Tobias – Assistente Acadêmico com *Generative AI* Aprimorada

João José Cunha Melo e Sousa Bentivi[1],
Pollyana Coelh da Silva
Notargiacomo[1]
[1]Faculdade de Computação e Informática (FCI)
Universidade Presbiteriana Mackenzie
São Paulo, SP – Brasil
<10395405@mackenzista.com.br>, <pollyana.notargiacomo@mackenzie.br>

20 de novembro de 2024

Resumo

A pesquisa se propõe a resolver problemas de precisão e confiabilidade enfrentados por modelos de IA generativa, principalmente em contextos acadêmicos. Isso requer a adoção de uma abordagem qualitativa para avaliar o desempenho desses modelos por meio de testes específicos e discutir suas limitações atuais, como alucinações de fatos e erros técnicos. A finalidade descritiva e metodológica busca desenvolver soluções aprimoradas, como a Geração Aumentada por Recuperação (RAG) para otimizar o processamento dos resumos de 222.259 artigos e aumentar a eficácia na interpretação e coleta de informações científicas. A pesquisa é fundamentada em uma revisão bibliográfica e documental abrangente, visando transformar o acesso e análise do conhecimento científico por meio de ferramentas computacionalmente robustas e confiáveis. Como resultado, espera-se obter respostas mais completas e a apresentação das referências para validação com informações sobre o tema da consulta.

Palavras-chave: Inteligência Artificial Generativa, Modelos de Linguagem de Grande Escala (LLMs), Geração Aumentada por Recuperação (RAG), Processamento de Texto Acadêmico, Gestão de Bibliotecas de Conteúdo Científico.

Abstract

The research aims to solve accuracy and reliability problems faced by generative AI models, mainly in academic contexts. Adopting a qualitative approach to evaluate the performance of these models through specific tests and discussing their current limitations, such as fact hallucinations and technical errors. The descriptive and

methodological purpose seeks to develop improved solutions, such as Retrieval Augmented Generation (RAG) to optimize the processing of abstracts of 222,259 articles and increase the effectiveness in the interpretation and collection of scientific information. The research is based on a comprehensive bibliographic and documentary review, aiming to transform the access and analysis of scientific knowledge through robust and reliable tools. As a result, it is expected to obtain more complete responses and the presentation of references for validation with information on the topic of the query.

Keywords: Generative Artificial Intelligence, Large Scale Language Models (LLMs), Retrieval Augmented Generation (RAG), Academic Text Processing, Scientific Content Library Management.

1 Introdução

O cérebro humano possui uma capacidade de reconhecer padrões e interpretar informações sensoriais eficiente, sobretudo se comparada ao que há atualmente na computação. Um exemplo disso diz respeito à, tarefa de identificar palavras escritas à mão. Apesar das variações nos estilos de caligrafia, distorções ou mesmo a falta de letras, é possível ler um exemplo disto na sentença “a úncia csoia ipromtatne é que a piermira e útlmia lretas etjeasm no lguar ctreo”. Essa capacidade decorre da complexa rede de neurônios do cérebro que processa e interpreta as entradas visuais.

Em contraste, programar um computador para executar a mesma tarefa de reconhecimento de palavras apresenta desafios significativos. Escrever algoritmos explícitos para levar em conta todas as variações possíveis na caligrafia é impraticável devido à vasta diversidade de estilos e representações. O problema deixa de ser trivial e adquire uma envergadura que requer, não só, processamento computacional, como a combinação de técnicas da área de inteligência artificial e de mineração de textos.

Assim, para lidar com essas tarefas complexas de reconhecimento de padrões, as redes neurais artificiais (RNAs) surgiram como um paradigma computacional. Inspiradas pela estrutura interconectada dos neurônios biológicos (LV et al., 2024), as RNAs são projetadas para imitar a capacidade do cérebro de aprender com dados e reconhecer padrões por meio de camadas de abstração. As redes neurais se tornaram, inclusive, parte integrante das aplicações modernas de inteligência artificial e aprendizado de máquina, demonstrando sucesso em reconhecimento de imagem e fala, processamento de linguagem natural e vários outros domínios.

O objetivo deste artigo será a comparação qualitativa de grandes modelos de linguagens (LLMs) baseados em redes neurais na resolução de perguntas complexas e sua capacidade de ser validada por referências em suas respostas. O documento abordará como as redes neurais conseguem "aprender", o funcionamento dos modelos de inteligência artificial generativa (Generative Ai) e as técnicas para aprimorar as suas respostas. No documento também será apresentada uma ferramenta chamada “Tobias” que tem como objetivo aprimorar a qualidade das respostas com referências para validação.

2 Conceitos Fundamentais

2.1 Redes Neurais

Uma rede neural consiste em unidades interconectadas chamadas *neurônios*, organizadas em camadas: uma *camada de entrada*, uma ou mais *camadas ocultas* e uma *camada de saída* (ZHAO et al., 2024). Cada neurônio contém um valor numérico conhecido como sua *ativação*, normalmente variando entre 0 e 1.

Já as camadas ocultas processam entradas para extrair características e padrões em vários níveis de abstração. Cada neurônio em uma camada oculta calcula sua ativação com base nas ativações de neurônios da camada anterior (ZHAO et al., 2024). Este cálculo pode envolver vários processos (ROSENBAUM, 2024), como:

1. Calcular uma soma ponderada das entradas.
2. Adicionar um termo de viés com o objetivo de desativar o neurônio caso não ultrapasse um certo valor.
3. Aplicar uma função de ativação para produzir a saída do neurônio.

Matematicamente, a ativação $a_j^{(l)}$ do j-ésimo neurônio na camada l é dada por:

$$a_j^{(l)} = \sigma\left(z_j^{(l)}\right), \tag{1}$$

onde $z_j^{(l)}$ é a entrada ponderada:

$$z_j^{(l)} = \sum_i w_{ji}^{(l)} a_i^{(l-1)} + b_j^{(l)}, \tag{2}$$

e:

- $w_{ji}^{(l)}$ é o peso que conecta o i-ésimo neurônio na camada $(l-1)$ ao j-ésimo neurônio na camada l.
- $b_j^{(l)}$ é o termo de viés para o j-ésimo neurônio na camada l.
- $\sigma(\cdot)$ é a função de ativação.

A função de ativação $\sigma(\cdot)$ introduz não linearidade na rede, permitindo que ela modele relacionamentos complexos. Uma função de ativação comumente usada no início da popularização das redes neurais é a *função sigmoide*:

$$\sigma(z) = \frac{1}{1+e^{-z}}. \tag{3}$$

A função sigmoide mapeia qualquer entrada de valor real para o intervalo $(0, 1)$, que pode ser interpretado como a taxa de disparo do neurônio ou nível de ativação.

No entanto, as redes neurais modernas geralmente empregam funções de ativação alternativas, como a unidade linear retificada (ReLU), definida como:

$$\sigma(z) = \max(0, z). \tag{4}$$

As funções ReLU tendem a ter melhor desempenho em redes profundas ao atenuar o problema do gradiente de desaparecimento e acelerar a convergência durante o treinamento, além de possuir um custo computacional menor (VASWANI et al., 2017).

Complementando os elementos anteriores, a camada de saída produz o resultado final da computação da rede. A ativação $a_k^{(L)}$ do k-ésimo neurônio na camada de saída representa a confiança da rede de que a imagem de entrada corresponde ao dígito k. O dígito previsto $\hat{y}$ é determinado por:

$$\hat{y} = \arg\max \left(a_k^{(L)} \right). \tag{5}$$

Cabe destacar, também, que as computações dentro de uma rede neural podem ser representadas compactamente usando notação matricial. Seja $a^{(l-1)}$ o vetor de ativações da camada $(l-1)$, $W^{(l)}$ a matriz de pesos que conecta as camadas $(l-1)$ e l, e $b^{(l)}$ o vetor de vieses para a camada l. Então, as equações 1 e 2 podem ser escritas como:

$$z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)}, \tag{6}$$

$$a^{(l)} = \sigma \left(z^{(l)} \right). \tag{7}$$

Aqui, $\sigma \left(z^{(l)} \right)$ denota a aplicação elemento a elemento da função de ativação ao vetor $z^{(l)}$.

A capacidade da rede de reconhecer padrões depende dos valores apropriados dos pesos $W^{(l)}$ e vieses $b^{(l)}$. O aprendizado envolve ajustar esses parâmetros para minimizar a discrepância entre as previsões da rede e os rótulos verdadeiros dos dados de treinamento. Isso é obtido por meio de algoritmos como *backpropagation* combinados com *gradient descent* (ZHAO et al., 2024), que atualizam iterativamente os pesos e vieses na direção que reduz o erro.

O erro é quantificado usando uma função de perda, como a perda de entropia cruzada para tarefas de classificação entre duas distribuições de probabilidade. Para um único exemplo de treinamento (x, y), onde x é a imagem de entrada e y é o rótulo verdadeiro, a perda $\mathcal{L}$ é dada por:

$$\mathcal{L} = -\sum_k y_k \log \left(a_k^{(L)} \right). \tag{8}$$

As redes neurais, conforme disposto acima, demonstraram, então, desempenho significantemente superior às demais abordagens computacionais em várias tarefas devido à sua capacidade de modelar relacionamentos complexos e não lineares e aprender representações de recursos hierárquicos.

2.1.1 Aprendizado por meio de gradiente descendente

Inicialmente, os pesos e vieses na rede são definidos como pequenos valores aleatórios. Com parâmetros aleatórios, a rede tem um desempenho ruim na tarefa de classificação, pois

ainda não aprendeu nenhum padrão significativo dos dados. Para melhorar seu desempenho, a rede passa por um processo de aprendizado envolvendo as seguintes etapas:

1. **Propagação direta**: os dados de entrada são passados pela rede para calcular as ativações de saída.

2. **Avaliação da função de custo**: a saída da rede é comparada aos rótulos verdadeiros usando uma função de custo C, como erro quadrático médio (MSE) ou perda de entropia cruzada, que quantifica o erro de previsão da rede.

3. **Propagação reversa (retropropagação)**: os gradientes da função de custo com relação a cada peso e viés são calculados usando a regra da cadeia do cálculo. Este processo calcula eficientemente como as mudanças nos parâmetros afetam o custo.

4. **Otimização de Descida de Gradiente**: Os parâmetros da rede são atualizados na direção que minimiza a função de custo. Para cada peso $w_{ji}^{(l)}$ e viés $b_j^{(l)}$, a regra de atualização é:

$$w_{ji}^{(l)} \leftarrow w_{ji}^{(l)} - \eta \frac{\partial C}{\partial w_{ji}^{(l)}} \tag{9}$$

$$b_j^{(l)} \leftarrow b_j^{(l)} - \eta \frac{\partial C}{\partial b_j^{(l)}}, \tag{10}$$

   onde η é a taxa de aprendizado, um hiperparâmetro que controla o tamanho do passo durante a otimização.

Este processo iterativo permite que a rede ajuste seus pesos e vieses para minimizar a função de custo em todos os exemplos de treinamento, aprendendo efetivamente com os dados.

Ampliando o que foi destacado anteriormente, a função de custo desempenha um papel crucial central na orientação do processo de aprendizagem. Ela mede a discrepância entre as previsões da rede e os rótulos reais em todo o conjunto de dados de treinamento (ZHAO et al., 2024). Uma escolha comum para a função de custo em redes neurais é o custo quadrático total:

$$C = \frac{1}{n} \sum_x \|y(x) - a^{(L)}(x)\|^2, \tag{11}$$

onde:

- n é o número de exemplos de treinamento,
- $y(x)$ é o verdadeiro vetor de rótulo para a entrada x,
- $a^{(L)}(x)$ é o vetor de saída da rede para a entrada x,
- $\|\cdot\|$ denota a norma euclidiana.

Minimizar C garante que as saídas da rede $a^{(L)}(x)$ sejam o mais próximo possível dos rótulos verdadeiros $y(x)$ para todos os exemplos de treinamento.

2.1.2 Desafios no treinamento de redes neurais

Para abordar desafios como **Mínimos locais e pontos de sela**, ***Overfitting*** e a **Complexidade Computacional**, uma das técnicas utilizadas é a *descida de gradiente estocástico* (SGD), onde gradientes são estimados usando subconjuntos aleatórios (mini-lotes) dos dados de treinamento. Isso acelera o treinamento e introduz estocasticidade que pode ajudar a escapar de mínimos locais no cenário de função de custo.

2.2 Transformador pré-treinado generativo (GPT)

Os modelos GPT são projetados especialmente para prever o próximo *token* em uma sequência. Os *tokens* são unidades fundamentais de dados que podem representar palavras, subpalavras ou caracteres. Em aplicativos baseados em texto, os *tokens* normalmente correspondem a palavras ou fragmentos significativos de subpalavras. O modelo processa uma sequência de entrada de *tokens* e gera uma distribuição de probabilidade sobre os próximos *tokens* possíveis, prevendo efetivamente o que vem a seguir na passagem (VASWANI et al., 2017).

O mecanismo de atenção em *Transformers* permite que o modelo pondere a relevância de diferentes partes dos dados de entrada dinamicamente, tornando-o adepto ao processamento de padrões e estruturas complexas inerentes à linguagem e outros dados sequenciais. O termo *Transformador pré-treinado generativo* (GPT) encapsula os principais recursos desses modelos:

- **Generativo**: os modelos GPT são capazes de gerar novos conteúdos, como texto, prevendo *tokens* subsequentes em uma sequência com base em padrões aprendidos de vastos conjuntos de dados.
- **Pré-treinado**: esses modelos passam por um pré-treinamento extensivo em grandes corpora de dados, aprendendo padrões estatísticos, estruturas de linguagem e representações, que podem ser posteriormente ajustados para tarefas específicas.
- **Transformador**: A arquitetura subjacente é baseada no modelo *Transformer*, um tipo de rede neural que utiliza mecanismos de atenção para processar dados de entrada de uma forma que captura relacionamentos entre elementos na sequência, independentemente de suas posições.

A arquitetura *Transformers* permite a paralelização ao processar sequências inteiras simultaneamente. Essa mudança arquitetônica não apenas melhora a eficiência computacional, mas também permite que os modelos capturem dependências de longo alcance dentro do texto, que são cruciais para entender o contexto e a semântica (VASWANI et al., 2017).

2.2.1 Transformer

O mecanismo de autoatenção é central para a arquitetura do *Transformer* e, por extensão, para os modelos GPT. Ele permite que o modelo pondere a influência de diferentes partes dos dados de entrada ao codificar uma palavra ou *token* específico. Ao calcular pontuações de atenção entre *tokens*, o modelo pode capturar relacionamentos contextuais e significados diferenciados (VASWANI et al., 2017).

As principais etapas presentes nos transformadores são (VASWANI et al., 2017):

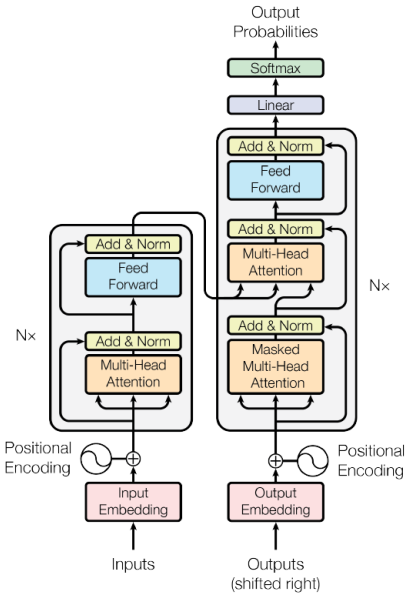

Figura 1 – Transformer (VASWANI et al., 2017, p. 3)

- ***Embedding***: as palavras ou *tokens* do texto de entrada são convertidos em vetores de *embeddings*. Esses *embeddings* são representações numéricas que capturam o significado das palavras em um espaço vetorial contínuo. Como os *Transformers* não têm uma estrutura sequencial interna (como em RNNs), é necessário adicionar informações sobre a posição dos *tokens* na sequência.
- **Camadas de Atenção**: este mecanismo permite que o modelo preste atenção a diferentes partes da sequência de entrada ao processar um determinado *token*. Ele calcula uma pontuação de atenção usando as representações dos *tokens* e distribui a atenção com base nessas pontuações, resultando em uma representação ponderada do contexto para cada *token* (VASWANI et al., 2017).
- **Camadas de Redes *Feed-Forward***: após a camada de atenção, cada *token* passa por uma rede neural *feed-forward* idêntica, mas aplicada separadamente a cada posição. Isso envolve transformações lineares e funções de ativação não-lineares, como ReLU (VASWANI et al., 2017).
- **Camadas de Normalização e Resíduos**: Para facilitar o treinamento, o Transformador utiliza conexões residuais em torno de cada subcamada (atenção e *feed-forward*) (VASWANI et al., 2017). Após as conexões residuais, é aplicada a normalização da camada para estabilizar e acelerar o treinamento.
- **Camada de saída**: Após várias repetições das camadas de atenção e *feed-forward*, a saída final é projetada de volta para o espaço original dos *tokens*, usando uma camada linear, seguida de uma *softmax* no caso de modelos de linguagem, para gerar a distribuição de probabilidade sobre o vocabulário.

No contexto dos *Transformers* (Figura 1), a etapa de *embedding* transforma palavras ou *tokens* em vetores numéricos que podem ser processados pelo modelo. Essa etapa é

dividida em duas partes principais: *input embedding* e *positional encoding* (VASWANI et al., 2017).

2.2.1.1 Criação do *Embedding*

A elaboração do *Embedding* envolve dois aspectos: *Input Embedding* e *Positional Encoding*. O propósito do *input embedding* é converter palavras ou *tokens* em vetores densos de tamanho fixo, permitindo que o modelo represente semanticamente cada termo em um espaço vetorial contínuo. Para cada *token* no vocabulário, um vetor de dimensão fixa é atribuído e aprendido durante o treinamento do modelo (VASWANI et al., 2017).

Já no *Positional Encoding*, ao contrário de modelos sequenciais como RNNs, os *Transformers* não têm um mecanismo intrínseco para lidar com a ordenação dos *tokens*. Para resolver isso, é necessário adicionar um vetor de *positional encoding* a cada *input embedding* para introduzir informações de posição. Os *positional encodings* são definidos, portanto, usando funções senoidais e cosenoidais de diferentes frequências, permitindo que o modelo capture a posição relativa dos *tokens* (VASWANI et al., 2017). Formalmente, para uma posição *pos* e uma dimensão do vetor $2i$, são definidos como:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d}}\right)$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d}}\right)$$

onde d é a dimensionalidade do vetor de *embedding*.

Essa combinação de *input embedding* e *positional encoding* permite que o *Transformer* represente e processe dados de texto levando em consideração tanto o significado das palavras, quanto a sua posição nas sequências de entrada. A matriz de *embedding*, portanto, é criada e ajustada durante o treinamento de um modelo *Transformer*. Este processo envolve a inicialização de pesos, o cálculo do erro durante o treinamento, e a atualização dos pesos usando otimização baseada em gradiente. Para a criação da matriz de *Embedding*, os seguintes passos são realizados.

- **Inicialização Aleatória dos Pesos**:

  A matriz de *embedding* $\mathbf{E}$ é criada com dimensões $V \times d$, onde V é o tamanho do vocabulário e d é a dimensão do *embedding*. Cada linha da matriz corresponde ao vetor de *embedding* de um *token* do vocabulário. Inicialmente, os elementos desta matriz são preenchidos com valores aleatórios pequenos. Isso pode ser feito usando uma distribuição normal ou uniforme. A inicialização aleatória é importante para quebrar a simetria entre os neurônios.

  $$\mathbf{E}_{ij} \sim \mathcal{U}(-\epsilon, \epsilon)$$

  onde ϵ é um pequeno valor escolhido para evitar valores extremos.

- **Passo de *Forward* e Cálculo de Erro**: Durante o treinamento, os *embeddings* são usados na rede para prever saídas (por exemplo, a próxima palavra em uma sequência). A previsão gerada pelo modelo é comparada à verdade do solo (*ground truth*), e uma função de perda $\mathcal{L}(\hat{y}, y)$ é calculada, onde $\hat{y}$ é a previsão e y é o valor real.

- **Cálculo do Gradiente**: O gradiente da função de perda em relação à matriz de *embedding* $\frac{\partial \mathcal{L}}{\partial \mathbf{E}}$ é calculado. Este gradiente indica a direção e magnitude pela qual cada peso da matriz de *embedding* deve ser ajustado para reduzir o erro.

- **Atualização de Pesos usando Otimização por Gradiente**: O algoritmo de otimização (por exemplo, Adam, SGD) é usado para ajustar os pesos da matriz de *embedding* com base no gradiente calculado. Um passo de atualização típico para cada elemento $\mathbf{E}_{ij}$ da matriz é dado por:

$$\mathbf{E}_{ij} = \mathbf{E}_{ij} - \eta \cdot \frac{\partial \mathcal{L}}{\partial \mathbf{E}_{ij}}$$

  onde η é a taxa de aprendizado, controlando o tamanho do passo em direção à minimização da perda.

- **Iteração e Convergência**: é realizado o processo de propagação reversa, onde é realizado o cálculo de gradiente e atualização de pesos se repete para muitas iterações (ou *epochs*) sobre o conjunto de dados de treinamento, aproximando vetores de *embeddings* de palavras com significados semelhantes. Com o tempo, a matriz de *embedding* aprende distribuições que refletem as semelhanças semânticas entre palavras. Por exemplo, palavras relacionadas como *cachorro* e *animal* tenham suas representações vetoriais mais próximas no espaço vetorial (ZHAO et al., 2024).

Este processo iterativo e adaptativo é o que permite que os *embeddings* capturem e representem a semântica do texto de maneira eficiente durante o treinamento do modelo. Neste sentido, a partir do foco deste trabalho – uso de métodos de aprimoramento de respostas do modelo voltado para a pesquisa acadêmica –, a seguir é caracterizada a arquitetura de recuperação de informações.

2.3 Retrieval-Augmented Generation (RAG)

O *Retrieval-Augmented Generation* (RAG) é uma arquitetura que combina modelos de recuperação de informações utilizando *embeddings* com modelos generativos para produzir respostas mais precisas e contextualizadas. A ideia central do RAG é "enriquecer" o processo de geração de texto com informações relevantes recuperadas de uma base de dados externa, permitindo que o modelo forneça respostas atualizadas e específicas, mesmo sobre tópicos que não foram amplamente cobertos durante seu treinamento inicial (FINARDI et al., 2024).

Assim, dada uma base de informações, o seu conteúdo deve ser convertido para uma estrutura matemática contendo seu valor semântico. Este processo é realizado ao separar o conteúdo textual em pedaços (*Chunks*) e os convertendo em *embeddings*, pois o tamanho do conteúdo não deve superar o do *embedding* utilizado, organizando essas informações em um vetor de conteúdos (*VectorStore*).

Desta forma, no momento que o usuário realiza uma requisição, esta é convertida em um *embedding*. A partir do vetor do *embedding*, é computada a similaridade cosseno entre o vetor da pesquisa e cada vetor de *embedding* armazenado na *VectorStore*. A distância cosseno é calculada utilizando a seguinte fórmula:

$$\text{distância cosseno} = 1 - \cos(\theta) = 1 - \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

Onde: - $\mathbf{A} \cdot \mathbf{B}$ é o produto escalar dos vetores $\mathbf{A}$ e $\mathbf{B}$. - $\|\mathbf{A}\|$ e $\|\mathbf{B}\|$ são as normas (ou magnitudes) dos vetores $\mathbf{A}$ e $\mathbf{B}$, respectivamente, calculadas como $\|\mathbf{A}\| = \sqrt{\sum_i A_i^2}$. Essa medida é especialmente útil quando os dados são vetoriais e o interesse principal está na direção dos vetores, ignorando suas magnitudes absolutas.

A partir disto, os documentos são, então, ordenados com base nos valores de similaridade, e os top-k documentos mais relevantes são selecionados. Com a pesquisa e coleta de informações no *VectorStore* concluída, as top-k pesquisas são adicionadas como contexto para a formulação da resposta final por meio de um modelo generativo.

Cabe destacar que, ao integrar informações recentes presentes nos dados, o RAG supera a limitação de modelos de linguagem que possuem um corte temporal de treinamento, fornecendo respostas baseadas nas informações mais atuais disponíveis. A recuperação de documentos também permite que o modelo acesse detalhes específicos que podem não estar presentes em seus parâmetros internos, enriquecendo as respostas com fatos e dados precisos (SAWARKAR; MANGAL; SOLANKI, 2024), possibilitando fornecer as fontes das informações usadas na geração da resposta e aumentando a confiança do usuário ao permitir a verificação dos dados.

Além disto, modelos generativos puros podem apresentar um fenômeno de "alucinar"ou gerar geração de informações inexatas. O RAG reduz esse problema ao ancorar as respostas em documentos reais ricos em conteúdo factual relevante.

Contudo, é importante ressaltar que, caso os documentos recuperados não forem relevantes ou contiverem informações incorretas, a qualidade da resposta será afetada. Outra dificuldade está em manter o conteúdo dos documentos atualizado e relevante necessita de esforços contínuos. Podendo até ser necessário lidar com grandes volumes de dados, garantindo sua qualidade (BRUCKHAUS, 2024). Como a consulta no *VectorStore* é feita a partir de um vetor gerado pela entrada do usuário, o modelo também é suscetível a pesquisas com baixa informação semântica. É visto na literatura apresenta casos em que RAG aprimorou as consultas (CUCONASU et al., 2024) e casos que não (BRUCKHAUS, 2024).

3 Metodologia

A presente investigação constitui uma pesquisa aplicada, de cunho explicativo, cuja abordagem é *ground theory* (GIL, 2022). Para a implementação da proposta foram utilizados os modelos **GPT-4o** e **o1-preview**. Estes foram escolhidos por apresentarem os melhores desempenhos no mercado. Isto pode ser observado nas Figuras 2 e 3.

Com base nas definições estabelecidas, a etapa dois envolveu a seleção da base de dados para a criação do *VectorStores*. Foi selecionada a biblioteca de artigos científicos ArXiv (ARXIV*, 2024), plataforma fundada em 1992 por Paul Ginsparg e mantida pela Universidade de Cornell. O objetivo da escolha foi a criação de uma base de dados com informações que não estão presentes no treino dos modelos atuais com documentos atualizado e relevantes por contar com uma curadoria acadêmica. Assim, a base de dados da presente investigação foi criada a partir dos *asbtracts* dos artigos publicados em 2024, totalizando 222.259 publicações armazenadas em uma *VectorStore* da biblioteca FAISS (*Facebook AI Similarity Search*), ocupando 2,84 GB em disco.

A partir dessa base de artigos confiável e ampla, o próximo passo para aprimorar a pesquisa (etapa 3) foi uma implementado de uma ampliação do contexto semântico

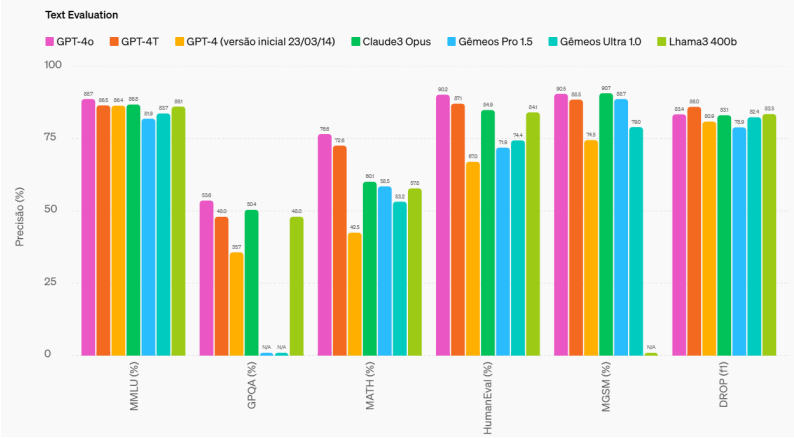

Figura 2 – Modelos (OPENAI*, 2024a)

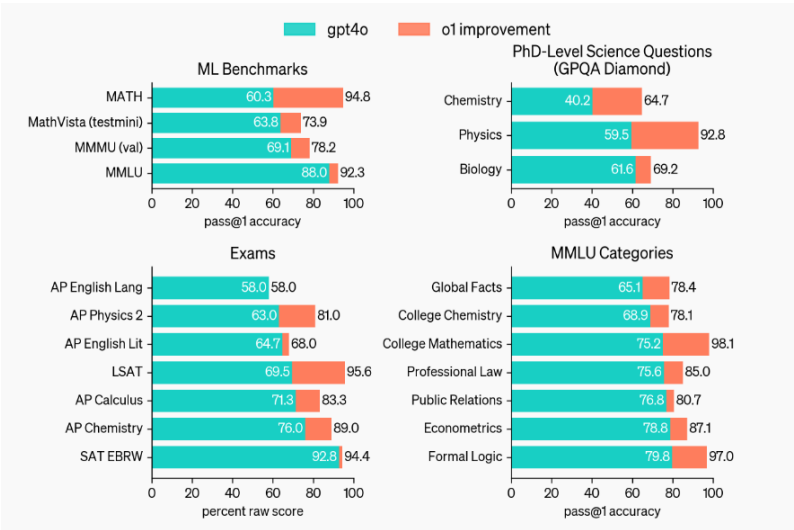

Figura 3 – Comparação dos modelos **GPT-4o** e **o1-preview** (OPENAI*, 2024b)

antes da busca de artigos no *VectorStore*. Isto foi viabilizado por meio de um *prompt* que é instruído a melhorar a pergunta do usuário, criando três diferentes consultas a partir da entrada inicial. Após feita a busca, como resultado é criada uma resposta provisória contendo um valor semântico ampliado. Esta resposta provisória é adicionada à pergunta inicial, permitindo que a consulta de dados relevantes no *VectorStore* seja refeita.

Com os dados de consultas relevantes e a pergunta inicial ampliada, se passou à etapa 4, em que se faz uso do modelo *o1-preview* para gerar a resposta final para o usuário. A ilustração do processo pode ser observada na figura 4.

Por último, foram definidos os modelos para a realização da comparação dos resultados. Foram selecionados os modelos **Meta AI**, **Gemini 1.5 flash** (estes dois possuem acesso à interente em suas consultas), **GPT-4o** e **o1-preview**.

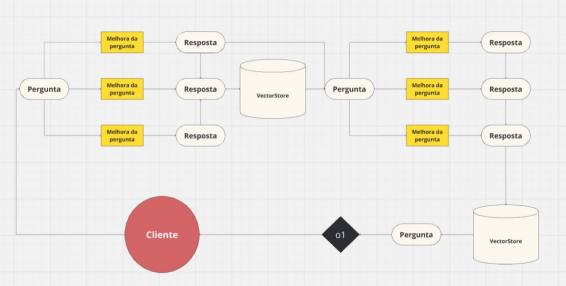

Figura 4 – Representação da metodologia de consulta (Elaborada pelo autor).

4 Resultados e discussão

Para o teste, foi realizada a pergunta: "Explique o paradoxo EPR e discuta suas implicações para o princípio de localidade na teoria quântica. Como experimentos recentes ajudam a esclarecer ou complicar o paradoxo?". Essa escolha foi feita por se tratar de um tema ainda considerado complexo na academia e por ser um conteúdo ainda pesquisado e premiado (Nobel de 2022).

Os resultados de cada modelo estão descritos abaixo (destacados em azul).

4.1 Meta AI

4.1.1 Resultados do Teste realizado com o Modelo Meta AI

O paradoxo EPR (Einstein-Podolsky-Rosen) é um experimento Gedanken que desafia o princípio de localidade na teoria quântica. Em 1935, Albert Einstein, Boris Podolsky e Nathan Rosen propuseram um cenário para demonstrar a suposta incompletude da mecânica quântica.

O Paradoxo EPR:

1. Dois sistemas quânticos, A e B, estão correlacionados de tal forma que suas propriedades estão entrelaçadas.

2. Medir uma propriedade em A instantaneamente afeta o estado de B, independentemente da distância entre eles.

Desafio ao Princípio de Localidade:

A teoria quântica sugere que a informação pode ser transmitida instantaneamente entre sistemas entrelaçados, violando o princípio de localidade.

Implicações:

1. Não-localidade: A teoria quântica permite a comunicação instantânea entre sistemas distantes.

2. Incompletude: O paradoxo EPR sugere que a mecânica quântica é incompleta.

Experimentos Recentes:

1. Experimento de Aspect (1982): Confirmou a violação da desigualdade de Bell.

2. Experimento de entrelaçamento quântico (1997): Demonstra a conexão entre partículas entrelaçadas.

3. Experimento de teleportação quântica (2006): Transfere informações quânticas entre sistemas.

Conclusão:

O paradoxo EPR desafia o princípio de localidade, e experimentos recentes reforçam a não-localidade da teoria quântica. Isso levanta questões fundamentais sobre a natureza da realidade e a compatibilidade entre relatividade e mecânica quântica.

Leitura Adicional:

- Einstein, A., Podolsky, B., Rosen, N. (1935). Can quantum-mechanical description of physical reality be considered complete? Physical Review, 47(10), 777-780.

- Bell, J. S. (1964). On the Einstein Podolsky Rosen paradox. Physics, 1, 195-200.

- Aspect, A. (1982). Bell's theorem: The naive view. Foundations of Physics, 12(10), 137-141.

A explicação do paradoxo EPR está muito superficial, com um erro na afirmação "A teoria quântica permite a comunicação instantânea entre sistemas distantes". Não há comunicação instantânea, pois a casualidade relativística ainda é preservada, mesmo sendo não-local (BENTIVI, 2018). Vemos que o modelo apresenta referências corretas, mas não conseguiu obter informações recentes. Nas menções dos experimentos realizados, o mais recente data de 2006.

4.2 Gemini 1.5 flash

4.2.1 Resultados do Teste realizado com o Modelo Gemini 1.5 flash

O Paradoxo EPR e suas Implicações para a Localidade Quântica

O paradoxo EPR, nomeado após Einstein, Podolsky e Rosen, é um experimento mental que explora as implicações bizarras da mecânica quântica. Ele desafia uma das nossas intuições mais profundas sobre o universo: a localidade.

O que é o Paradoxo EPR?

Imagine duas partículas entrelaçadas, como se fossem gêmeas quânticas. Elas são preparadas de forma que o estado de uma está instantaneamente correlacionado com o estado da outra, independentemente da distância que as separa. Se medirmos uma propriedade de uma partícula, instantaneamente sabemos o valor da mesma propriedade da outra partícula, não importa quão longe ela esteja.

Por que é um paradoxo?

Esse fenômeno parece violar o princípio da localidade, que afirma que um objeto só pode ser influenciado por seu entorno imediato. A ideia de que uma ação em um lugar possa ter um efeito instantâneo em outro lugar, sem nenhuma interação física entre eles, desafia nossa compreensão clássica de causa e efeito.

Implicações para a Localidade:

O paradoxo EPR nos força a reconsiderar o conceito de localidade na teoria quântica. Ele sugere que a realidade quântica pode ser não local, ou seja, as partículas entrelaçadas podem estar conectadas de uma maneira que transcende o espaço e o tempo.

Experimentos Recentes e suas Implicações:

Nos últimos anos, diversos experimentos foram realizados para testar as previsões do paradoxo EPR. Esses experimentos confirmaram repetidamente a existência do entrelaçamento quântico e sua natureza não local.

Testes de Bell: Esses experimentos, inspirados nas desigualdades de Bell, foram cruciais para descartar explicações alternativas ao entrelaçamento, como variáveis ocultas locais. Os resultados experimentais sempre violaram as desigualdades de Bell, confirmando a natureza não local do entrelaçamento.

Teletransporte Quântico: Essa técnica permite transferir um estado quântico de uma partícula para outra, mesmo que estejam separadas por grandes distâncias. O teletransporte quântico se baseia no entrelaçamento e demonstra a capacidade de transferir informação de forma instantânea.

Redes de Entrelaçamento: Cientistas têm construído redes cada vez maiores de partículas entrelaçadas, demonstrando a possibilidade de realizar cálculos quânticos distribuídos e comunicação quântica segura.

O que esses experimentos significam?

Os experimentos recentes reforçam a ideia de que a mecânica quântica é uma teoria fundamental da natureza e que o entrelaçamento quântico é um fenômeno real. Eles desafiam nossa intuição clássica sobre o mundo e abrem novas possibilidades para a computação quântica, a criptografia e a nossa compreensão fundamental da realidade.

Em resumo:

O paradoxo EPR é um dos aspectos mais intrigantes da mecânica quântica. Ele nos mostra que a realidade quântica é muito mais estranha e complexa do que podemos imaginar. Embora a não localidade seja um conceito desafiador, os experimentos recentes fornecem evidências sólidas de sua existência e abrem novas perspectivas para a ciência e a tecnologia.

Para aprofundar seus conhecimentos, você pode pesquisar sobre:

Desigualdades de Bell: Um conjunto de desigualdades matemáticas que delimitam o quanto as correlações entre sistemas físicos podem ser explicadas por teorias locais e realistas.

Teoria da Informação Quântica: Um campo que estuda a transmissão, processamento e proteção de

informação quântica.

Computação Quântica: Um paradigma de computação baseado nos princípios da mecânica quântica, com potencial para resolver problemas que são intratáveis para computadores clássicos.

Referências:

<https://www3.unicentro.br/petfisica/2018/08/28/paradoxo-de-einstein-podolsky-rosen-aplicado-em-um-sistema-de-centenas-de-#:~:text=O%20paradoxo%20de%20Einstein%2DPodolsky,trata%20sobre%20a%20natureza%20da>

<https://www.gta.ufrj.br/grad/07_1/quantica/ParadoxoEPR.html#:~:text=O%20paradoxo%20EPR%20refuta%20o,tem%20um%20efeito%20n%C3%A3o%20local>

<https://www.fisica.net/conceitos/2009/06/paradoxo-epr.html>

<https://pt.wikipedia.org/wiki/Experi%C3%AAncias_dos_testes_de_Bell>

<https://www.inovacaotecnologica.com.br/noticias/noticia.php?artigo=teletransporte-envolvendo-materia&id=010110200623#:~:text=O%20teletransporte%20qu%C3%A2ntico%20envolve%20duas,teleporta%22%20instantaneamente%20seu%20estado%20para>

<https://www.maisconhecer.com/tecnologia/9832/Provando-que-o-entrelacamento-quantico-e-real>

<https://m.youtube.com/watch?v=IGEqhWZwjBQ>

<https://pt.wikipedia.org/wiki/Ci%C3%AAncia_da_informa%C3%A7%C3%A3o_qu%C3%A2ntica#:~:text=A%20ci%C3%AAncia%20da%20informa%C3%A7%C3%A3o%20qu%C3%A2ntica,de%20informa%C3%A7%C3%B5es%20usando%20princ%C3%ADpios%20da>

<https://aws.amazon.com/pt/what-is/quantum-computing/#:~:text=Computa%C3%A7%C3%A3o%20qu%C3%A2ntica%20%C3%A9%20um%20campo,que%20utiliza%20a%20mec%C3%A2nica%20qu%C3%A2ntica>

O modelo possui uma explicação razoável acerca do paradoxo EPR, contudo suas referências são providas de páginas da internet com informações incompletas de experimentos realizados, com a mais recente sendo de 2020 (Teletransporte quântico). Devido a este conteúdo, é criado um tom de escrita semelhante ao de blogs e revistas online. Também vemos um erro na afirmação “demonstra a capacidade de transferir informação de forma instantânea.” pois, como mencionado anteriormente, não há transferência de informação instantânea, pois a casualidade relativística ainda é preservada, mesmo sendo não-local (BENTIVI, 2018).

4.3 GPT-4o

4.3.1 Resultados do Teste realizado com o Modelo GPT-4o

O paradoxo EPR (Einstein-Podolsky-Rosen) foi formulado em 1935 por Albert Einstein, Boris Podolsky e Nathan Rosen. Eles questionavam a completude da mecânica quântica e introduziram o paradoxo para sugerir que a teoria quântica talvez fosse inadequada para descrever a realidade física por completo. A base do paradoxo está na propriedade do entrelaçamento quântico (ou emaranhamento), onde partículas podem ter seus estados correlacionados de tal forma que a medição do estado de uma imediatamente determina o estado da outra, independentemente da distância entre elas.

Estrutura do Paradoxo EPR O paradoxo EPR propõe o seguinte experimento de pensamento:

Considere duas partículas entrelaçadas, A e B, criadas juntas e separadas por uma grande distância. De acordo com a mecânica quântica, o estado das partículas é descrito por uma função de onda coletiva, e as propriedades individuais, como momento e posição, só são determinadas no ato da medição. Se medimos uma propriedade (como o momento) de uma das partículas (A), instantaneamente sabemos o valor correspondente da propriedade na outra partícula (B) sem precisar medi-la. Essa correlação instantânea parece contradizer o princípio da localidade, que afirma que objetos distantes não podem influenciar-se mutuamente instantaneamente, respeitando o limite da velocidade da luz imposto pela relatividade especial. Einstein interpretava essa não-localidade aparente como um indício de que a mecânica quântica não era uma descrição completa da realidade, sugerindo que haveria "variáveis ocultas"determinando os resultados das medições de forma independente da função de onda quântica.

Implicações para o Princípio de Localidade O paradoxo EPR questiona a compatibilidade entre a mecânica quântica e o princípio da localidade. Se a medição em uma partícula afeta instantaneamente o estado da outra, isso sugeriria uma violação da localidade. Essa questão gerou intensos debates sobre se a mecânica quântica implica uma forma de "não-localidade"fundamental ou se há uma interpretação alternativa da teoria (como a teoria das variáveis ocultas) que preservaria a localidade.

Experimentações Recentes e Desigualdades de Bell Na década de 1960, o físico John Bell formulou as

desigualdades de Bell, uma série de desigualdades matemáticas que qualquer teoria de variáveis ocultas locais deveria satisfazer. Essas desigualdades permitiram uma maneira experimental de testar o paradoxo EPR: se as desigualdades de Bell fossem violadas, então qualquer teoria baseada em variáveis ocultas locais estaria incorreta.

Experimentos realizados desde os anos 1970, especialmente os mais recentes usando técnicas avançadas e "loophole-free"(sem brechas) envolvendo entrelaçamento de partículas e testes rigorosos de Bell, confirmaram repetidamente a violação das desigualdades de Bell. Esses experimentos mostram que a mecânica quântica permite correlações entre partículas que violam o princípio da localidade de uma maneira incompatível com qualquer teoria de variáveis ocultas locais. Contudo, não significam necessariamente transmissão de informação a uma velocidade superior à da luz, pois não é possível usar o entrelaçamento para enviar informações sem uma coordenação clássica.

Implicações e Interpretações Atuais Esses resultados reforçam a visão de que o universo quântico possui uma não-localidade inerente que desafia a intuição clássica. A interpretação dessa não-localidade ainda é tema de debate. As duas abordagens principais são:

Interpretação de Copenhagen: Aceita que a mecânica quântica é completa e que a não-localidade é uma característica fundamental da natureza, mas evita discutir o "realismo"dos estados entrelaçados antes da medição. Interpretação de Muitos Mundos: Sugere que todas as possibilidades quânticas ocorrem em universos paralelos, eliminando a necessidade de "colapso"instantâneo e a violação de localidade, embora implique um multiverso. Os experimentos recentes complicam o paradoxo EPR ao mostrar que, embora a mecânica quântica pareça não-local, ainda não é possível usá-la para violar a causalidade no sentido clássico, uma vez que o entrelaçamento não permite a comunicação mais rápida que a luz. No entanto, esses resultados desafiam nossa compreensão de realidade e causam novas reflexões sobre o significado profundo da teoria quântica.

Apresenta uma explicação razoável acerca do paradoxo EPR, contudo não apresenta referências e não possui informações de experimentos recentes a cerca do tema.

4.4 o1-preview

4.4.1 Resultados do Teste realizado com o Modelo 01-preview

O Paradoxo EPR (Einstein-Podolsky-Rosen) é um argumento desenvolvido em 1935 pelos físicos Albert Einstein, Boris Podolsky e Nathan Rosen. Eles buscaram demonstrar que a mecânica quântica, conforme entendida na época, era uma teoria incompleta. O paradoxo explora as implicações do entrelaçamento quântico e questiona o princípio de localidade na teoria quântica.

Descrição do Paradoxo EPR Einstein, Podolsky e Rosen consideraram um par de partículas que interagem e depois se separam, mas permanecem em um estado quântico entrelaçado. Nesse estado, as propriedades físicas (como posição e momento, ou spin) das partículas estão correlacionadas de tal forma que a medição em uma partícula imediatamente determina o estado da outra, independentemente da distância que as separa.

Eles argumentaram o seguinte:

Realismo Local: A ideia de que as propriedades físicas existem independentemente da observação e que nenhuma influência pode se propagar mais rápido que a luz (princípio de localidade da relatividade especial).

Previsibilidade Perfeita: Com base na função de onda quântica, é possível prever com certeza o resultado de uma medição em uma partícula ao medir sua parceira entrelaçada.

Dessa forma, EPR concluíram que:

Se é possível prever o valor de uma propriedade física sem interagir diretamente com a partícula, então essa propriedade deve ter um valor real antes da medição. Portanto, a mecânica quântica não fornece uma descrição completa da realidade física, sugerindo a existência de variáveis ocultas que determinam os resultados das medições. Implicações para o Princípio de Localidade O paradoxo EPR desafia o princípio de localidade, que afirma que eventos ocorrendo em pontos separados no espaço não podem influenciar-se instantaneamente. Na interpretação tradicional da mecânica quântica (interpretação de Copenhague), a medição do estado de uma partícula entrelaçada parece afetar instantaneamente o estado da outra, mesmo que separadas por distâncias arbitrariamente grandes. Isso sugere um tipo de não-localidade ou "ação fantasmagórica à distância", como Einstein descreveu.

Essa aparente violação da localidade levanta questões fundamentais sobre a compatibilidade entre a mecânica quântica e a relatividade especial. Se a mecânica quântica permitir influências instantâneas, isso contraria o limite de velocidade universal imposto pela relatividade (a velocidade da luz).

O Papel do Teorema de Bell Em 1964, o físico John Bell formulou o Teorema de Bell, que propõe uma maneira de testar experimentalmente se a natureza obedece ao realismo local ou se é inerentemente não-local, conforme previsto pela mecânica quântica.

Desigualdades de Bell: Bell derivou desigualdades matemáticas que devem ser obedecidas por qualquer teoria de variáveis ocultas locais.

Predições Quânticas: A mecânica quântica prevê violações dessas desigualdades em certos experimentos com partículas entrelaçadas.

Experimentos Recentes e Suas Implicações Desde a proposição do Teorema de Bell, diversos experimentos foram realizados para testar as desigualdades de Bell. Os resultados têm consistentemente apoiado as predições da mecânica quântica, indicando violações das desigualdades de Bell e, portanto, sugerindo que a natureza não pode ser descrita por teorias de variáveis ocultas locais.

Principais Experimentos: Experimentos de Aspect (1982): Alain Aspect e colegas realizaram experimentos que demonstraram violações das desigualdades de Bell, fortalecendo as evidências contra o realismo local.

Experimentos Sem Lacunas (2015):

B. Hensen et al. (Delft University): Realizaram o primeiro teste de Bell sem lacunas significativas, fechando tanto a "lacuna de detecção"quanto a "lacuna de localidade". Utilizaram qubits em diamantes separados por 1,3 km. L. K. Shalm et al. e M. Giustina et al.: Conduziram experimentos independentes usando fótons e também fecharam as principais lacunas, reforçando os resultados anteriores. Impacto dos Experimentos: Refutação de Variáveis Ocultas Locais: Os resultados experimentais indicam que não é possível explicar as correlações quânticas com base em teorias que mantenham simultaneamente o realismo e a localidade.

Aceitação da Não-Localidade Quântica: A mecânica quântica parece permitir correlações não-locais, onde a medição em uma partícula influencia instantaneamente o estado da outra, sem transferência de sinal ou energia.

Causalidade Preservada: Apesar da não-localidade, a mecânica quântica não permite comunicação mais rápida que a luz. As correlações entre partículas não podem ser usadas para transmitir informação de forma superluminal, preservando a causalidade relativística.

Esclarecimentos e Complicações do Paradoxo Esclarecimentos: Natureza das Correlações Quânticas: Os experimentos ajudaram a clarificar que as correlações quânticas são mais fortes do que qualquer correlação clássica permitida por teorias de variáveis ocultas locais.

Interpretação da Mecânica Quântica: Os resultados estimulam debates sobre as diferentes interpretações da mecânica quântica, como a interpretação de Copenhague, teoria de ondas piloto (Bohmiana) ou a interpretação dos muitos mundos.

Complicações: Desafios Conceituais: A aceitação da não-localidade quântica obriga os físicos a repensarem conceitos fundamentais sobre a natureza da realidade, separabilidade e independência espacial.

Questões Abertas: Apesar dos avanços experimentais, questões sobre o que exatamente significa "realidade"quântica e como conciliar mecânica quântica com relatividade geral permanecem em aberto.

Conclusão O paradoxo EPR e os experimentos subsequentes desempenham um papel crucial na compreensão dos fundamentos da mecânica quântica. Eles mostram que:

A Mecânica Quântica é Inerentemente Não-Local: As predições quânticas e os resultados experimentais indicam que a natureza exibe não-localidade quântica, o que desafia o princípio de localidade clássico.

Abandono do Realismo Local: Devemos abandonar a noção de que as propriedades físicas têm valores definidos independentes da observação (realismo) ou aceitar a não-localidade.

Preservação da Causalidade: Mesmo com a não-localidade, a mecânica quântica não viola a causalidade relativística, pois não permite comunicação superluminal.

Os experimentos recentes esclarecem o paradoxo EPR ao fornecer evidências empíricas que apoiam a mecânica quântica e suas implicações não-locais. No entanto, eles também complicam nossa compreensão intuitiva do mundo, desafiando conceitos arraigados sobre espaço, tempo e realidade. A exploração contínua desses fenômenos é fundamental para o desenvolvimento de teorias mais profundas que unifiquem a mecânica quântica e a gravidade.

Apresenta uma explicação mais profunda do paradoxo e apresenta referências precisas. Contudo, não foi apresentado nenhuma referência mais recente que 2015.

4.5 Tobias

4.5.1 Resultados do Teste realizado com o Tobias

O paradoxo EPR, proposto por Einstein, Podolsky e Rosen em 1935, é uma argumentação que questiona a completude da mecânica quântica e a compatibilidade desta com o princípio de localidade. O paradoxo considera dois sistemas quânticos que interagiram e, posteriormente, se separaram por uma grande distância. De acordo com a mecânica quântica, esses sistemas podem estar em um estado emaranhado, no qual as propriedades de um estão intrinsecamente ligadas às do outro, independentemente da separação espacial entre eles.

Implicações para o Princípio de Localidade na Teoria Quântica:

O princípio de localidade, fundamentado na teoria da relatividade de Einstein, estabelece que um objeto

é influenciado diretamente apenas pelo seu entorno imediato e que nenhuma influência pode se propagar mais rapidamente que a velocidade da luz. No contexto do paradoxo EPR, a medição realizada em um dos sistemas aparentemente influencia instantaneamente o estado do outro sistema distante. Isso sugere uma violação do princípio de localidade, pois implica em uma conexão instantânea entre as partículas, independentemente da distância que as separa.

Essa aparente violação levou ao debate sobre a natureza da realidade na mecânica quântica e se esta teoria seria completa ou deveria ser complementada por "variáveis escondidas"que restabeleceriam o realismo local. John Bell, em 1964, formulou as desigualdades que levam seu nome, demonstrando que nenhuma teoria de variáveis escondidas locais pode reproduzir todas as previsões da mecânica quântica.

Contribuições dos Experimentos Recentes:

Experimentos recentes têm sido cruciais para testar as previsões da mecânica quântica e esclarecer o paradoxo EPR. A violação das desigualdades de Bell em experimentos de laboratório fornece evidências empíricas de que a natureza não obedece ao realismo local.

Por exemplo, o artigo "*Violation of Bell's Inequality in the Clauser-Horne-Shimony-Holt Form with Entangled Quantum States Revisited*"de Carlo Cafaro et al. (2024) revisita a violação da desigualdade de Bell na forma Clauser-Horne-Shimony-Holt (CHSH) usando estados quânticos emaranhados. Os autores discutem como a extensão dessa violação depende tanto da orientação dos polarizadores quanto do grau de emaranhamento dos estados quânticos. Eles também destacam a importância dos experimentos realizados por Aspect, Grangier e Roger em 1982, que forneceram evidências robustas da violação das desigualdades de Bell, fortalecendo a compreensão da não-localidade quântica.

Outro exemplo significativo é o trabalho "*Filtering one-way Einstein-Podolsky-Rosen steering*"de Ze-Yan Hao et al. (2024). Neste estudo, os autores investigam o efeito da reutilização de partículas descartadas em operações de filtragem locais. Eles observam todas as configurações de condução EPR simultaneamente e evoluções intrigantes da não-localidade quântica assimétrica, como a inversão da direção da condução EPR unidirecional. Este trabalho fornece uma nova perspectiva sobre a utilização da condução quântica como um recurso e demonstra ferramentas práticas para manipular sistemas quânticos assimétricos com potencial para aplicações em tarefas de informação quântica.

Esses experimentos avançam nossa compreensão sobre o paradoxo EPR ao demonstrar que os efeitos quânticos não podem ser explicados por teorias locais realistas. No entanto, eles também complicam o paradoxo, pois reforçam que a mecânica quântica desafia nossas intuições clássicas sobre separabilidade e causalidade. A não-localidade quântica não permite comunicação superluminal, mas indica que as partículas emaranhadas compartilham correlações que não podem ser explicadas por interações locais convencionais.

Em resumo, o paradoxo EPR e os experimentos subsequentes revelam que o princípio de localidade não é absoluto no nível quântico. A mecânica quântica, confirmada por experimentos como os discutidos, exige uma reavaliação dos conceitos de localidade e realidade, indicando que, no domínio quântico, as partes de um sistema podem estar profundamente conectadas de maneiras que não se encaixam na compreensão clássica. A pesquisa contínua nessa área busca não apenas esclarecer esses fundamentos teóricos, mas também explorar suas implicações para o desenvolvimento de tecnologias quânticas emergentes.

Referências:

[Distância semântica $L2^2$= 0.831] Titulo: Violation of Bell's Inequality in the Clauser-Horne-Shimony-Holt Form with Entangled Quantum States Revisited Autores: Carlo Cafaro, Christian Corda, Philip Cairns, Ayhan Bingolbali Data: 2024-05-02 Doi: Não disponivel

Abstract: Scientific imagination and experimental ingenuity are at the heart ofphysics. One of the most known instances where this interplay between theory(i.e., foundations) and experiments (i.e., technology) occurs is in thediscussion of Bell's inequalities. In this paper, we present a revisitation ofthe violation of Bell's inequality in the Clauser-Horne-Shimony-Holt (CHSH)form with entangled quantum states. First, we begin with a discussion of the1935 Einstein-Podolski-Rosen (EPR) paradox (i.e., incompleteness of quantummechanics) that emerges from putting the emphasis on Einstein's locality andthe absolute character of physical phenomena. Second, we discuss Bell's 1971derivation of the 1969 CHSH form of the original 1964 Bell inequality in thecontext of a realistic local hidden-variable theory (RLHVT). Third, identifyingthe quantum-mechanical spin correlation coefficient with the RLHVT one, wefollow Gisin's 1991 analysis to show that quantum mechanics violates Bell'sinequality when systems are in entangled quantum states. For pedagogicalpurposes, we show how the extent of this violation depends both on theorientation of the polarizers and the degree of entanglement of the quantumstates. Fourth, we discuss the basics of the experimental verification ofBell's inequality in an actual laboratory as presented in the original 1982Aspect-Grangier-Roger (AGR) experiment. Finally, we provide an outline of someessential take home messages from this wonderful example of physics at itsbest.

[Distância semântica $L2^2$= 0.900] Titulo: Filtering one-way Einstein-Podolsky-Rosen steering Autores: Ze-Yan Hao, Yan Wang, Jia-Kun Li, Yu Xiang, Qiong-Yi He, Zheng-Hao Liu, Mu Yang, Kai Sun, Jin-Shi Xu, Chuan-Feng Li, and Guang-Can Guo Data: 2024-01-04 Doi: Não disponivel

Abstract: Einstein-Podolsky-Rosen (EPR) steering, a fundamental concept of quantumnonlocality, describes one observer's capability to remotely affect anotherdistant observer's state by local measurements. Unlike quantum entanglement andBell nonlocality, both associated with the symmetric quantum correlation, EPRsteering depicts the unique asymmetric property of quantum nonlocality. Withthe local filter operation in which some system

components are discarded,quantum nonlocality can be distilled to enhance the nonlocal correlation, andeven the hidden nonlocality can be activated. However, asymmetric quantumnonlocality in the filter operation still lacks a well-rounded investigation,especially considering the discarded parts where quantum nonlocal correlationsmay still exist with probabilities. Here, in both theory and experiment, weinvestigate the effect of reusing the discarded particles from local filter. Weobserve all configurations of EPR steering simultaneously and other intriguingevolution of asymmetric quantum nonlocality, such as reversing the direction ofone-way EPR steering. This work provides a perspective to answer "What is theessential role of utilizing quantum steering as a resource?", and demonstratesa practical toolbox for manipulating asymmetric quantum systems withsignificant potential applications in quantum information tasks.

[Distância semântica L2²= 0.914] Titulo: Can Quantum Nonlocality be the Consequence of Faster-Than-Light Interactions? Autores: Luiz Carlos Ryff Data: 2024-01-18 Doi: Não disponivel

Abstract: It has been advocated by Bell and Bohm that the Einstein-Podolsky-Rosen (EPR)correlations are mediated through faster-than-light (FTL) interactions. In aprevious paper a way to avoid causal paradoxes derived from this FTL hypothesis(via the breakdown of Lorentz symmetry) has been suggested. Lorentztransformations would remain valid, but there would be no equivalence betweenactive and passive Lorentz transformations in the case of EPR correlations.Some counterintuitive consequences of this assumption are briefly examinedhere.

[Distância semântica L2²= 0.9176] Titulo: Quantum Reality Erasure with Spacelike-Separated Operations Autores: J. S. Ara'ujo, Diego S. Starke, A. S. Coelho, J. Maziero, G. H. Aguilar and R. M. Angelo Data: 2024-07-11 Doi: Não disponivel

Abstract: In 1935, Einstein, Podolsky, and Rosen argued that quantum mechanics isincomplete, based on the assumption that local actions cannot influenceelements of reality at a distant location (local realism). In this work, usinga recently defined quantum reality quantifier, we show that Alice's localquantum operations can be correlated with the erasure of the reality ofobservables in Bob's causally disconnected laboratory. To this end, weimplement a modified optical quantum eraser experiment, ensuring that Alice'sand Bob's measurements remain causally disconnected. Using an entangled pair ofphotons and quantum state tomography, we experimentally verify that, even withthe total absence of any form of classical communication, the choice of quantumoperation applied by Alice on her photon is correlated with the erasure of aspatial element of reality of Bob's photon. In this case, it is shown thatBob's photon can entangle two extra non-interacting degrees of freedom, thusconfirming that Bob's photon path is not an element of physical reality.

[Distância semântica L2²= 0.9220] Titulo: Einstein-Podolsky-Rosen steering paradox "2=1" for N qubits Autores: Zhi-Jie Liu, Jie Zhou, Hui-Xian Meng, Xing-Yan Fan, Mi Xie, Fu-lin Zhang, and Jing-Ling Chen Data: 2024-06-26 Doi: <https://doi.org/10.1142/S0217732324500305>

Abstract: Einstein-Podolsky-Rosen (EPR) paradox highlights the absence of a localrealistic explanation for quantum mechanics, and shows the incompatibility ofthe local-hidden-state models with quantum theory. For N-qubit states, ormore importantly, the N-qubit mixed states, we present the EPR steeringparadox in the form of the contradictory equality "2=1". We show that thecontradiction holds for any N-qubit state as long as both the pure staterequirement and the measurement requirement are satisfied. This also indicatesthat the EPR steering paradox exists in more general cases. Finally, we givespecific examples to demonstrate and analyze our arguments.

[Distância semântica L2²= 0.93] Titulo: My discussions of quantum foundations with John Stewart Bell Autores: Marian Kupczynski Data: 2024-03-26 Doi: <https://doi.org/10.1007/s10699-024-09946-z>

Abstract: In 1976, I met John Bell several times in CERN and we talked about a possibleviolation of optical theorem, purity tests, EPR paradox, Bell inequalities andtheir violation. I review our discussions, and explain how they were related tomy earlier research. I also reproduce handwritten notes, which I gave to Bellduring our first meeting and a handwritten letter he sent to me in 1982. Wehave never met again, but I have continued to discuss BI-CHSH inequalities andtheir violation in several papers. The research stimulated by Bell papers andexperiments performed to check his inequalities led to several importantapplications of quantum entanglement in quantum information and quantumtechnologies. Unfortunately, it led also to extraordinary metaphysical claimsand speculations about quantum nonlocality and retro-causality, which in ouropinion John Bell would not endorse today. BI-CHSH inequalities are violated inphysics and in cognitive science, but it neither proved the completeness ofquantum mechanics nor its nonlocality. Quantum computing advantage is not dueto some magical instantaneous influences between distant physical systems.Therefore one has to be cautious in drawing far-reaching philosophicalconclusions from Bell inequalities.

[Distância semântica L2²= 0.935] Titulo: Quantum Entanglement without nonlocal causation in (3,2)-dimensional spacetime Autores: Marco Pettini Data: 2024-05-24 Doi: Não disponivel

Abstract: This work aims at exploring whether the nonlocal correlations due to quantumentanglement could exist without nonlocal causation. This is done with the aidof a toy model to investigate whether the ability of two quantum entangledparticles to "correlate" their behaviors even at very large distances and inthe absence of any physical connection can be seen as due to an exchange ofinformation through an extra-temporal dimension. Since superluminal informationexchange is forbidden in our (3,1) spacetime, an extra-temporal dimension isneeded to recover the physical picture of finite velocity information exchangebetween entangled entities. Assuming that the geometry of space-time ofdimension (3,2) is described by a metric containing a warping factor, theconfinement of the massive particles in the extra time dimension follows.Therefore, why we do not experience an infinitely large extra time dimensioncan be explained. The toy model proposed here is defined by borrowingBohm-Bub's proposal to describe the wavefunction collapse using nonlinear(non-unitary) dynamical equations and then elaborating this

approach for anentangled system. The model so obtained is just speculative without any claimof being robust against any criticism, nevertheless, it satisfies the purposeof giving the possibility to the hypotheses formulated above to be verifiedexperimentally; in fact, it proposes an experiment potentially interestingwhich would otherwise be immediately dismissed as manifestly trivial. Theproposed experiment would consist of checking the possible violation of Bell'sinequality between two identical but independent systems under appropriateconditions. Beyond its theoretical interest, entanglement is a key topic inquantum computing and quantum technologies, so any attempt to gain a deeperunderstanding of it could be useful.

[Distância semântica L2²= 0.9428] Titulo: Einstein locality: An ignored core element of quantum mechanics Autores: Sheng Feng Data: 2024-03-20 Doi: Não disponivel

Abstract: Quantum mechanics is commonly accepted as a complete theory thanks toexperimental tests of non-locality based on Bell's theorem. However, wediscover that the completeness of the quantum theory practically suffered fromdetrimental ignorance of a core element – Einstein locality. Without thiselement, important experimental results of relevance could hardly receive fullunderstanding or were even completely misinterpreted. Here we present thediscovery with a theory of Einstein locality developed to recover thecompleteness of quantum mechanics. The developed theory provides a unifiedframework to account for the results of, e.g., Bell experiments (on Bellnon-locality) and double-slit experiments with entangled photons (onwave-particle duality). The theory reveals the dynamics of Bell non-localityand the principle of biased sampling in measurement in the double-slitexperiments, which otherwise will be impossible tasks without introducingEinstein locality. Worse still, ignorance of this element has causedmisinterpretation of observations in the double-slit experiments, leading toperplexing statements of duality violation. Einstein locality also manifestsindispensability in theory by its connection to the foundations of otherfundamental concepts and topics (e.g., entanglement, decoherence, and quantummeasurement) and may advance quantum technology by offering a promisingapproach to optimizing quantum computing hardware.

[Distância semântica L2²= 0.9441] Titulo: Experimental Test of Irreducible Four-Qubit Greenberger-Horne-Zeilinger Paradox Autores: Zu-En Su, Wei-Dong Tang, Dian Wu, Xin-Dong Cai, Tao Yang, Li Li, Nai-Le Liu, Chao-Yang Lu, Marek .Zukowski, and Jian-Wei Pan Data: 2024-07-11 Doi: <https://doi.org/10.1103/PhysRevA.95.030103>

Abstract: Bell's theorem shows a profound contradiction between local realism andquantum mechanics on the level of statistical predictions. It does not involvedirectly Einstein-Podolsky-Rosen (EPR) correlations. The paradox ofGreenberger-Horne-Zeilinger (GHZ) disproves directly the concept of EPRelements of reality, based on the EPR correlations, in an all-versus-nothingway. A three-qubit experimental demonstration of the GHZ paradox was achievednearly twenty years ago, and followed by demonstrations for more qubits. Still,the GHZ contradictions underlying the tests can be reduced to three-qubit one.We show an irreducible four-qubit GHZ paradox, and report its experimentaldemonstration. The reducibility loophole is closed. The bound of athree-setting per party Bell-GHZ inequality is violated by $7sigma$. Thefidelity of the GHZ state was around 81%, and an entanglement witnessreveals a violation of the separability threshold by $19sigma$.

[Distância semântica L2²= 0.9454] Titulo: Generalized Einstein-Podolsky-Rosen Steering Paradox Autores: Zhi-Jie Liu, Xing-Yan Fan, Jie Zhou, Mi Xie, and Jing-Ling Chen Data: 2024-06-06 Doi: Não disponivel

Abstract: Quantum paradoxes are essential means to reveal the incompatibility betweenquantum and classical theories, among which the Einstein-Podolsky-Rosen (EPR)steering paradox offers a sharper criterion for the contradiction betweenlocal-hidden-state model and quantum mechanics than the usual inequality-basedmethod. In this work, we present a generalized EPR steering paradox, whichpredicts a contradictory equality $2_Q = left(1 + deltaight)_C(0leqdelta < 1)$ given by the quantum (Q) and classical (C) theories. Forany N-qubit state in which the conditional state of the steered party ispure, we test the paradox through a two-setting steering protocol, and findthat the state is steerable if some specific measurement requirements aresatisfied. Moreover, our construction also enlightens the building of EPRsteering inequality, which may contribute to some schemes for typical quantumteleportation and quantum key distributions.

Apresenta uma explicação profunda do paradoxo e apresenta uma grande quantidade de referências precisas e recentes (2024). O modelo também conseguiu desenvolver conclusões acerca dos experimentos recentes aplicadas ao paradoxo.

5 Conclusões Parciais

A escolha da pergunta é complexa, pois estudos recentes continuam a não atingir um consenso. Um exemplo é o questionamento do papel do observador e sua função no colapso da realidade persistem a discordar (FRAUCHIGER; RENNER, 2018) (BRUKNER, 2018) (AL., 2019), não chegando a nenhum avanço experimental que contradiz o já conhecido. Além disto, experimentos realizados podem ser mal compreendidos devido a sua natureza.

Dos modelos utilizados, o Meta AI apresentou o pior desempenho. Os modelos Meta AI e Gemini 1.5 flash apresentaram erros em suas respostas. Os modelos Gpt-4o e o1-preview apresentaram desempenhos superiores, se comparados com os modelos anteriores. Contudo, a falta de referências ou apresentação de artigos anteriores a 2015, impossibilitou que eles respondessem ao requisito presente na pergunta de teste. Já o Tobias apresentou a explicação do paradoxo EPR com as implicações para a não localidade e analisou como os experimentos recentes estão abordando o tema, dispondo 10 referências de artigos pertinentes ao tema.

O uso de modelos generativos estão cada vez mais presentes no meio acadêmico. Contudo, problemas como informações incorretas, não verificáveis ou desatualizadas as tornam não confiáveis, além de possíveis brechas de informações dependendo de onde o modelo é utilizado. Com ferramentas como o Tobias, é possível solucionar tais problemas por trabalhar com um grande volume de dados acadêmicos de um repositório confiável, assim como aprimorar a consulta de informações no *VectorStore* para obter as melhores referências dada uma entrada do usuário. Isso pode ser associado aos modelos comerciais de melhor desempenho, tendo as informações da consulta protegidas de brechas de informação por não serem utilizadas para o treino de novos modelos.

Referências

AL., M. P. et. Experimental test of local observer independence. In: . [S.l.]: Sci. Adv.5,eaaw9832, 2019.

ARXIV*. Open access research sharing platform. acesso em: 12 de novembro de 2024. In: . [S.l.]: Disponível em https://info.arxiv.org/help/bulk$_{d}ata/index.html$., 2024.

BENTIVI, J. J. B. M. e S. Obtenção da equação de einstein pelo método diferencial e variacional. 2018. 35 f. In: . [S.l.]: Centro de Ciências, Universidade Federal do Ceará, Fortaleza, 2018.

BRUCKHAUS, T. Rag does not work for enterprises. In: . [S.l.]: Strative.ai, 2024.

BRUKNER Časlav. A no-go theorem for observer-independent facts. In: . [S.l.]: Entropy, 20(5), 350, 2018.

CUCONASU, F. et al. The power of noise: Redefining retrieval for rag systems. In: . [S.l.: s.n.], 2024.

FINARDI, P. et al. The chronicles of rag: The retriever, the chunk and the generator. In: . [S.l.: s.n.], 2024.

FRAUCHIGER, D.; RENNER, R. Quantum theory cannot consistently describe the use of itself. In: . [S.l.]: Nat Commun 9, 3711, 2018.

GIL, A. C. Como elaborar projetos de pesquisa. 7. ed. São Paulo: Atlas, 2022.

LV, C. et al. Towards biologically plausible computing: A comprehensive comparison. In: . [S.l.: s.n.], 2024.

OPENAI*. Hello GPT-4o, 13 Maio 2024. Acesso em: 12 de novembro de 2024. [S.l.]: Disponível em https://openai.com/index/hello-gpt-4o ., 2024.

OPENAI*. Learning to reason with llms, 12 setembro 2024. acesso em: 12 de novembro de 2024. In: . [S.l.]: Disponível em https://openai.com/index/learning-to-reason-with-llms/ ., 2024.

ROSENBAUM, R. On the relationship between predictive coding and backpropagation. In: . [S.l.: s.n.], 2024.

SAWARKAR, K.; MANGAL, A.; SOLANKI, S. R. Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. In: . [S.l.: s.n.], 2024.

VASWANI, A. et al. Attention is all you need. In: . [S.l.: s.n.], 2017.

ZHAO, Y. et al. Poor man's training on mcus: A memory-efficient quantized back-propagation-free approach. In: . [S.l.: s.n.], 2024.