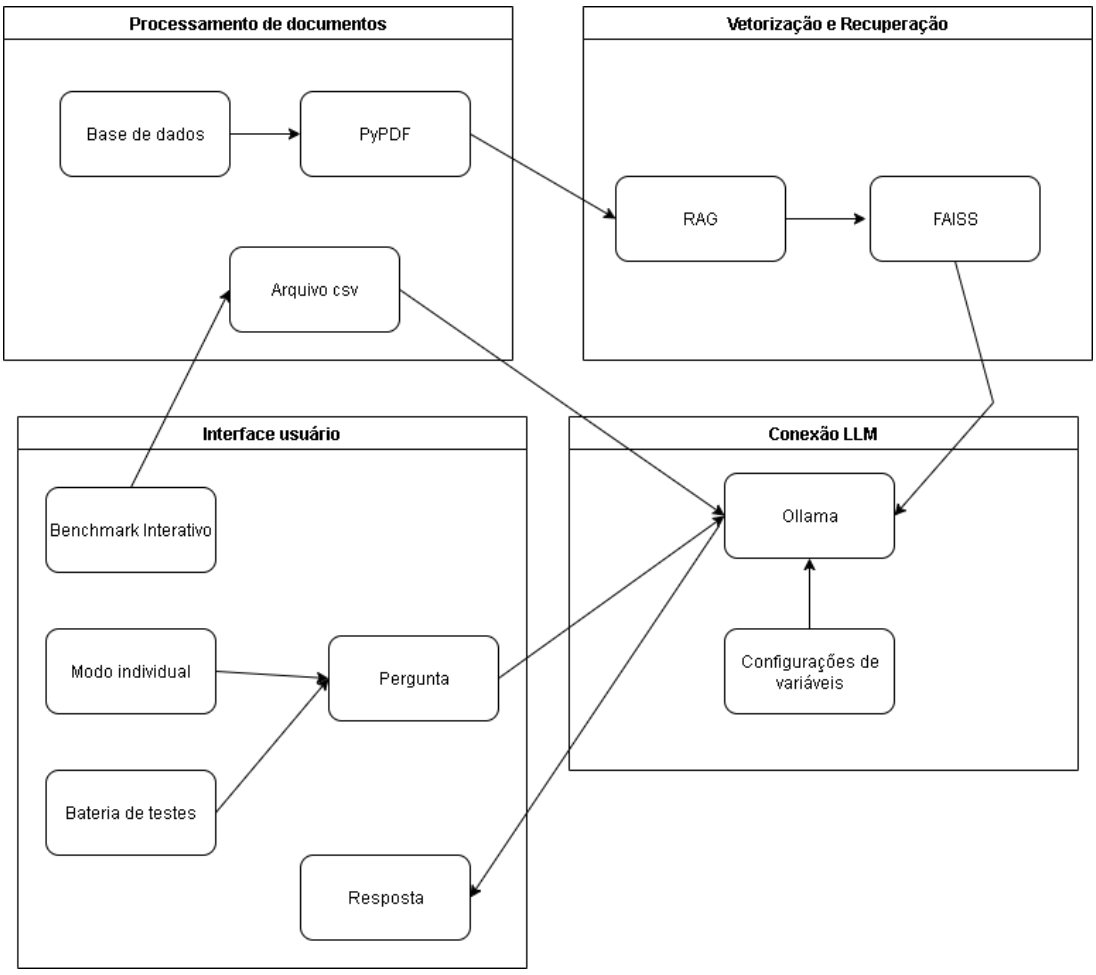

Figura 5.1: Arquitetura do sistema

A base técnica utilizada na pesquisa é composta exclusivamente por documentos em formato PDF armazenados localmente. Os arquivos foram obtidos de diversas fontes distintas. Inicialmente, foi proposto a utilização de documentação e especificações técnicas de documentos privados, contudo, devido a limitação de acesso gratuito aos recursos, não foi possível realizar os testes com documentos sigilosos, com isso, a base de dados foi carregada somente com documentos públicos, garantindo um direcionamento eficaz do LLM.

Durante o pré-processamento, os arquivos são divididos em segmentos menores (chunks) com tamanho e sobreposição ajustáveis, para facilitar a vetorização e recuperação contextualizada.

Os *system prompts* foram configurados para orientar os LLMs conforme sua função. Nos modelos com acesso à base interna, o prompt enfatiza a utilização exclusiva dos dados fornecidos, requerendo rigor técnico e referência às fontes. Já no modelo de controle (sem acesso à base), o prompt é adaptado para estimular explicações didáticas baseadas no conhecimento geral do modelo.

- **Prompt LLM com acesso a base interna**: "Você é um assistente acadêmico especializado em redes sem fio. Suas respostas devem ser em português do Brasil e baseadas estritamente nos documentos fornecidos, com análise crítica e redação original. Responda com foco em Wi-Fi 7 (IEEE 802.11be), explicando os conceitos de maneira clara, objetiva e tecnicamente precisa, sem recorrer a conhecimento externo. Utilize apenas as informações presentes nos documentos, mantendo o rigor técnico, evitando jargões excessivos e garantindo que a resposta seja interpretável por profissionais da área técnica ou estudantes avançados."

- **Prompt LLM sem acesso a base interna**: "Você é um assistente técnico especializado em redes sem fio, com foco em Wi-Fi 7 (IEEE 802.11be). Responda sempre em português do Brasil de forma clara, objetiva e acessível, utilizando termos corretos, mas explicando conceitos de maneira simples. Evite jargões excessivos. Seja direto, didático e preciso, como se estivesse explicando para um profissional da área técnica que deseja respostas rápidas, porém compreensíveis por qualquer pessoa com conhecimento básico em tecnologia. Responda com conhecimento geral (não

use documentos específicos).”

A variável de temperatura foi configurada como 0.3 em todos os modelos, garantindo assim um padrão de retorno do LLM mais técnico e determinístico, contudo, ainda permitindo um grau de criatividade para que as respostas fossem mais claras.

Foram implementados três modos principais de operação e uma opção de sair do sistema, conforme figura 5.2 e descrição:

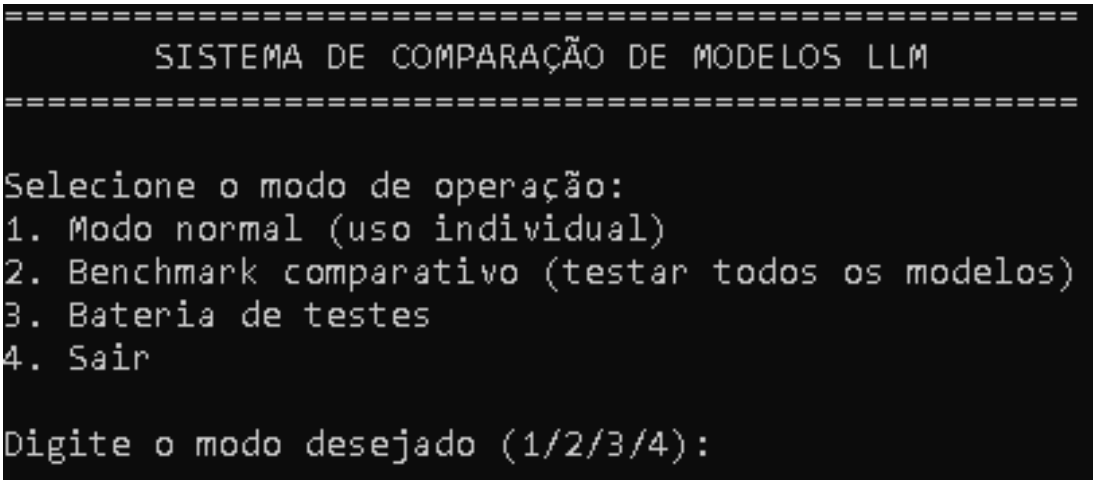

Figura 5.2: Modos de uso do sistema

1. Modo normal Permite a seleção individual de cada LLM disponível, fornecendo a resposta para a pergunta desejada, o tempo de execução e as fontes utilizadas. Esse modo foi o primeiro passo da integração do sistema com o LLM, garantindo que os retornos seriam satisfatórios para a análise de resultados.

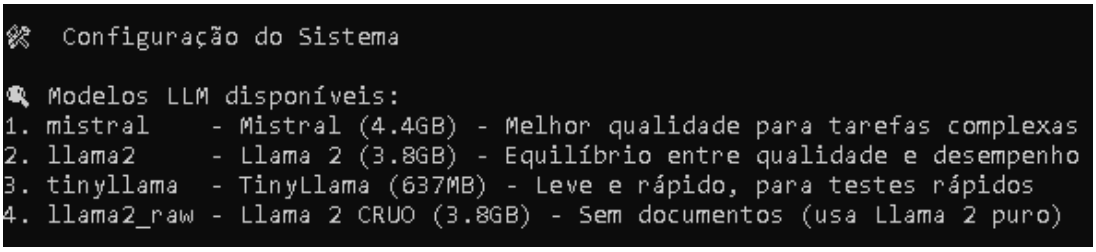

Figura 5.3: Escolha de modelo do LLM

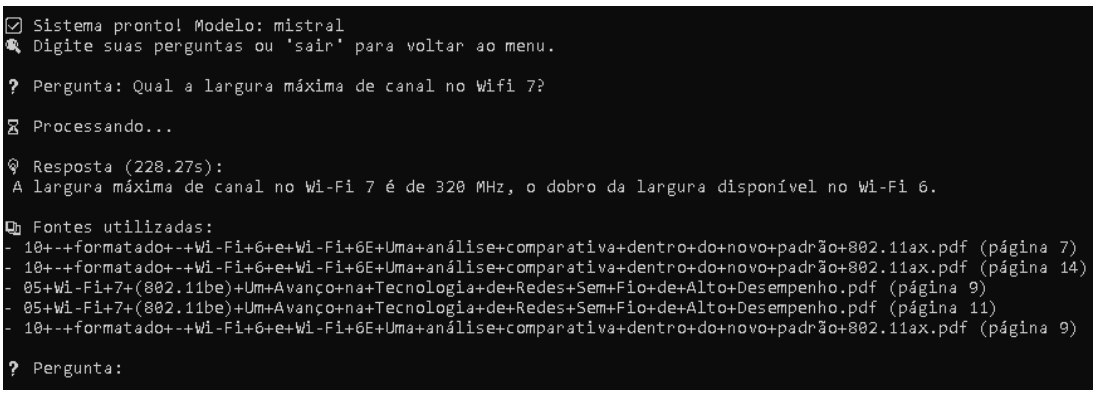

Figura 5.4: Resposta do modo normal

2. Benchmark comparativo: Segunda etapa do desenvolvimento dos LLM, permite o envio simultâneo de uma mesma pergunta para os quatro modelos analisados, o sistema irá exibir o nome do modelo e o tempo de execução para cada pergunta e possibilitará o usuário que realize novas perguntas, ao escolher "sair"será exibido em tela os resultados e feita a exportação do resultado em um arquivo CSV. Esse modo foi essencial para comparação de qualidade e precisão entre os LLMs e garantindo que todos os modelos acessavam corretamente os mesmos conteúdos.

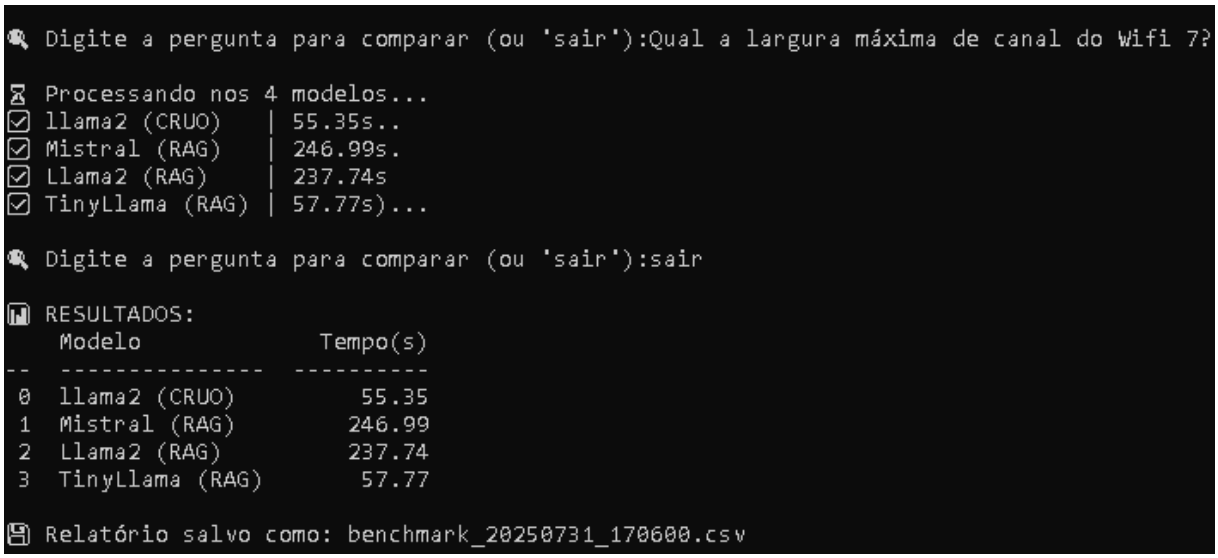

Figura 5.5: Visão modo Benchmark comparativo

A	B	C	D	E	F	G	H	I	J
Modelo	Pergunta	Tempo(s)	Resposta	Fontes					
llama2 (CRUO)	Qual a largura m	55.35	O Wi-Fi 7, també	N/A (resposta crua)					
Mistral (RAG)	Qual a largura m	246.99	A largura máxim	10+-+formatado+-+Wi-Fi+6+e+Wi-Fi+6E+Uma+análise+comparativa+dentro+do+novo+padrão+802.11ax.					
Llama2 (RAG)	Qual a largura m	237.74	De acordo com a	10+-+formatado+-+Wi-Fi+6+e+Wi-Fi+6E+Uma+análise+comparativa+dentro+do+novo+padrão+802.11ax.					
TinyLlama (RAG)	Qual a largura m	57.77	O texto contém t	10+-+formatado+-+Wi-Fi+6+e+Wi-Fi+6E+Uma+análise+comparativa+dentro+do+novo+padrão+802.11ax.					

Figura 5.6: Visão retorno do Benchmark comparativo

A figura 5.6 ilustra um exemplo de arquivo de retorno salvo após execução da bateria de testes, exibindo na integra os modelos, pergunta gerada pelo usuário, tempo de execução e as fontes utilizadas por cada modelo.

3. Bateria de testes: Última etapa do desenvolvimento, possibilita a leitura de um arquivo .CSV externo com diversas perguntas a serem enviadas individualmente para um modelo de LLM escolhido previamente. Em tela é exibido o total de perguntas carregadas, o progresso da execução, a pergunta e o tempo de resposta, ao final é exportado um arquivo CSV registrando o nome do modelo, pergunta, tempo de resposta, resposta do LLM e fonte do conteúdo que foi utilizada como contexto.

```
Modelos disponíveis:
1. mistral    - Mistral (4.4GB) - Melhor qualidade para tarefas complexas
2. llama2     - Llama 2 (3.8GB) - Equilíbrio entre qualidade e desempenho
3. tinyllama  - TinyLlama (637MB) - Leve e rápido, para testes rápidos
4. llama2_raw - Llama 2 CRUO (3.8GB) - Sem documentos (usa Llama 2 puro)

Selecione o modelo (1-4): 1
☑ 100 perguntas carregadas com sucesso
```

Figura 5.7: Visão de escolha do modo Bateria de testes

```
Iniciando teste com mistral para 2 perguntas

Progresso: 1/2
   Pergunta: Qual a largura máxima de canal no Wi-Fi 7?...
   ☑ Respondido em 160.99s

Progresso: 2/2
   Pergunta: Quantas bandas de frequência o Wi-Fi 7 utiliza?...
   ☑ Respondido em 174.27s

Arquivo salvo: resultados_mistral_20250731_1725.csv
Total de perguntas respondidas: 2/2
```

Figura 5.8: Visão do retorno do sistema no modo Bateria de testes

A	B	C	D	E	F	G	H	I
Modelo	Pergunta	Tempo(s)	Tempo corrigido	Resposta	Fontes			
				Wi-Fi 7 é o nomeado para a versão atual do				
				A evolução do Wi-Fi 7 representa um marco				
llama2	O que significa V	226.18	226,18	Em resumo, Wi-Fi 7 é a versão atual do pad	05+Wi-Fi+7+(802.11be)+Um+Avanço+na+Tecnologia+de+Redes+S			
llama2	Qual é o nome té	100.3	100,30	O nome técnico do Wi-Fi 7 é IEEE 802.11be	05+Wi-Fi+7+(802.11be)+Um+Avanço+na+Tecnologia+de+Redes+S			
llama2	Qual a largura m	154.2	154,20	Com base nos documentos fornecidos, a lar	05+Wi-Fi+7+(802.11be)+Um+Avanço+na+Tecnologia+de+Redes+S			
				Com base nos documentos fornecidos, pode				
				Ao contrário do Wi-Fi 5, que tem limitada ca				
llama2	Quantas bandas	196.41	196,41	Em relação à questão de quantas bandas de	05+Wi-Fi+7+(802.11be)+Um+Avanço+na+Tecnologia+de+Redes+S			
				O Wi-Fi 7 (802.11be) suporta um máximo de				
				Ao usar canais de até 320 MHz, o Wi-Fi 7 p				
				Além disso, o Wi-Fi 7 oferece suporte a uma				
llama2	Qual é o máximo	224.54	224,54	Em resumo, o Wi-Fi 7 (802.11be) é capaz de	05+Wi-Fi+7+(802.11be)+Um+Avanço+na+Tecnologia+de+Redes+S			
				O ganho teórico de velocidade do Wi-Fi 7 em				
				Em relação ao Wi-Fi 6, o Wi-Fi 7 apresenta				
llama2	Qual o ganho teó	201.08	201,08	Em resumo, o Wi-Fi 7 oferece um ganho sig	10+-+formatado+-+Wi-Fi+6+e+Wi-Fi+6E+Uma+análise+comparativ			
				O Multi-Link Operation (MLO) é uma tecnolo				
				MLO permite que um único dispositivo, com				
				A arquitetura de MLO é baseada em disposi				
				Existem dois tipos de classificação de um M				
				A implementação do STR consiste na assoc				
llama2	O que é Multi-Lir	252.85	252,85	Em resumo, o Multi-Link Operation (MLO) é	AnaliseDesempenhoWIFI7MLOns3.pdf (p.4); 05+Wi-Fi+7+(802.11b			
llama2	Qual é a modula	152.36	152,36	De acordo com as informações fornecidas n	05+Wi-Fi+7+(802.11be)+Um+Avanço+na+Tecnologia+de+Redes+S			

Figura 5.9: Visão do arquivo de resultado da Bateria de testes

4. Sair : Fecha o programa e finaliza a execução

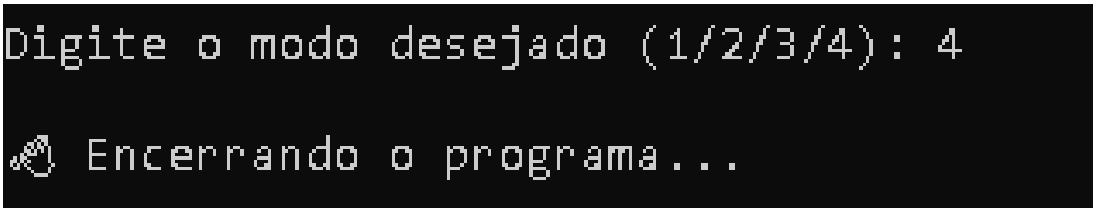

Figura 5.10: Encerramento do sistema

Através do monitoramento do consumo de RAM durante a execução de uma bateria de perguntas para cada modelo, foi possível realizar uma análise quantitativa aprofundada do desempenho de cada LLM.

Como parte complementar ao sistema principal de avaliação dos modelos LLM, foi desenvolvido um mecanismo de monitoramento do consumo de memória RAM com o objetivo de quantificar o impacto de cada modelo sobre os recursos do sistema. Essa funcionalidade foi implementada por meio de arquivos em lote (.bat), os quais acionavam *scripts* em *PowerShell* responsáveis pela coleta periódica dos dados. O monitoramento era realizado a cada minuto, ao longo de um período de duas horas, durante o qual os

modelos processavam as 100 perguntas definidas para os testes. As métricas coletadas incluíam a memória total disponível, memória livre e memória utilizada. Ao término da execução, os dados eram exportados automaticamente em formato .csv, permitindo posterior análise e comparação entre os modelos.

5.3 Execução dos Testes

Para a avaliação dos modelos, foi adotada uma bateria de 100 perguntas técnicas objetivas relacionadas ao padrão Wi-Fi 7. Exemplos de perguntas incluem:

- "O que é Multi-Link Operation (MLO)?"
- "Qual a largura máxima de canal no Wi-Fi 7?"
- "Qual a latência teórica do Wi-Fi 7?"

Os testes foram realizados de forma sequencial para cada um dos quatro modelos LLM, todos executados no mesmo dia, com o notebook em estado de inércia, ou seja, com apenas os serviços essenciais em funcionamento. Essa abordagem visou reduzir interferências externas, garantindo maior controle sobre o ambiente de teste. Durante a execução, foram registrados logs automáticos de consumo de memória RAM, que foram exportados preservando o nome do modelo testado, data e hora da execução, possibilitando a análise de consumo de memória RAM antes, durante e após a execução do sistema.

Tabela 5.3: Monitoramento do consumo de RAM durante execução do LLM Mistral

Timestamp	**Total RAM (GB)**	**RAM Usada (GB)**	**RAM Livre (GB)**
2025-07-23 17:37:54	15,77	11,18	4,59
2025-07-23 17:38:54	15,77	11,14	4,63
2025-07-23 17:39:54	15,77	11,14	4,63
2025-07-23 17:40:54	15,77	11,13	4,64
2025-07-23 17:41:55	15,77	11,13	4,64
2025-07-23 17:42:55	15,77	11,05	4,72
2025-07-23 17:43:55	15,77	11,04	4,73

As respostas e o tempo necessário para cada resposta dos modelos testados foram exportadas individualmente e organizadas em arquivos separados. Em seguida, foi reali-

zada uma análise da média do tempo de resposta e uma análise qualitativa, considerando critérios de assertividade e clareza. Essa avaliação contou com o apoio do orientador, a fim de garantir maior rigor técnico e consistência na análise dos retornos.

A tabela 5.3 exibe alguns exemplos reais dos testes que embasaram os resultados dessa pesquisa e de como os dados foram coletados a partir dos logs e exportação de arquivos.

Tabela 5.4: Exemplo de tabela exportada comparando as respostas dos modelos

Modelo	**Pergunta**	**Tempo resposta (s)**	**Resposta**	**Fontes**
llama2_raw	Qual a largura máxima de canal no Wi-Fi 7?	45.17	O Wi-Fi 7, também conhecido como IEEE 802.11be, tem uma largura de canal máxima de 160 MHz...	N/A (resposta crua)
llama2	Qual a largura máxima de canal no Wi-Fi 7?	154.2	Com base nos documentos fornecidos, a largura máxima de canal no Wi-Fi 7 é de 320 MHz...."	05+Wi-Fi+7+(802.11be)+Um+Avanço+na+Tecnologia+de+Red...
tinyllama	Qual a largura máxima de canal no Wi-Fi 7?	38.3	Na resposta acadêmica acima, o texto diz que o Wi-Fi 7 permite maior capacidade de transmissão de dados...	05+Wi-Fi+7+(802.11be)+Um+Avanço+na+Tecnologia+de+Red...
mistral	Qual a largura máxima de canal no Wi-Fi 7?	121.05	A largura máxima de canal no Wi-Fi 7 é de 320 MHz.	05+Wi-Fi+7+(802.11be)+Um+Avanço+na+Tecnologia+de+Red...

5.4 Resultados Obtidos

Os resultados obtidos foram analisados comparativamente entre os três modelos que utilizaram documentação interna e o modelo de controle, todos foram avaliados seguindo os critérios abaixo:

5.4.1 Assertividade e Clareza das Respostas

A avaliação qualitativa das respostas geradas pelos modelos LLM foi realizada com apoio do orientador e foi conduzida com base em dois critérios principais: **assertividade** e **clareza**. Cada critério recebeu uma pontuação de 1 a 4, conforme descrito abaixo:

- **Assertividade**: avalia o grau de correção da resposta em relação ao conteúdo da pergunta.
  - **Nota 1**: Resposta errada – incorreta ou contradiz a pergunta.
  - **Nota 2**: Incompleta ou inadequada – apresenta apenas parte da resposta correta ou está fora de contexto.
  - **Nota 3**: Parcialmente correta – resposta certa, mas com omissões relevantes ou pequenas imprecisões.
  - **Nota 4**: Correta e completa – tecnicamente precisa e totalmente coerente com a pergunta.
- **Clareza**: avalia a estrutura, linguagem e facilidade de compreensão da resposta.
  - **Nota 1**: Confusa ou ininteligível – mal estruturada, ambígua ou difícil de entender.
  - **Nota 2**: Pouco clara – explicação vaga, técnica demais ou com pouca fluidez.
  - **Nota 3**: Clara – estrutura adequada e compreensível para a maioria dos leitores.
  - **Nota 4**: Muito clara e didática – linguagem simples, objetiva e fácil de entender.

  A tabela 5.5 resume a pontuação média de cada modelo:

Tabela 5.5: Pontuação de assertividade e clareza das respostas.

Modelo	**Assertividade**	**Clareza**
Mistral	3,40	3,70
Llama 2	3,30	3,74
TinyLlama	2,30	2,33
Llama 2-controle	2,93	3,46

Além da análise qualitativa baseada em notas de assertividade e clareza, foram registradas as métricas de tempo de resposta de cada modelo para cada pergunta. A partir desses dados, foram gerados gráficos de dispersão com o objetivo de explorar possíveis correlações entre o desempenho qualitativo (em termos de qualidade da resposta) e o tempo de execução do LLM:

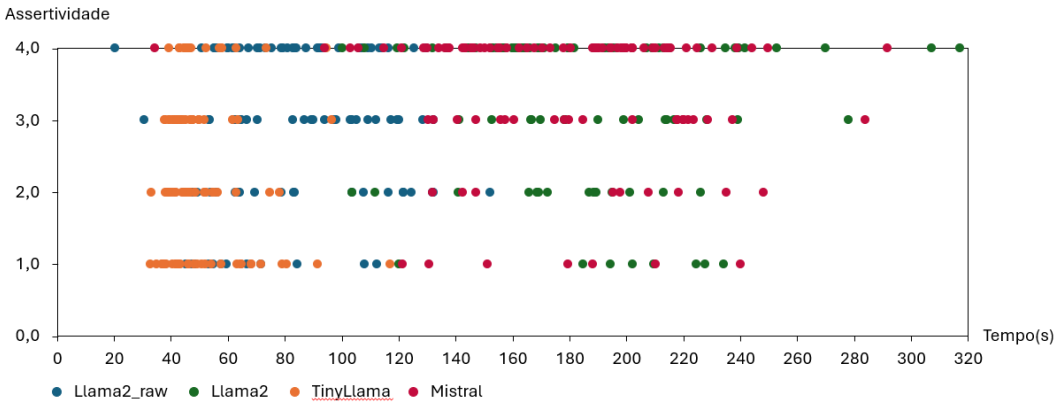

Figura 5.11: Dispersão dos resultados de Assertividade x Segundo

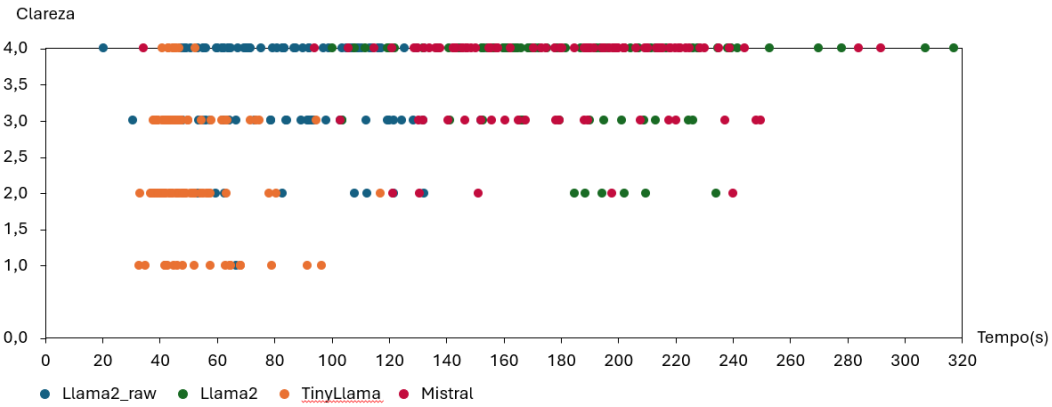

Figura 5.12: Dispersão dos resultados de Clareza x Segundo

As Figuras 5.11 e 5.12 representam a dispersão dos resultados de assertividade e clareza em comparação ao tempo de resposta para cada pergunta, onde é possível observar que quanto maior o tempo de resposta, menor é a incidência de respostas erradas e de difícil compreensão, porém, a maior concentração de resultados na escala de excelência ocorrem em um tempo mediano, entre 100 e 160 segundos.

5.4.2 Eficiência e Custo Computacional

A tabela 5.6 apresenta os tempos médios de resposta de cada modelo, bem como o consumo total de memória RAM durante a execução dos testes. Um monitoramento prévio foi conduzido para aferir o consumo base de RAM do computador em inércia antes da execução do software, que apresentou uma média de 6,80 GB de RAM sendo consumido. No entanto, para garantir a integridade e rastreabilidade dos dados obtidos, os valores apresentados na tabela correspondem ao consumo total registrado durante os testes, conforme extraído diretamente dos arquivos de log.

Tabela 5.6: Eficiência computacional dos modelos

Modelo	**Tempo Médio (s)**	**Consumo de Memória RAM(GB)**
Mistral	180,41	11,16
Llama 2	184,05	11,53
TinyLlama	50,76	8,49
Llama 2-controle	84,99	10,22

Durante os testes, foi observado que o tempo de resposta variava de forma significativa entre as perguntas, especialmente em função da complexidade dos temas abordados, conforme Figura 5.11 e Figura 5.12. No entanto, o consumo de memória RAM manteve-se relativamente estável ao longo de toda a execução para cada modelo, apresentando pequenas variações apenas em momentos pontuais, contudo, pelos testes não foi possível identificar a causa desses pequenos momentos de variação, este comportamento pode ser visualizado através do gráfico conforme Figura 5.13:

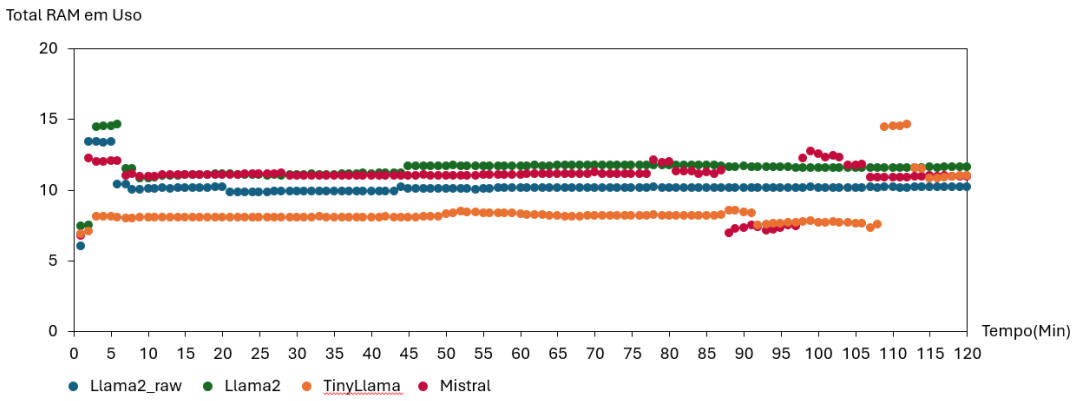

Figura 5.13: Dispersão de consumo de RAM por Minuto

5.4.3 Discussão dos Resultados

A avaliação qualitativa das respostas geradas pelos modelos de linguagem revelou diferenças significativas entre os modelos.

O modelo **Mistral** com média de 3,40 em assertividade e 3,70 em clareza, se caracterizou como um modelo com respostas parcialmente corretas e claras, obtendo o melhor equilíbrio entre precisão e qualidade. Sua pontuação elevada indica que, na maioria dos casos, forneceu respostas corretas e bem estruturadas. A nota elevada em clareza demonstra que suas respostas foram, em geral, bastante didáticas e acessíveis, e o desempenho na assertividade conferem a este modelo um ótimo resultado para cenários onde a assertividade e clareza são essenciais.

O modelo **Llama 2** que fez uso da documentação interna também apresentou bons níveis de assertividade, embora tenha ficado ligeiramente abaixo do Mistral, e obteve a melhor média de clareza entre os modelos testados, a proximidade da nota máxima (4.0) demonstra que este modelo é valioso em alguns cenários onde a clareza da resposta possui alta relevância.

O modelo **Llama 2** utilizado como ponto de controle obteve uma assertividade consideravelmente inferior aos modelos Mistral e Llama 2 com uso de RAG, evidenciando que o uso de documentação interna como contexto contribui para um retorno correto e claro do LLM, porém, por mais que as respostas por vezes fosse incompletas, o modelo apresentou bons resultados de clareza, gerando respostas claras e didáticas mesmo que em alguns casos, inconsistente com o retorno esperado.

O modelo **TinyLlama** teve um desempenho significativamente inferior aos demais nos dois quesitos de qualidade, se caracterizando como um modelo que geralmente fornece respostas incompletas ou inadequadas e com pouca clareza, misturando conceitos importante e não realizando a tradução de alguns materiais.

Portanto, considerando os quesitos de qualidade do retorno, avaliando os dados com relação as métricas de assertividade e clareza, os modelos Mistral e Llama 2 que utilizaram documentação interna como contexto obtiveram os melhores resultados e são capazes de fornecer retornos claros e assertivos para a maior parte das perguntas realizadas.

A análise dos resultados de eficiência computacional revela diferenças significativas entre os modelos avaliados, considerando os critérios de tempo médio de resposta e consumo de memória RAM.

O modelo **TinyLlama** destacou-se como o mais eficiente, apresentando o menor tempo médio de resposta (50,76 segundos) e o menor consumo de memória entre os modelos avaliados (8,49 GB). Esses resultados indicam que o TinyLlama é especialmente vantajoso em cenários com limitações de recursos computacionais, sendo uma alternativa viável para aplicações locais de baixo custo e que precisam obter uma resposta rápida.

Por outro lado, os modelos **Mistral** e **LLaMA 2** apresentaram tempos médios bastante próximos (180,41 s e 184,05 s, respectivamente), mas com consumo de memória ligeiramente superior ao dos demais, superando os 11 GB. Isso sugere que, embora sejam modelos mais robustos em termos de capacidade linguística, são menos eficientes para uso local, exigindo hardware com maior desempenho.

O **Llama 2** de controle, por sua vez, obteve um desempenho intermediário, com tempo médio de 84,99 segundos e consumo de 10,22 GB de RAM, demonstrando que a utilização de LLMs carregados com contexto de documentos internos gera um consumo maior de memória e eleva o tempo de resposta de forma significativa.

O TinyLlama se mostra ideal para soluções mais leves, enquanto os demais modelos exigem maior capacidade de processamento, podendo oferecer respostas potencialmente mais elaboradas ao custo de maior tempo e uso de memória, principalmente quando carregados com documentação interna.

Por fim, levando em consideração todos os quesitos avaliados, o modelo TinyLlama se apresentou como um modelo leve e com respostas rápidas, porém, imprecisas e confusas, sendo indicado somente para cenários onde velocidade e consumo de recursos são as métricas mais relevantes, já os modelos Mistral e Llama 2, exigiram mais tempo de processamento e consumo de recursos, porém, demonstraram bons níveis de clareza e assertividade em relação ao modelo de controle, sendo assim, indicados para uso para análise de documentações internas.

6 Conclusão

Neste trabalho foi desenvolvido e avaliado um sistema capaz de utilizar LLMs para interpretação e resposta com base em documentação técnica interna. A proposta foi aplicar três modelos de código aberto (LLaMA 2, Mistral e TinyLlama) a uma base de dados específica, com foco no Wi-Fi 7, de forma controlada e sem acesso a fontes externas.

Ao longo da execução, foi possível verificar a viabilidade da abordagem adotada, demonstrando que mesmo modelos executados localmente podem gerar respostas relevantes e contextualizadas quando corretamente integrados a uma base documental bem estruturada. Além disso, a análise comparativa entre os modelos permitiu observar diferentes comportamentos em relação as métricas envolvidas.

Os resultados obtidos demonstraram que a utilização de documentação interna como contexto através da aplicação de tecnologias de tratamento geram um ganho considerável de assertividade e clareza nas respostas do modelo ao custo de elevar o tempo de execução e consumo de memória RAM.

Espera-se que este estudo possa contribuir com a comunidade acadêmica e profissional, servindo como referência para soluções futuras em que a IA seja aplicada à leitura e interpretação de informações especializadas. Entre os possíveis desdobramentos, destacam-se aplicações em instituições de ensino, onde assistentes possam oferecer respostas personalizadas com base em regimentos internos, manuais acadêmicos e dados administrativos; em empresas, para automatizar o suporte técnico com base na documentação de software; e na área da saúde, auxiliando na análise de grandes volumes de laudos e prontuários.

Trabalhos futuros poderão expandir o escopo da pesquisa utilizando bases de dados mais diversas, documentos privados que não podem ser acessados por LLMs convencionais, integração com outras ferramentas de vetorização e indexação, ajustes de diferentes *prompts* e variações de temperatura, além da exploração de novos modelos de linguagem com maior capacidade de compreensão semântica e geração textual.

Bibliografia

Associação de Padrões IEEE (IEEE SA). *A evolução da tecnologia e dos padrões Wi-Fi*. 2023. Accessed: 2025-07-30. Disponível em: ⟨https://standards.ieee.org/beyond-standards/the-evolution-of-wi-fi-technology-and-standards/⟩.

BROWN, T. B. et al. Language models are few-shot learners. *Advances in neural information processing systems*, v. 33, p. 1877–1901, 2020.

CARDOSO, L. B.; ZAMBERLAN, A.; COMPUTAÇÃO, C. d. C. da. Módulo LLM local em sistema web para geração de relatórios textuais via prompt.

CHEN, B. et al. *Integrating Access Control with Retrieval-Augmented Generation*. 2025.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. [S.l.]: MIT press, 2016.

GUARIZI, D. D.; OLIVEIRA, E. V. Estudo da inteligência artificial aplicada na área da saúde. In: *Colloquium Exactarum*. [S.l.: s.n.], 2014. v. 6, p. 26–37.

JOHNSON, J.; DOUZE, M.; JÉGOU, H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, IEEE, v. 7, n. 3, p. 535–547, 2019.

JÚNIOR, J. D. A. P. L. et al. Lightweight LLMs for 3GPP specifications: Fine-tuning, retrieval-augmented generation and quantization. In: IEEE. *2025 IEEE 11th International Conference on Network Softwarization (NetSoft)*. [S.l.], 2025. p. 37–42.

KAN, K. B. et al. Mobile-llama: Instruction fine-tuning open-source LLM for network analysis in 5G networks. *IEEE Network*, IEEE, v. 38, n. 5, p. 76–83, 2024.

KLESEL, M.; WITTMANN, H. F. Retrieval-augmented generation RAG. *Business & Information Systems Engineering*, 2025.

LANGCHAIN. *LangChain Documentation*. 2023. Online. Disponível em: ⟨https://docs.langchain.com⟩.

LEWIS, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in neural information processing systems*, v. 33, p. 9459–9474, 2020.

LIU, X. et al. Ieee 802.11be wi-fi 7: Feature summary and performance evaluation. *arXiv preprint arXiv:2309.15951*, 2023.

MAKRIDAKIS, S. The forthcoming artificial intelligence AI revolution: Its impact on society and firms. *Futures*, Elsevier, v. 90, p. 46–60, 2017.

MCKINNEY, W. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, v. 445, p. 51–56, 2010.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, v. 26, 2013.

REVIEW, N. T. Features of ieee 802.11 and the wi-fi alliance. *NTT Technical Review*, v. 21, n. 4, p. 25–32, 2023. Análise da padronização IEEE 802.11 e papel da Wi-Fi Alliance na interoperabilidade.

SANTOS. *Mistral 7B Quantizado vs TinyLlama para Sistemas com Poucos Recursos*. [S.l.]: LinkedIn, 2024. ⟨https://www.linkedin.com/pulse/mistral-7b-quantizado-vs-tinyllama-para-sistemas-com-recursos-santos-kkurf/⟩. Acesso em: jul. 2025.

TOUVRON, H. et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

TOUVRON, H. et al. Tinyllama: Efficient small-scale language model. In: *NeurIPS Workshop on Efficient NLP*. [S.l.: s.n.], 2024.

VASWANI, A. et al. Attention is all you need. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2017. v. 30.

VENKATESWARAN, S. K. Target wake time in ieee 802.11 wlans: Survey and challenges. *Computer Networks*, v. 236, p. 108441, 2025.

WOLF, T. et al. Introducing mistral: A high-performance open-source llm. *Journal of Artificial Intelligence Research*, v. 72, p. 1234–1256, 2023.