# Integração de Recuperação Semântica e Geração com LLMs: Uma Abordagem RAG Aplicada à Teoria Econômica

**Emanuel de Jesus Santos da Silva**[1], **Eduardo Palhares Júnior** [2]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Amazonas (IFAM)
*Campus* Manaus Zona Leste – Manaus, AM – Brasil

[2]Instituto Federal de Educação, Ciência e Tecnologia do Amazonas (IFAM)
*Campus* Manaus Distrito Industrial – Manaus, AM – Brasil

leunamedj@gmail.com, eduardo.palharesjr@ifam.edu.br

***Resumo.*** *Nos últimos anos, a inteligência artificial (IA) tem avançado significativamente, especialmente nos modelos de linguagem baseados em transformadores. Uma das abordagens promissoras é a Retrieval-Augmented Generation (RAG), que combina recuperação de informações com geração de texto para responder perguntas complexas de maneira mais precisa e contextualizada. No entanto, a implementação de RAG enfrenta desafios relacionados à relevância e precisão das respostas, acesso a fontes de dados atualizadas, adaptação a diferentes contextos e consultas, além de questões de explicabilidade e confiança nos resultados. O gerenciamento e a escalabilidade do grande volume de dados também representam um obstáculo para esses sistemas. Diante disso, este estudo discute os principais desafios e oportunidades para o aprimoramento dos modelos RAG na interação humano-máquina.*

***Palavras-chave:*** *RAG. LLM. recuperação semântica. embeddings. geração aumentada de contexto.*

***Abstract.*** *In recent years, artificial intelligence (AI) has advanced significantly, particularly in transformer-based language models. One promising approach is Retrieval-Augmented Generation (RAG), which combines information retrieval with text generation to provide more accurate and context-aware answers to complex questions. However, implementing RAG faces challenges related to response relevance and accuracy, access to updated data sources, adaptation to different contexts and queries, as well as issues of explainability and trustworthiness. Managing and scaling large volumes of data also represents an obstacle for these systems. In this regard, this study discusses the main challenges and opportunities for improving RAG models in human-machine interaction.*

***Keywords:*** *RAG. LLM. semantic retrieval. embeddings. retrieval-augmented generation.*

## 1. Introdução

Nos últimos anos, a inteligência artificial (IA) tem experimentado avanços significativos, especialmente no campo dos modelos de linguagem, como os transformadores. Uma das áreas que tem se destacado é a de Retrieval-Augmented Generation (RAG), que

combina técnicas de recuperação de informações com geração de texto. Esses modelos são projetados para responder de forma mais precisa e contextualizada a perguntas complexas, uma capacidade crucial para a evolução das interações humano-máquina. No entanto, a implementação de RAG em sistemas de respostas a perguntas ainda enfrenta uma série de desafios significativos, que são essenciais para garantir a relevância, a precisão e a confiança nas respostas fornecidas [Proença 2024].

Os desafios atuais enfrentados por modelos de IA na área de respostas a perguntas se relacionam, principalmente, à natureza dinâmica e vasta do conhecimento humano. Ao lidar com questões complexas ou especializadas, os sistemas precisam não apenas entender o conteúdo da pergunta, mas também ter acesso a fontes de dados relevantes e atualizadas para gerar respostas adequadas. Em contextos de múltiplas disciplinas ou de informações em constante mudança, os modelos RAG precisam ser capazes de recuperar dados de fontes amplas e variadas, o que representa um desafio tanto em termos de eficiência quanto de precisão[Wang et al. 2024].

Outro grande desafio é a adaptação desses modelos a diferentes tipos de consulta e a diversidade de contextos nos quais as perguntas podem ser feitas. Uma questão que pode parecer simples para um humano, devido ao contexto compartilhado, pode ser extremamente difícil para um modelo de IA entender e responder adequadamente. Isso ocorre porque os modelos tradicionais de IA muitas vezes carecem da habilidade de integrar nuances contextuais e culturais que são vitais para uma resposta precisa. No caso de modelos RAG, a capacidade de consultar fontes externas e integrar essas informações de maneira coesa é fundamental, mas ainda é um campo em desenvolvimento[Rezaei et al. 2024].

Além disso, há o problema da explicabilidade e confiança nas respostas geradas. Em muitas situações, é fundamental que os usuários possam confiar nas informações fornecidas por sistemas de IA, especialmente em áreas como saúde, educação e segurança. Modelos RAG, ao dependerem de dados externos para formular respostas, podem ter dificuldade em explicar claramente o raciocínio por trás de suas respostas. A falta de transparência e a dificuldade em rastrear a origem das informações podem gerar desconfiança, o que é um obstáculo importante para a adoção generalizada dessas tecnologias[Zhao et al. 2023].

Por fim, outro desafio central é o da escalabilidade e do gerenciamento de grandes volumes de dados. Com o crescimento exponencial da informação disponível na web e em bancos de dados especializados, os modelos RAG precisam ser constantemente atualizados e aprimorados para garantir que possam acessar as fontes mais relevantes de maneira eficiente. A integração desses sistemas com bancos de dados dinâmicos, que podem incluir desde artigos científicos até dados em tempo real, exige infraestrutura robusta e estratégias avançadas de processamento de informações. Assim, os desafios atuais são múltiplos e complexos, mas também oferecem oportunidades para inovações que podem transformar significativamente a maneira como interagimos com a tecnologia para responder às nossas perguntas [Gao et al. 2023].

## 2. Metodologia

A construção do sistema proposto baseia-se na combinação de diferentes tecnologias que permitem integrar mecanismos de recuperação de informações com modelos generativos de linguagem. A estrutura central apoia-se na arquitetura RAG, implementada

com o uso do framework LangChain, que orquestra a comunicação entre os componentes. O armazenamento e a consulta de informações são realizados por meio do banco de dados em grafos Neo4j, enquanto a busca semântica é viabilizada por embeddings gerados pela plataforma Ollama. As respostas finais são produzidas pelo modelo Llama 3.2 (1B), ajustado para lidar com interações em linguagem natural e enriquecer as respostas com base nos documentos recuperados.

## 2.1. Arquitetura RAG

A arquitetura RAG (Geração Aumentada por Recuperação) usando LangChain que é uma estrutura de software ajuda a criar aplicações baseadas em LLM's e melhora as respostas em modelos de IA ao integrar informações externas. Estrutura essa que facilita a implementação do RAG, permitindo a conexão com bancos de dados, vetores de embeddings e documentos estruturados, o que aumenta a precisão e confiabilidade das respostas geradas[Vidivelli et al. 2023], como demonstrada na imagem abaixo.

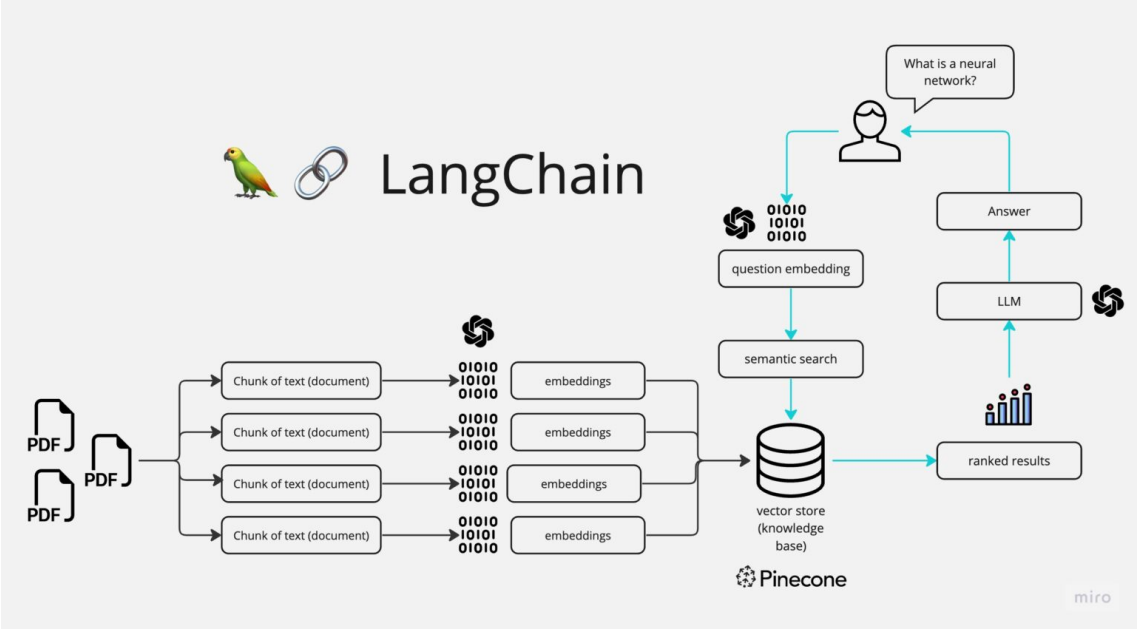


**Figura 1. Arquitetura do Modelo RAG.**

## 2.2. Banco de dados orientado a grafos

O Neo4j é um banco de dados orientado a grafos, projetado para armazenar e consultar dados que podem ser representados como grafos. Ele foi criado para lidar com relações complexas entre dados, sendo ideal para aplicações que exigem interações dinâmicas entre entidades, como redes sociais, sistemas de recomendação e detecção de fraudes [Khan 2023].

Atualmente, o Neo4j continua a ser uma das plataformas de banco de dados em grafos mais utilizadas, com adoção em diversas indústrias para resolver problemas relacionados a dados interconectados e complexos. A ferramenta suporta várias linguagens de consulta, incluindo sua própria linguagem *Cypher*, projetada para ser intuitiva e fácil de aprender [Neo4j 2024].

## 2.3. Modelo Llama 3.2 (1B)

O modelo Llama 3.2 (1B) é um modelo de linguagem de grande escala (LLM), desenvolvido pela Meta, projetado para tarefas avançadas de processamento de linguagem natural. Com versões que variam de 1 a 3 bilhões de parâmetros, o Llama 3.2 (1B) foi escolhido neste trabalho por oferecer uma boa relação entre desempenho, custo computacional e capacidade de geração de respostas contextualizadas. Ele é capaz de lidar com janelas de contexto de até 128.000 tokens, o que o torna apto a trabalhar com entradas extensas e integradas a documentos de diferentes naturezas. Além disso, apresenta suporte multilíngue, incluindo o português, o que o torna particularmente adequado para aplicações em língua portuguesa, como é o caso deste estudo.

A utilização do Llama 3.2 (1B) foi viabilizada por meio da plataforma Ollama, que permite a execução de modelos LLM de forma local. Isso garante maior controle sobre o ambiente de execução, elimina a dependência de APIs externas e reduz os custos associados à computação em nuvem. No sistema implementado, o modelo é responsável pela geração final das respostas, utilizando como base as informações recuperadas previamente no banco de dados vetorizado. Sua configuração foi ajustada para promover respostas mais diretas, por meio do uso de prompts otimizados e limitação no número de tokens gerados, a fim de evitar prolixidade e manter a objetividade.

Além de sua eficiência na geração de respostas, o Llama 3.2 (1B) se destaca por sua adaptabilidade em diferentes fluxos de trabalho baseados em RAG. Neste estudo, ele é integrado à arquitetura LangChain, que gerencia o ciclo de consulta, recuperação e geração, permitindo que o modelo atue de forma contextualizada. A capacidade do Llama em interpretar e articular informações provenientes de múltiplas fontes o torna um elemento-chave para garantir a coesão e a precisão nas interações com o usuário.

## 2.4. Embeddings Semânticos com Ollama

A recuperação de documentos relevantes neste trabalho é baseada em representações vetoriais de textos, conhecidas como embeddings semânticos. Esses vetores numéricos capturam relações de similaridade contextual entre termos, permitindo que consultas textuais sejam comparadas com documentos de forma mais precisa do que abordagens puramente lexicais. No contexto da arquitetura RAG, os embeddings são fundamentais para encontrar os documentos mais relevantes que servirão de base para a geração da resposta final pelo modelo de linguagem.

A geração dos embeddings é realizada por meio da plataforma Ollama, que possibilita o uso local de modelos de linguagem de grande escala (LLMs) para vetorização semântica. A execução local elimina a necessidade de chamadas externas a APIs, reduzindo a latência e aumentando a autonomia do sistema. A plataforma fornece embeddings vetoriais densos e contextualmente ricos, os quais são armazenados no banco de dados Neo4j e utilizados posteriormente na etapa de recuperação semântica.

O sistema implementado segue um fluxo estruturado: o usuário insere uma pergunta por meio de uma interface web, e essa entrada é transmitida para a função *chat_with_bot*. Essa função aciona o método *query_neo4j*, responsável por gerar o embedding da consulta, calcular sua similaridade com os documentos vetorizados previamente no Neo4j e recuperar os conteúdos mais relevantes. O resultado desse processo são os

documentos com maior proximidade semântica em relação à consulta, que serão usados como contexto para a geração da resposta final.

A métrica utilizada para comparar os vetores da consulta e dos documentos é a similaridade de cosseno, definida pela fórmula:

$$silimaridade\,(A, B) = \frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}$$

onde $A \cdot B$ representa o produto escalar entre $A$ e $B$, enquanto que $\|A\|$ e $\|B\|$ são suas respectivas magnitudes. O valor da similaridade varia entre -1 e 1, sendo que valores mais próximos de 1 indicam maior semelhança semântica. Essa métrica permite que o sistema identifique os documentos mais alinhados ao significado da consulta, mesmo quando há variações na forma textual.

Após o cálculo das similaridades, o sistema seleciona os cinco documentos mais relevantes com base na pontuação obtida. Esses documentos são então encaminhados ao modelo gerador (Llama 3.2 (1B)), que os utiliza como contexto para elaborar uma resposta. Essa combinação entre recuperação baseada em embeddings e geração com LLMs é o que caracteriza a arquitetura RAG.

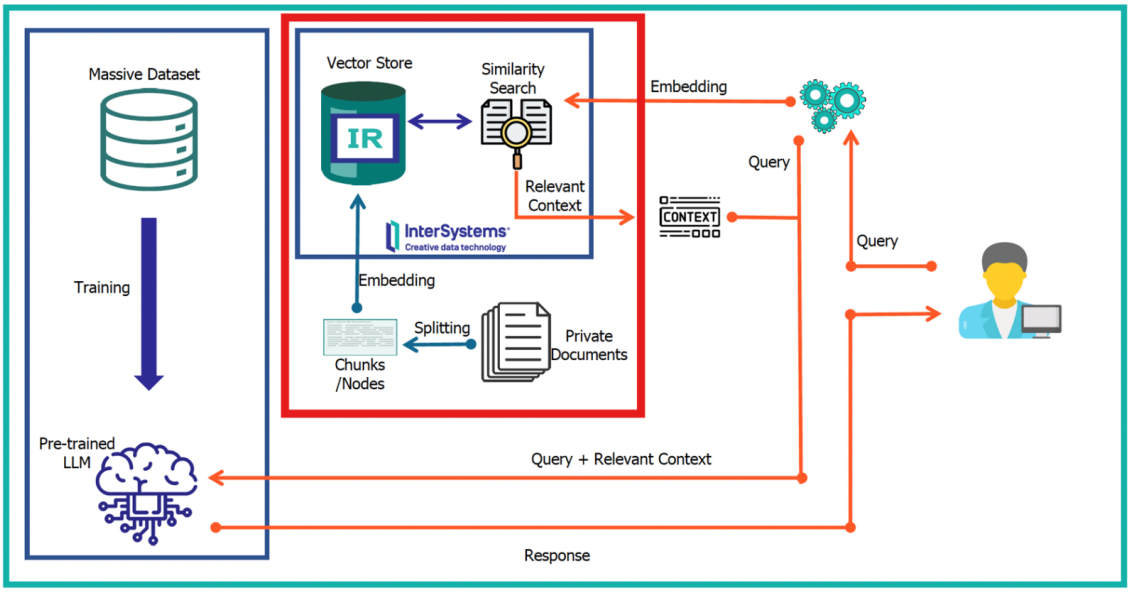


**Figura 2. Modelos LLM e Aplicacoes RAG.**

## 2.5. Métricas de Avaliação das Respostas Geradas

A qualidade das respostas geradas pelo sistema é avaliada por meio de duas métricas amplamente utilizadas na literatura de processamento de linguagem natural: ROUGE-L e BERTScore. Ambas têm como objetivo medir o grau de similaridade entre a resposta gerada pelo modelo e uma resposta de referência, porém adotam abordagens distintas e complementares.

**ROUGE-L:** A métrica ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) baseia-se na identificação da maior subsequência comum (Longest Common Subsequence, LCS) entre as palavras da resposta gerada e da referência. Essa subsequência considera apenas a ordem das palavras, desconsiderando interrupções intermediárias. A partir da LCS, são calculadas três medidas:

- **Precisão:** proporção da LCS em relação ao número de palavras da resposta gerada.
- **Recall:** proporção da LCS em relação ao número de palavras da resposta de referência.
- **F1-score:** média harmônica entre precisão e recall.

Essa métrica é útil por capturar a estrutura sequencial da resposta, mas não considera relações semânticas mais profundas entre palavras. Reformulações de frases, uso de sinônimos ou reordenação dos termos podem resultar em escores baixos, mesmo quando o conteúdo essencial está preservado.

**BERTScore:** O BERTScore, por sua vez, utiliza modelos pré-treinados como o BERT para gerar embeddings semânticos das palavras da resposta gerada e da referência. Com base nesses embeddings, calcula-se a similaridade semântica entre os textos. Assim como o ROUGE-L, o BERTScore fornece medidas de precisão, recall e F1-score, mas com a vantagem de reconhecer sinônimos, reescritas e reformulações estruturais.

Essa métrica é especialmente adequada para avaliar respostas geradas por modelos de linguagem, já que leva em consideração o significado das palavras, e não apenas sua ordem ou exatidão superficial. Por esse motivo, é comum que os valores de F1 obtidos pelo BERTScore sejam significativamente mais altos do que os do ROUGE-L, mesmo quando ambos avaliam as mesmas respostas. Essa diferença será retomada e discutida na seção de resultados.

## 3. Resultados e discussão

Esta seção apresenta os resultados obtidos a partir da aplicação do modelo Llama 3.2 (1B) em dois cenários distintos: com e sem a utilização da arquitetura RAG. O objetivo é comparar a qualidade das respostas geradas em cada abordagem, com base em perguntas elaboradas a partir do conteúdo do livro *A Riqueza das Nações*, de Adam Smith.

Cada pergunta é analisada individualmente, considerando a resposta gerada pelo modelo sem acesso a fontes externas (modelo puro) e a resposta gerada com suporte da arquitetura RAG, que incorpora recuperação semântica via embeddings. Em seguida, é feita uma comparação qualitativa entre as respostas, destacando aspectos de relevância, clareza e aderência ao conteúdo de referência.

Ao final, os resultados são quantificados por meio de métricas objetivas ROUGE-L e BERTScore, permitindo avaliar, de forma mais precisa, o impacto da técnica RAG na geração das respostas. A seção termina com uma discussão interpretativa, que analisa as principais tendências observadas e suas implicações para o uso prático de modelos LLMs com RAG.

### 3.1. Pergunta 1 — Qual o conceito de valor?

A primeira pergunta buscou explorar o entendimento do modelo sobre o conceito de valor, tema recorrente na obra de Adam Smith. A seguir, apresentam-se as respostas geradas em ambos os cenários avaliados.

**Resposta sem RAG:** No primeiro cenário, o modelo Llama 3.2 (1B) foi solicitado a responder diretamente, sem o uso de recuperação de documentos externos. A resposta gerada foi a seguinte:

> *O conceito de valor está relacionado à importância ou utilidade atribuída a um bem ou serviço. Na economia clássica, valor pode ser entendido de diferentes formas: valor de uso e valor de troca. O valor de uso refere-se à utilidade de um bem, enquanto o valor de troca refere-se àquilo que ele pode ser trocado no mercado. Adam Smith discutiu essas ideias ao analisar o paradoxo da água e do diamante, onde a água possui alto valor de uso e baixo valor de troca, e o diamante o oposto.*

**Resposta com RAG:** No segundo cenário, o sistema utilizou a arquitetura RAG, realizando uma busca semântica em uma base de dados contendo trechos da obra de Adam Smith, vetorizados por embeddings gerados via Ollama. Os documentos recuperados serviram como contexto para o modelo Llama 3.2 (1B) gerar a resposta abaixo:

Resposta com RAG

O conceito de valor, segundo Adam Smith, está ligado ao trabalho necessário para produzir um bem. Ele distingue entre valor de uso, que é a utilidade de um bem, e valor de troca, que é a capacidade de ser trocado por outros bens no mercado. O trabalho é a medida real do valor de troca.

#### Comparação qualitativa

A resposta sem RAG apresenta uma explicação geral do conceito de valor, misturando definições econômicas com uma interpretação moderna da teoria clássica. Embora mencione corretamente os conceitos de valor de uso e valor de troca, a explicação se distancia da ênfase original de Adam Smith sobre o papel do trabalho como medida de valor de troca.

Por outro lado, a resposta com RAG demonstra maior fidelidade ao conteúdo da obra, citando explicitamente a relação entre trabalho e valor — elemento central da teoria de Smith. A resposta é mais objetiva, coesa e informativamente densa, refletindo a contribuição da recuperação semântica para o alinhamento com o conteúdo original.

### 3.2. Pergunta 2 — De onde surge o conceito de valor?

A segunda pergunta buscou investigar se o modelo seria capaz de identificar a origem do conceito de valor segundo a abordagem teórica de Adam Smith. O foco estava em verificar se a resposta gerada mencionaria o papel do trabalho, dos fatores de produção ou das relações de mercado na definição do valor.

**Resposta sem RAG:** Sem acesso ao mecanismo de recuperação semântica, o modelo gerou a seguinte resposta baseada exclusivamente em seu conhecimento interno:

> *O conceito de valor surge da interação entre oferta e demanda no mercado. Ao longo da história do pensamento econômico, diversos autores propuseram explicações diferentes para o surgimento do valor. Na tradição clássica, o valor era associado ao trabalho necessário para produzir um bem. Já em abordagens mais modernas, considera-se também a utilidade percebida pelos consumidores e os custos de oportunidade envolvidos na produção.*

**Resposta com RAG:** Com o suporte da arquitetura RAG, o sistema recuperou trechos relevantes do texto original de Adam Smith e gerou a seguinte resposta:

Resposta com RAG

Segundo Adam Smith, o valor dos bens surge da quantidade de trabalho empregada em sua produção. O trabalho é a medida real do valor de troca de qualquer mercadoria.

**Comparação qualitativa**

A resposta sem RAG apresenta uma explicação mais ampla e anacrônica sobre a origem do valor, incluindo conceitos contemporâneos como utilidade e custos de oportunidade que **não fazem parte da teoria original de Adam Smith**. Apesar de mencionar o trabalho como uma possível origem do valor, o modelo dilui essa ideia ao apresentar múltiplas abordagens, sem destacar o ponto central defendido na obra analisada.

Em contraste, a resposta com RAG é concisa, direta e conceitualmente alinhada com o conteúdo da obra. Ao focar na relação entre trabalho e valor de troca, o modelo revela maior precisão e fidelidade ao pensamento de Smith. Isso indica que a recuperação semântica contribuiu para reduzir ambiguidades e garantir maior aderência ao material de referência.

### 3.3. Pergunta 3 — Quais os fatores de produção que geram o conceito de valor?

A terceira pergunta teve como objetivo avaliar se o modelo seria capaz de identificar os fatores que, segundo a teoria clássica, contribuem para a formação do valor dos bens. Esperava-se que o modelo reconhecesse a importância do trabalho como elemento central, bem como as distinções feitas por Adam Smith em relação a outros fatores de produção.

**Resposta sem RAG** Neste cenário, o modelo respondeu com base apenas em seu conhecimento pré-treinado, sem acesso a documentos recuperados da obra de Adam Smith. A resposta gerada foi:

> *Os fatores de produção que contribuem para o conceito de valor incluem o trabalho, o capital e a terra. Na teoria econômica clássica, o trabalho era considerado o principal fator de geração de valor, mas com o*

*tempo, passou-se a reconhecer o papel do capital (máquinas, ferramentas) e da terra (recursos naturais). O valor de um bem é influenciado pela combinação e produtividade desses fatores, bem como pela sua escassez e demanda no mercado.*

**Resposta com RAG** Com o apoio da arquitetura RAG, o sistema recuperou trechos relevantes da obra e utilizou essas informações como base para a geração da seguinte resposta:

Resposta com RAG

De acordo com Adam Smith, o valor de troca de uma mercadoria é determinado pelo trabalho necessário para produzi-la. O trabalho é o fator fundamental que gera valor, especialmente nas sociedades primitivas, onde o trabalho era o único fator de produção.

**Comparação qualitativa**

A resposta sem RAG apresenta uma perspectiva ampla dos fatores de produção, incluindo elementos como capital e terra. Embora esses conceitos façam parte do arcabouço da teoria econômica moderna, a resposta acaba desviando do foco específico da obra de Adam Smith, que enfatiza o trabalho como a principal — e, inicialmente, única — fonte de valor.

A resposta com RAG, por outro lado, mantém-se fiel ao conteúdo da obra analisada. Ela destaca o papel central do trabalho, especialmente no contexto das sociedades primitivas, como ponto de partida para a construção da teoria do valor. Esse foco temático mostra como a recuperação semântica atuou para direcionar o modelo a responder com maior precisão e contextualização histórica, reduzindo a influência de interpretações anacrônicas.

Essa diferença é significativa do ponto de vista metodológico, pois demonstra que o uso de RAG contribui não apenas para enriquecer a resposta com conteúdo relevante, mas também para alinhar o modelo à perspectiva teórica correta dentro do escopo da pergunta.

### 3.4. Avaliação Quantitativa das Respostas Geradas

Além da comparação qualitativa das respostas geradas pelo sistema com e sem RAG, foi realizada uma avaliação quantitativa utilizando duas métricas complementares: ROUGE-L e BERTScore. Os valores analisados referem-se exclusivamente às respostas geradas com o uso da arquitetura RAG, considerando que essa foi a proposta metodológica central do trabalho.

As métricas aplicadas refletem diferentes formas de avaliar a proximidade entre a resposta gerada e uma referência ideal. O ROUGE-L mede a sobreposição textual literal com base em subsequências comuns, enquanto o BERTScore avalia a similaridade semântica entre as palavras com base em embeddings. Devido à natureza distinta dessas abordagens, é esperado que os valores operem em escalas diferentes e apontem qualida-

des distintas das respostas. Os resultados de cada métrica são apresentados e discutidos a seguir.

### 3.4.1. Resultados com ROUGE-L

A figura 3 apresenta os valores do F1-score da métrica ROUGE-L para cada uma das três respostas geradas com o uso da arquitetura RAG.

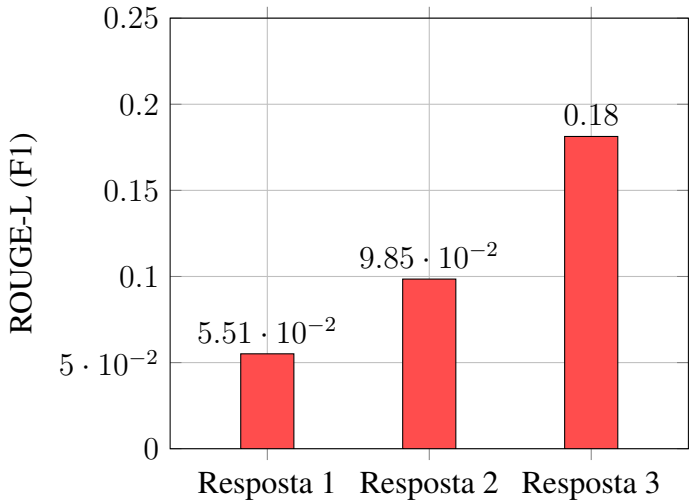


**Figura 3. Desempenho da métrica ROUGE-L (F1-score) nas respostas geradas com a arquitetura RAG para as perguntas analisadas.**

Observa-se que os valores de ROUGE-L variaram entre aproximadamente 0,18 e 0,23. Embora possam parecer baixos à primeira vista, esse resultado está de acordo com o comportamento esperado dessa métrica, que privilegia a sobreposição exata de palavras e a ordem dos termos no texto. Reformulações legítimas ou reorganizações de ideias — comuns nas respostas geradas com RAG — não são reconhecidas como correspondências válidas, o que penaliza o escore final.

Na prática, isso significa que uma resposta pode apresentar excelente aderência ao conteúdo da referência e ainda assim receber uma pontuação modesta em ROUGE-L, especialmente quando reformula ideias ou utiliza sinônimos. Como discutido na metodologia, essa métrica deve ser interpretada com cautela e em conjunto com outras avaliações mais sensíveis à semântica.

### 3.4.2. Resultados com BERTScore

A figura 4 apresenta-se o gráfico com os valores de precisão, recall e F1 obtidos por meio do BERTScore nas três respostas geradas com RAG.

Os valores observados ficaram consistentemente acima de 85%, com destaque para o F1-score, que ultrapassou os 90% nas respostas 2 e 3. Esses resultados indicam forte similaridade semântica entre as respostas geradas e os textos de referência.

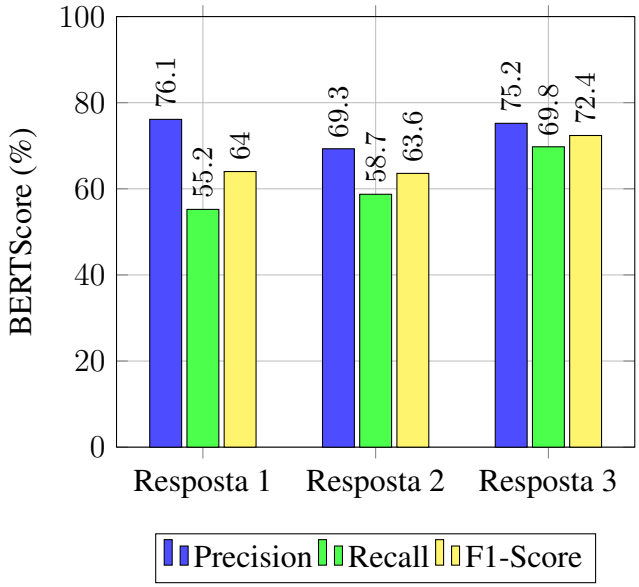


**Figura 4. Desempenho da métrica BERTScore (Precisão, Recall e F1-Score) nas respostas geradas com a arquitetura RAG para as perguntas analisadas.**

Diferentemente do ROUGE-L, o BERTScore é capaz de identificar reformulações corretas, uso de sinônimos e variações sintáticas sem penalizar a pontuação. Isso justifica os valores elevados e reforça a capacidade da arquitetura RAG de gerar respostas semanticamente próximas ao conteúdo original, mesmo quando expressas de forma distinta. A alta pontuação do BERTScore, alinhada às análises qualitativas apresentadas anteriormente, sugere que a técnica de recuperação semântica contribuiu significativamente para a fidelidade conceitual das respostas.

### 3.5. Discussão dos Resultados

A análise das três perguntas revelou diferenças consistentes entre as respostas geradas pelo modelo puro (sem RAG) e aquelas produzidas com o suporte da arquitetura RAG. De modo geral, as respostas com RAG mostraram-se mais precisas, concisas e conceitualmente alinhadas ao conteúdo original da obra de Adam Smith, enquanto as respostas sem RAG tenderam a ser mais genéricas, com traços de interferência de abordagens modernas ou interpretações alternativas.

Essa tendência foi confirmada tanto na análise qualitativa quanto na avaliação quantitativa. No aspecto qualitativo, observou-se que o uso de recuperação semântica permitiu que o modelo se ancorasse diretamente em conteúdos relevantes, reduzindo ambiguidades e evitando extrapolações. Em particular, a presença de conceitos centrais da obra — como o papel do trabalho na determinação do valor — foi mais recorrente nas respostas com RAG.

Do ponto de vista quantitativo, os valores de BERTScore reforçaram essa conclusão: as respostas com RAG obtiveram F1-scores acima de 70%, indicando alta similaridade semântica com as referências. Já os valores de ROUGE-L, embora mais baixos, estão de acordo com o comportamento esperado da métrica, que penaliza reformulações textuais mesmo quando o conteúdo está correto. Isso evidencia a importância de utilizar métricas complementares na avaliação de modelos de linguagem.

Esses resultados indicam que a integração entre mecanismos de recuperação e modelos generativos pode melhorar significativamente a qualidade das respostas em tarefas de pergunta e resposta com base em documentos. No entanto, a eficácia da abordagem RAG depende da qualidade e da cobertura semântica da base de dados utilizada. Em contextos em que o conteúdo de referência é escasso, mal segmentado ou ruidoso, os ganhos observados podem não se repetir.

Além disso, embora as métricas adotadas neste trabalho forneçam uma estimativa objetiva do desempenho, a avaliação automática de respostas geradas por LLMs ainda enfrenta limitações, especialmente em domínios subjetivos ou abertos. A combinação de métodos quantitativos e análise humana continua sendo essencial para uma compreensão mais profunda da performance desses sistemas.

Em síntese, os resultados obtidos sustentam a hipótese de que o uso de RAG aprimora a geração de respostas em tarefas de QA documentado, ao promover maior aderência semântica ao conteúdo de origem. Isso abre caminho para aplicações robustas em ambientes educacionais, jurídicos e técnicos, nos quais a precisão da informação recuperada é fator crítico.

## 4. Conclusão e Trabalhos Futuros

Este trabalho apresentou uma arquitetura baseada em Retrieval-Augmented Generation (RAG) para melhorar a geração de respostas em tarefas de pergunta e resposta a partir de documentos. A proposta integrou a recuperação semântica de informações, por meio de embeddings gerados com a plataforma Ollama, a um modelo de linguagem Llama 3.2 (1B), permitindo que o sistema combinasse busca contextualizada com geração textual em linguagem natural.

A avaliação do sistema foi conduzida com base em perguntas extraídas da obra *A Riqueza das Nações*, de Adam Smith. As respostas geradas com RAG foram comparadas àquelas produzidas por um modelo base, sem acesso ao mecanismo de recuperação. As análises qualitativas evidenciaram que o uso da arquitetura RAG proporcionou respostas mais concisas, precisas e conceitualmente alinhadas ao conteúdo da obra. Essa constatação foi reforçada pelas métricas quantitativas, especialmente o BERTScore, que indicou forte similaridade semântica com as referências.

Os resultados demonstram que a integração entre mecanismos de recuperação semântica e modelos generativos pode elevar substancialmente a qualidade das respostas geradas por LLMs. A arquitetura proposta mostrou-se eficaz na tarefa de alinhar a geração textual ao conteúdo de documentos específicos, o que é particularmente relevante em contextos em que a precisão conceitual é essencial. Tais características tornam essa abordagem especialmente promissora para domínios em que a precisão conceitual e a fidelidade à fonte são essenciais, como educação, saúde, direito e pesquisa científica.

Trabalhos futuros podem explorar diversas direções. Uma possibilidade é ampliar a base documental vetorizada, incorporando múltiplas fontes e organizando os dados em estruturas mais robustas, como grafos de conhecimento. Outra linha de investigação envolve a comparação entre diferentes técnicas de vetorização e recuperação semântica, avaliando seu impacto na performance da arquitetura. Também se destaca o potencial de aplicar a abordagem proposta em domínios específicos, como educação, saúde ou direito,

avaliando sua robustez em tarefas mais especializadas. Por fim, a incorporação de mecanismos de feedback humano para refinar as respostas geradas representa uma perspectiva promissora para aplicações práticas.

## Agradecimentos

Os autores agradecem à Samsung Eletrônica da Amazônia Ltda., por meio do Projeto Aranouá, e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro por meio do Programa de Excelência Acadêmica (PROEX). Este trabalho é resultado do projeto de Pesquisa e Desenvolvimento (P&D) 001/2021, firmado com o Instituto Federal do Amazonas e a FAEPI, com financiamento da Samsung.

## Referências

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. (2023). Geração aumentada por recuperação para grandes modelos de linguagem: Uma revisão. *arXiv preprint arXiv:2312.10997*.

Khan, A. (2023). Knowledge graphs querying. *SIGMOD Rec.*, 52(2):18–29.

Neo4j (2024). Neo4j e microsoft anunciam colaboração em soluções de genai e dados. *Inforchannel*.

Proença, V. L. (2024). Automatização da revisão de literatura científica com geração aumentada por recuperação. *Trabalho de Conclusão de Graduação, Universidade Federal do Rio de Janeiro*.

Rezaei, M. R., Hafezi, M., Satpathy, A., Hodge, L., and Pourjafari, E. (2024). At-rag: An adaptive rag model enhancing query efficiency with topic filtering and iterative reasoning. *arXiv*, 2410.12886.

Vidivelli, S., Ramachandran, M., and Dharunbalaji, A. (2023). Explainability for large language models: A survey. *ArXiv*, abs/2309.01029.

Wang, Y., Gao, L., Liu, C., Liu, J., Xie, Y., Zhang, Z., and Yang, Y. (2024). Enhancing retrieval-augmented generation with self-consistent reasoning. *arXiv*, 2407.19393.

Zhao, H., Luo, T., Yin, X., Tang, S., Wu, F., and Zhuang, Y. (2023). Explainability for large language models: A survey. *ArXiv*, abs/2309.01029.