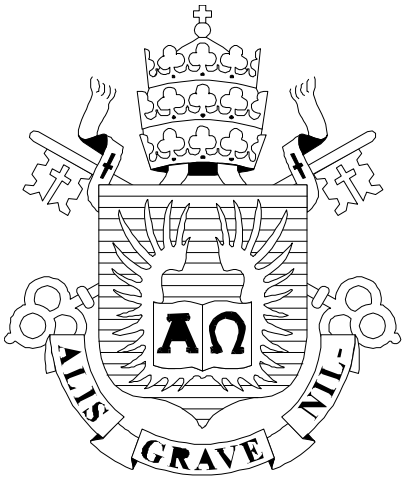

PUC

Monografia

RAG (Reinforced Augmented Generation) no Ambiente Corporativo: Oportunidades e Desafios.

Alexandre Fettermann Coutinho

Orientador: Pedro Gomes

Especialização em Transformação Digital

Junho/2025

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO

RUA MARQUÊS DE SÃO VICENTE, 225 - CEP 22451-900

RIO DE JANEIRO - BRASIL

RAG (Reinforced Augmented Generation) no Ambiente Corporativo: Oportunidades e Desafios.

Alexandre Fettermann Coutinho

alexfett@gmail.com

Abstract. This paper aims to present the use of AI (Artificial Intelligence) through the application of the RAG (Reinforced Augmented Generation) model as an alternative for decision-making support in large corporations. Initially, the fundamental technological aspects of RAG are presented, followed by a discussion of the opportunities provided by the model in the current corporate landscape. Among the benefits are increased accuracy, efficiency in data analysis, and personalization of user interactions, which enhance decision-making. Next, the challenges faced in adopting RAG are addressed, such as the need for robust technological infrastructure, managing large volumes of data dispersed across various databases, and ensuring information security. Finally, the work concludes that, despite the challenges, the adoption of RAG in the corporate environment offers a significant opportunity for improving decision support, becoming a valuable tool for companies seeking to stand out in an increasingly competitive market.

Keywords: Artificial Intelligence, Data Management, Decision Support, Information Security, Large Language Model, Natural Language Processing, Technological Infrastructure, Reinforced Augmented Generation

Resumo. Este trabalho tem como objetivo apresentar a utilização da IA (Inteligência Artificial) através da aplicação do modelo RAG (Reinforced Augmented Generation) como uma alternativa para o suporte à tomada de decisões em grandes corporações. Inicialmente são apresentados os aspectos tecnológicos fundamentais do RAG, seguidos de uma discussão sobre as oportunidades proporcionadas pelo modelo no cenário atual das grandes corporações. Entre os benefícios destacam-se a maior precisão, eficiência na análise de dados e personalização de interações com usuários, que potencializam a melhoria na tomada de decisões. Em seguida, são abordados os desafios enfrentados na adoção do RAG, como a necessidade de uma infraestrutura tecnológica robusta, a gestão de grandes volumes de dados dispersos em diversas bases e a garantia de segurança das informações. Por fim, o trabalho conclui que, apesar dos desafios, a adoção do RAG no ambiente corporativo oferece uma oportunidade significativa para a melhoria do suporte à decisão, tornando-se uma ferramenta valiosa para empresas que buscam se destacar em um mercado cada vez mais competitivo..

Palavras-chave: Inteligência Artificial, Reinforced Augmented Generation, Gestão de Dados, Grande Modelo de Linguagem, Infraestrutura tecnológica, Processamento de Linguagem Natural, Segurança da Informação, Suporte à Decisão.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

API – Application Programming Interface

AI – Artificial Inteligence

BPTT – Back Propapgation Through Time

CAPEX – Capital Expenditure

CEO – Chief Executive Officer

CPU – Central Processing Unit

CXO – Chief Experience Officer

IA – Inteligência Artificial

GenAI – Generative Artificial Intelligence

GPT – Generative Pre-Trained Transformer

GPU – Graphics Processing Unit

LLM – Large Language Model

MIPS – Maximum Inner Product Search

ML – Machine Learning

NLP – Natural Language Processing

PDF – Printable Document Format

RAG – Reinforced Augmented Generation

RNN – Recurral Neural Network

seq-2-seq – Sequence to Sequence

ROI – Return Over Investment

TI – Tecnologia da Informação

TX – Total Experience

1 Introdução

Desde de 2022, (com grande crescimento em 2023), as tecnologias de Inteligência Artificial (IA) generativas têm ganho destaque no ambiente corporativo, transformando a maneira como as empresas trabalham (Michael Chui, 2023). Em uma pesquisa reportada pela consultoria McKinsey & Company® esse aumento é de cinco vezes, com um incremento de 4% para 22% de utilização em funções corporativas (HEIMES, SHIRALI, WOODCOCK, & GOSWAMI, 2024). Já a Delloite® através de seu centro de pesquisa, informa que aproximadamente 80% dos líderes empresariais e de TI esperam que a IA generativa impulsione uma transformação significativa em suas indústrias nos próximos três anos, com investimentos que saíram de cerca de US$ 3 bilhões em 2022 para US$ 25 bilhões em 2023, com projeções de atingir aproximadamente US$ 150 bilhões até 2027 (SHACT, KREIT, VERT, HOLDOWSKY, & BUCKLEY, 2024)

Apesar de grandes consultorias e empresas de TI exaltarem os benefícios na adoção dessas tecnologias, a geração de valor diretamente relacionada a tomada de decisão ainda não é uma realidade a partir do uso de IA. A maioria das implementações estão focadas em eficiência como por exemplo: atendimento a clientes e automação de tarefas, ao invés da eficácia de processo (HEIMES, SHIRALI, WOODCOCK, & GOSWAMI, 2024) tais como: análise de dados operacionais, assistência à tomada de decisão e geração de relatórios personalizados. Porém, no médio prazo naturalmente as empresas buscarão aplicações mais complexas. A figura 1, retirada de (HEIMES, SHIRALI, WOODCOCK, & GOSWAMI, 2024) traz uma síntese de uma pesquisa realizada com 276 *Chief Experience Officers* -CXO (Diretor de Experiência do Usuário) quanto a abordagem do uso de IA Generativa relativamente à eficiência ou eficácia de processos.

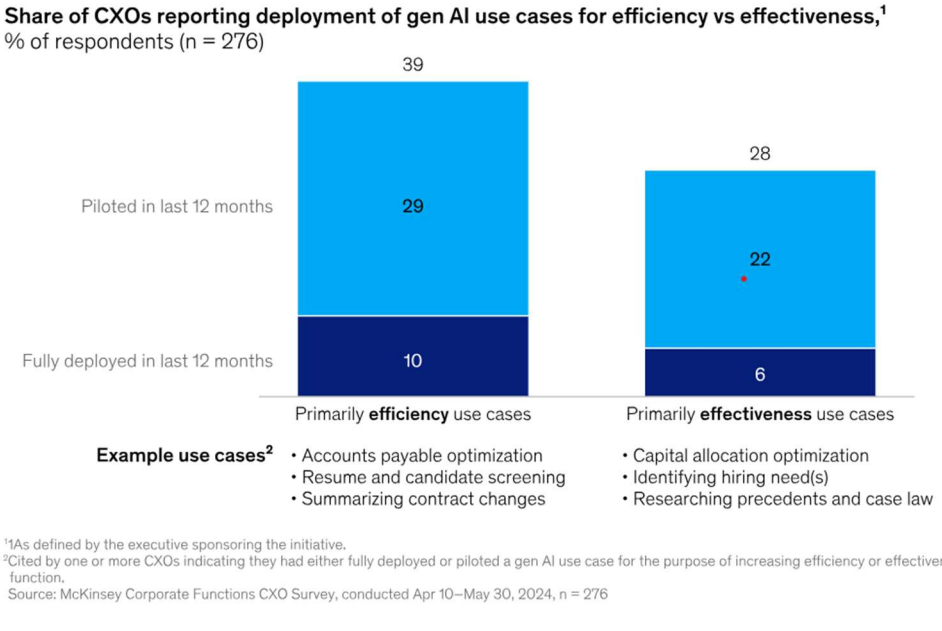

Figura 1 - Os usos mais frequentes da IA generativa nas funções corporativas são aqueles focados em eficiência, em vez de eficácia.

Os Modelos de Linguagem de Grande Escala (*LLMs*, do inglês Large Language Models) e de Recuperação Aumentada por Geração (*RAG*, do inglês Retrieval-Augmented Generation) são duas tecnologias centrais da IA Generativa (*GenAI*, do inglês Generative Artificial Inteligence) e são o objeto de análise deste estudo. Outro aspecto fundamental que deve ser considerado na aplicação da *GenAI* está relacionado aos principais problemas intrínsecos a esta tecnologia, dos quais os principais são: segurança e privacidade de dados (Jeff Pollard, 2023), falta de rastreabilidade (Litan, 2024) e incertezas regulatórias (Walsh, 2023), sendo o primeiro considerado por muitos como barreira para a adoção dessa tecnologia (Jeff Pollard, 2023) e (Stamford, 2025).

Como ponto de partida, apresenta-se a seguir de maneira resumida numa tabela comparativa os principais aspectos relevantes para esse estudo entre os modelo RAG e LLMs pré-treinadas sem o RAG. A tabela é um resumo das referências: (Jeff Pollard, 2023), (Litan, 2024) e (Stamford, 2025)

Aspecto	RAG	Somente LLM
Custo	Reduz a necessidade ou elimina. Se utiliza de modelos pré-treinados e específicos. Pode ser expandido sem necessidade de treinar o modelo	Necessidade de treinamento sobre grandes quantidades de dados.
Privacidade	É alta já que se utiliza de dados internos à corporação.	É baixa. Pode ser acessado por APIs externas
Segurança	É alta já que o acesso pode ser controlado aos dados.	É baixa já que o risco de exposição aos dados é mais amplo
Rastreabilidade	É alta pois a fonte dos dados está disponível e pode ser curada pela organização.	É baixa já que a geração do conteúdo não é bem definida (caixa-preta)

Tabela 1 - Comparativo sumário de alguns pontos críticos entre LLM e RAG nos ambientes corporativos

Nesse contexto, um dos modelos que surge com uma alternativa importante no auxílio da geração de valor para os processos decisórios recai sobre um dos modelos da IA generativa RAG, que traz algumas vantagens importantes para utilização de LLM em um contexto empresarial, no intuito de prover informações a decisores em tempo real de maneira mais segura, com rastreabilidade nas informações, especificidade em temas e com custos menores de treinamento dos modelos.

Esse trabalho traz uma sugestão de utilização de RAG em ambientes corporativos como ferramenta de apoio para à tomada de decisão.

2 Contextualização

2.1 A IA Generativa (GenAI) em ambientes corporativos

Pode se definir de maneira sucinta o termo *GenAI* como um ramo da inteligência artificial que se utiliza de tecnologias específicas para criar novos conteúdos, a partir de dados existentes (Microsoft, 2025). No pilar dessa tecnologia estão os modelos generativos, também conhecidos como modelos de base ou fundação (*foundation models*[1]). Assim, a IA Generativa é capaz de produzir desde de textos, imagens, a vídeos e músicas que não haviam sido criados anteriormente, a partir de conteúdos pré-existentes.

A aplicação dessa tecnologia em contextos empresariais estruturados, principalmente em fluxos de trabalho com base conteúdos específicos e confiáveis, trazem grande oportunidade de otimização de processos e geração de valor. Conforme descrito anteriormente na Introdução, o grande entusiasmo e a explosão na oferta de várias ferramentas demonstram uma grande corrida, não só dos fornecedores em prover ferramentas mais atrativas para esse mercado, como também da própria liderança em gerar valor nos negócios através de sua adoção. No entanto, segundo a pesquisa do Grupo Gartner "*Hype Cycle for Artificial Intelligence*", publicada em Agosto de 2024, (STAMFORD, Conn., 2024), para muitas organizações a *GenAI* entrou no vale da desilusão.

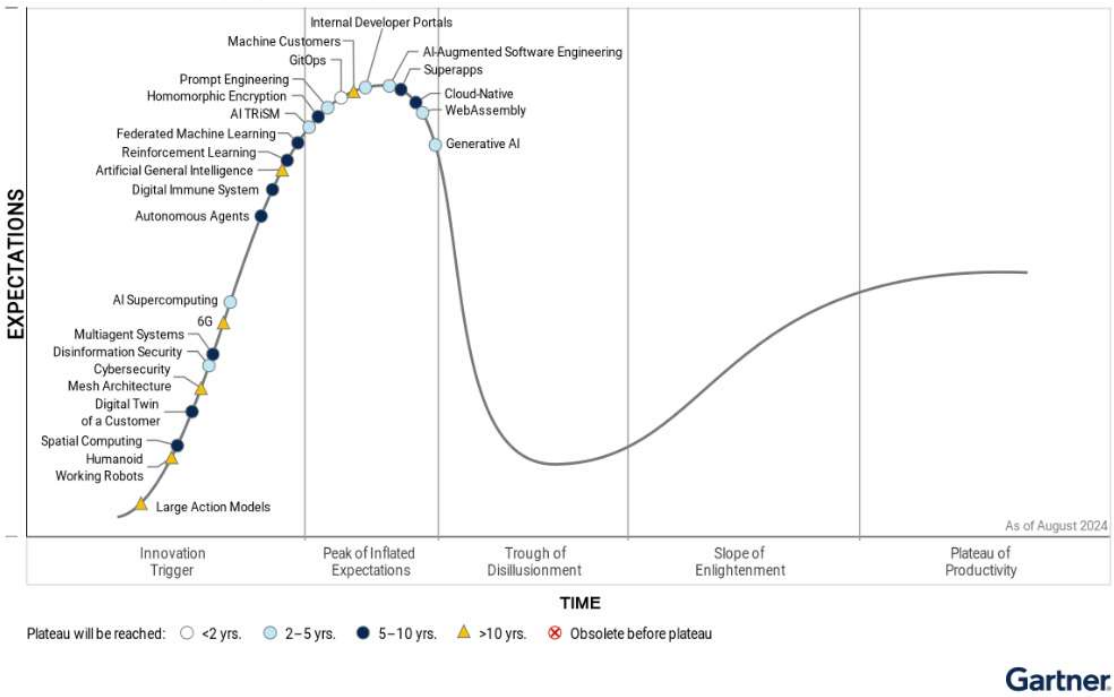

Figura 2 - Hype Cycle para as Tecnologias emergentes, 2024[2] *(STAMFORD, Conn., 2024)*

Em um primeiro momento, essa desaceleração pode demonstrar uma pequena perda de importância, mas segundo o executivo do próprio Gartner Arun Chandrasekaran, (Distinguished VP Analyst): "*GenAI está além do Pico das Expectativas Infladas, à medida que o foco dos negócios continua a mudar do entusiasmo em torno dos modelos fundamentais para*

1 What are foundation models? – IBM - https://www.ibm.com/think/topics/foundation-models

2 Até a produção deste trabalho este era o último relatório disponível com o tema: "*Hypecicle for Generative AI.*" O relatório atualizado para 2025 será publicado em 20 de Junho de 2025 - The 2025 Gartner Group

casos de uso que geram retorno sobre o investimento (ROI)". Portanto, essa tecnologia parece estar saindo de um estágio de experimentação para entrar num novo estágio de maturidade aonde as empresas buscam agregar valor mensurável dentro de seus fluxos de trabalhos. Ainda segundo o executivo, nesse mesmo estudo, (STAMFORD, Conn., 2024) o amadurecimento da utilização de *GenAI* em ambientes empresarias aponta quatro tópicos emergentes: IA Autônoma, aumento de produtividade de desenvolvedores, Total Experience (TX) e, Segurança e Privacidade Centrada no Usuário sendo que aquele que traz interesse para esse estudo é a chamada IA Autônoma em conjunto com modelo RAG. A combinação entre agentes autônomos de IA e modelos de recuperação aumentada surge então como uma proposta poderosa no suporte a tomada de decisão.

3 RERENCIAL TEÓRICO

3.1 LLMs

Para que possamos contextualizar de maneira adequada as chamadas LLMs, é necessário percorrer um caminho no mundo da AI e o ponto inicial, no sentido mais amplo, que é a IA Generativa.

Segundo (Wikipedia, 2025), "*Inteligência Artificial Generativa (IA Generativa, GenAI ou GAI) é um ramo da inteligência artificial voltado para a criação de novos conteúdos com base em padrões aprendidos a partir de grandes volumes de dados de treinamento. Essa tecnologia permite gerar textos, imagens, áudios, vídeos e até códigos de software, a partir de comandos em linguagem natural (prompts), imagens ou vídeos.*" A geração de conteúdo pode ser estar classificada em dos itens listados a seguir:

- **Texto para texto** (*text-to-text*), como na geração de respostas, resumos, traduções ou até mesmo livros inteiros;
- **Texto para imagem** (*text-to-image*), criando imagens a partir de descrições textuais, como fazem ferramentas populares tais como: DALL·E, Midjourney e Stable Diffusion;
- **Texto para código** (*text-to-code),* geração de código-fonte em linguagens de programação e scripts a partir de descrições textuais de escopo e descrição de erros tendo como entrada trecho de código-fonte. Exemplo GitHub Copilot.
- **Imagem para imagem** (*image-to-image*), transformação ou estilização de imagens tendo como base, outras imagens.
- **Vídeo para vídeo** (*video-to-video*), transformando ou aprimorando conteúdos visuais com base em entradas de vídeo.
- **Texto para vídeo** *(text-to-video)*: criação de vídeos realistas, ou artísticos a partir de descrições textuais.
- **Geradores de áudio** para criação de vozes sintéticas, músicas e efeitos sonoros, como Lyrebird.

Dentro desse contexto é preciso conhecer também os Modelos Generativos (Generative Models), que são capazes identificar padrões e distribuições nos dados de treinamento para gerar novos conteúdos com base em entradas fornecidas pelos usuários. Durante o treinamento, o modelo reconhece relações estatísticas entre características dos

dados e desenvolve uma lógica interna que orienta a criação de novas amostras semelhantes às originais.

Usualmente, tais modelos são treinados com grandes volumes de dados não rotulados, e a partir destes gera seus próprios rótulos. Por isso, são denominados de modelos autosupervisionados, que contrastam os modelos supervisionados em que os rótulos são entregues para o modelo (Bergman, s.d.). Por fim, a avaliação do desempenho desse modelo é feita por uma função de perda (*loss function*) (Wikipedia, Loss Function, 2025), que mede a diferença entre os resultados gerados e os reais. A estratégia é minimizar a diferença entre o os dados gerados e os reais, tornando as saídas cada vez mais próxima dos dados reais.

A geração de conteúdo é um processo probabilístico e portanto, o modelo não "sabe" como um humano, mas prevê a saída mais provável com base nas regras matemáticas aprendidas. Essa abordagem probabilística, também pode produzir eventualmente saídas que parecem estar fora do contexto, (IBMAIHALL, 2023), quando o modelo gera conteúdo que não se considera um resultado realista.

A definição da arquitetura de um modelo generativo irá definir uma série de características que impactará seu funcionamento. Para o objeto desse estudo o destaque é para os chamados *deep generative models* que são um subtipo dos modelos generativos que se baseiam em redes neurais profundas, com múltiplas camadas, *deep neural networks*. A classificação mais usual de modelos generativos é disposta a seguir, como em (Belcic, 2024)

- **Modelos autorregressivos** (*autoregressive models*) preveem o próximo ponto de dados em uma sequência com base nas instâncias anteriores.
- **Transformers** (*transformers*) se destacam em tarefas de processamento de linguagem natural (NLP) devido à sua capacidade aprimorada de compreender e utilizar o contexto.
- **Modelos de difusão** (*difusion models*) geram novos dados adicionando ruído gradualmente a um conjunto de dados e, em seguida, aprendem a remover esse ruído para produzir uma saída original.
- **Redes adversariais generativas** (*generatives adversarial networks* - GANs) combinam um modelo gerador e um discriminador em uma espécie de competição, onde o objetivo do gerador é criar saídas que enganem o discriminador.
- **Autoencoders variacionais** (*varational Autoencoders - VAEs*) comprimem os dados de entrada por meio de um *encoder* e depois os reconstroem com um *decoder*, gerando dados semelhantes aos originais.
- **Modelos baseados em fluxo (flow-based models)** aprendem as relações entre distribuições simples e complexas de dados por meio de operações matemáticas reversíveis.

Baseado no *papper* dos pesquisadores do Google lançado em junho de 2017, *Attenction is all you need* (Ashish Vaswani, 2017), (Figura 4) os *transformes* trouxeram inovações importantes para que se fossem superadas algumas limitações que o modelo até então utilizado comumente e que se baseava em um tipo de modelo autorgressivo, (*redes neurais recorrentes* ou *recurrent neural networks – RNN*). As RNNs, até então, eram utilizadas na

maioria das implementações de modelos *seq-2-seq*[3]. O funcionamento básico está descrito a seguir e está ilustrado na figura 3. Nesse exemplo, a RNN irá receber uma sequência *X* e irá produzir um sequência de saída *Y*:

1) Incialmente uma sequência X é quebrada (*split*) em sequências menores X_1, X_2, X_3,..., X_n denominados tokens de entrada (*iput tokens*)
2) A RNN recebe o *input token* X_1
3) O primeiro item X_1 juntamente com o estado inicial <0> (geralmente zeros) produz o token de saída (*output token*) Y_1
4) Na iteração seguinte, dado o estado oculto (*hidden state*) anterior e o próximo *input token* X_2 a RNN produz o *output token* Y_2.
5) As iterações são repetidas *N* vezes, tanto quanto forem os número de *input tokens*
6) Ao término serão gerados *N outputs token Y*

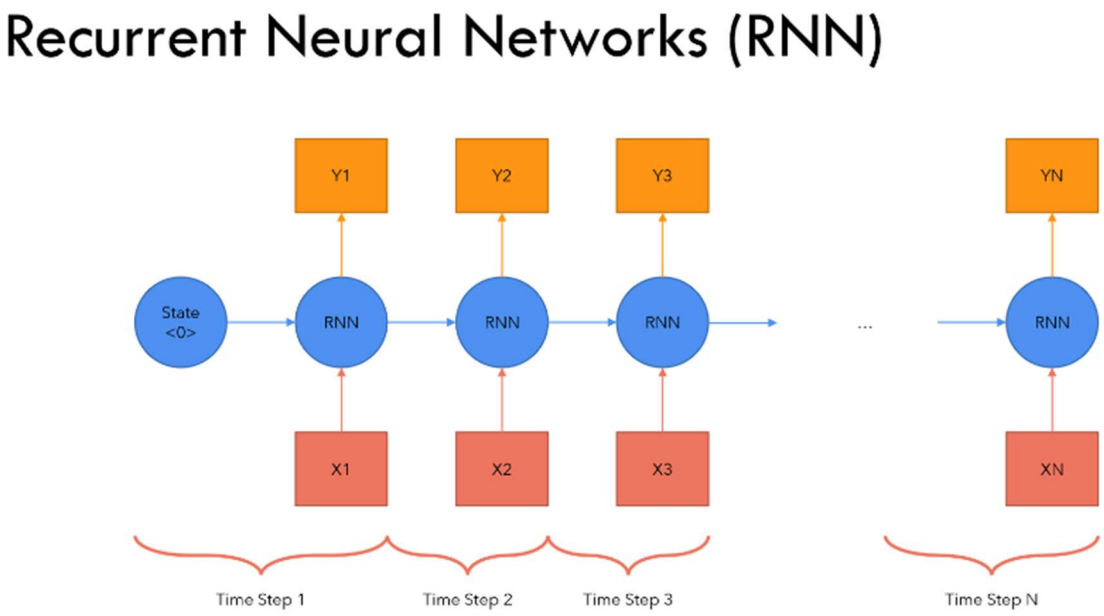

Figura 3 - RNN genérica. Fonte: (Jamil, 2023)

Vale ressaltar que o exemplo acima é genérico e que normalmente uma RNN gera *N output tokens*, dado *N input tokens*, mas na prática existem situações em que isso não ocorre. Existem cenários em que a RNN não gera uma saída por entrada:

- **Codificador-decodificador** (*encoder-decoder*): como em tradução automática, a RNN codificadora recebe *N input tokens*, mas o *decoder* pode gerar uma sequência de saída de comprimento diferente.
- **Classificação de sequência**: a RNN processa toda a sequência de entrada e gera **uma única saída** (por exemplo, para classificar o sentimento de uma frase).

Os pontos-chave envolvidos são:

3 Definição seq2seq na Wikipedia https://en.wikipedia.org/wiki/Seq2seq; Artigo introdutório (Ilya Sutskever,Oriol Vinyals,Quoc V. Le, 2014)

- **Token**: é a menor unidade significativa de entrada ou saída em uma sequência processada pela RNN, como uma palavra, caractere ou um pedaço de um palavra.
- **Hidden state**: carrega a memória da sequência anterior.
- **RNN**: processa a sequência passo a passo, atualizando o *Hidden State*.
- **Saída**: uma previsão por token de entrada.
- **Initial state**: é o vetor que representa a memória da RNN antes de processar o primeiro token da sequência — geralmente inicializado com zeros.

A partir dos pontos citados, podemos enumerar as principais desvantagens nas implementações de *seq-2-seq* baseadas em RNN:

1. **Processamento sequencial não paralelizável:** tanto as RNNs (se implementadas com mecanismos de atenção) quanto os *tansofrmers* têm complexidade **O(n²).** No entanto, as RNNs não são paralelizáveis e os *transformers* são. Esse paralelismo que as RNNs não possuem as torna ineficientes na prática para grandes volumes de dados.
2. **Dificuldade com dependências de longo prazo:** no exemplo anterior (figura 3) nota-se que o processamento é feito um token por vez e armazenado em cada *hidden state*. Em cadeias muito longas os últimos estados podem não mais sofrer influência dos estados iniciais. Portanto a contribuição dos estados iniciais praticamente desaparece para os estados finais. Para maiores detalhes o trabalho de (Olah, 2015) é uma referência bem conhecida.
3. **Problemas com gradientes:** para compreender esse problema é preciso entender que os *frameworks* que geralmente implementam RNNs, como por exemplo o *PyTorch* (Pytorch, 02)*)*, convertem as redes em grafos computacionais. Outro aspecto importante é que os algoritmos precisam calcular as derivadas das *loss funciton* (Wikipedia, Loss Function, 2025) em relação a cada um dos pesos ao longo do percurso desse grafo. Esse mecanismo é chamado *Backpropagation Through Time* (Wikipedia, Back Propagation Through Time, 02) *ou BPTT*. No cerne da questão está a chamada regra da cadeia (Wikipedia, Regra da cadeia, 2025), cujo cálculo das derivadas pode tornar o efeito do peso a ser aplicado muito pequeno (*vanishing*) ou muito grande (*gradiente explode*) ao ponto de extrapolar a precisão de GPUs (*Graphical processing Unit*) ou CPUs (*Central Processing Unit*) modernas, usualmente de 64 bits. Uma referência mais detalhada e com exemplos práticos pode ser encontrada em (GeekToGeek, 2025).

Os t*ransformers* introduziram duas inovações fundamentais em relação à RNNs que os tornaram o padrão de fato para modelos de linguagem de grande escala (LLMs) em IA generativa:

- **Processamento paralelo**: Ao contrário das RNNs, os t*ransformers* processam todos os elementos de uma sequência simultaneamente, o que aumenta significativamente a eficiência. Isso permite treinar modelos com grandes volumes de dados em menos tempo.
- **Mecanismos de autoatenção (***self-attention***)**: permite que o modelo avalie a importância relativa de cada item da sequência em relação aos demais. Com isso, os *transformers* conseguem capturar relações contextuais entre elementos distantes, o que é essencial para tarefas de processamento de linguagem natural, como geração de texto e tradução automática. O conceito chave reside em determinar

quais partes dos dados o modelo deve focar em cada momento, permitindo que ele processe grandes volumes de informação de forma eficiente.

- **Escalabilidade** (*scalability*): consegue lidar com grandes volumes de dados

Entre os três tipos de modelos *transformer* —**codificadores (*encoders*)**, **decodificadores (*decoders*)** e **codificador-decodificador (*encoder-decoder*)**— os dois últimos contêm componentes autoregressivos. O funcionamento básico é os seguinte: O encoder transforma a entrada em uma representação vetorial, enquanto o decoder reconstrói essa representação no mesmo tipo de dado da entrada original. Os *decoders* são responsáveis pela geração de conteúdo, utilizando autoregressão para prever o próximo *token* com base nos *tokens* já gerados.

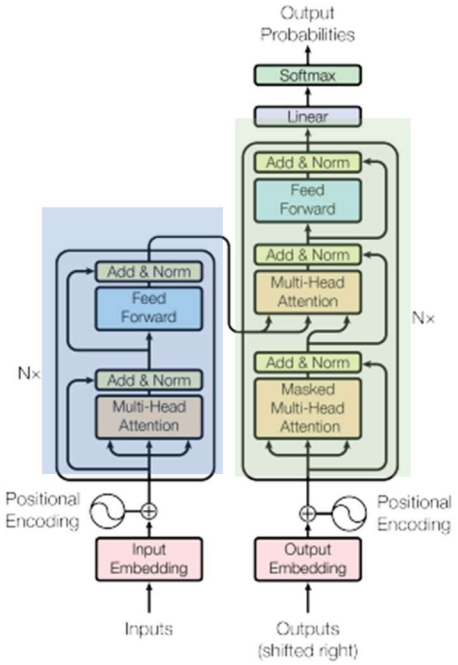

Figura 4 - Arquitetura do modelo transformer. Fonte: (Ashish Vaswani, 2017)

Na figura 4 que foi retirada do artigo de 2017 da Meta (Ashish Vaswani, 2017) destacamos os dois principais componentes dos *transformers*: o *encoder* (em azul) e o *deeocder* (verde). O processo de geração é basicamente descrito como a seguir:

1. Tokenização da Entrada (*Inputs*)

- **O que acontece**: A sequência de entrada (por exemplo, uma frase) é dividida em unidades menores chamadas *tokens* (palavras, *subpalavras* ou caracteres).
- **Por que é importante**: Isso permite que o modelo processe a linguagem de forma estruturada e numérica de forma individualizada.

2. Camada de Embedding (*Embedding Layer*)

- **O que acontece**: Cada token é convertido em um vetor denso (*embedding*) que transforma a linguagem natural para um espaço vetorial contínuo para que os *transformers* possam analisar seu valor semântico. De forma bem simplificada, os embedings irão transformar os *tokens* em um formato

matemático que possibilita o uso pelos *transormers*. Para um detalhamento maior desse processo, um bom documento a ser consultado é (Khan, 2025)

- **Por que é importante**: A partir dos vetores densos é possível que os LLMs capturem nuances semânticas e contextuais, permitindo assim a identificação de padrões complexos nos textos. Esse mecanismo possibilita que os LLMs interpretem o contexto, gere respostas coerentes e realize tarefas como tradução automática, sumarização de conteúdo e busca semântica com alta precisão, por exemplo. Um outro artigo que traz uma boa explicação pode ser acessado em (Jr, 2024)

3. Codificação Posicional (*Positional Encoding***)**

- **O que acontece**: Uma das grandes vantagens dos *transformers* é a capacidade de processamento paralelo. Com isso cada *token* é transferido do *encoder* para o *decoder* sem que se preserve o a posição de cada um. Consequentemente, se faz necessário incorporar a ordem das palavras no modelo, caso contrário o sentido semântico fica comprometido por completo.
- **Por que é importante**: Sem essa codificação, os *transformers*, ao processarem os *tokens* em paralelo, não conseguiriam "compreender" a posição de cada palavra na frase, o que é essencial para o significado de cada palavra no contexto em que estão inseridas.

4. Pilha de Encoders (*encoder stack***) -** *usado em modelos encoder-decoder como BERT[4] ou T5[5])*

Cada camada do *encoder* inclui:

- **Atenção Automática Multi-Cabeça** (*Multi-Head Self-Attention*): A compreensão do modelo de atenção de produto escalar é necessária para entender como o modelo de multi-cabeças funciona. De maneira simplificada, a auto atenção calcula pontuações de atenção a partir de um vetor de entrada para determinar quanto foco cada elemento da sequência deve ter sobre os demais. Isso é feito usando três matrizes-chave: **Consulta (Q) –** que possui o relacionamento da palavra atual com outras; **Chave (K)** que representa as palavras que estão sendo comparadas e **Valor (V)** que contém as representações reais das palavras. A auto atenção então é calculada como[6]:

$$Atenção\ (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V$$

No mecanismo de multi-cabeças, várias cabeças de atenção são usadas paralelamente, o que permite que o modelo se concentre em diferentes partes da sequência de entrada simultaneamente. Para o detalhamento do funcionamento desse mecanismo, uma fonte interessante pode ser encontrada em (Raschika, 2024)

4 BERT – Bidirectional encoder representation from transformers – https://en.wikipedia.org/wiki/BERT_(language_model).

5 T5 – Text-to-Text Decoder - https://huggingface.co/docs/transformers/en/model_doc/t5

6 Scale Dot-Product Attention: https://paperswithcode.com/method/scaled

- **Rede Neural Feedforward**: Sua principal função é refinar a saída da camada de atenção preparando-a adequadamente como entrada para a próxima etapa do processamento. Isso é feito através de ajuste de pesos durante o treinamento, aplicando a mesma matriz de transformação de forma independente a cada posição de token
- **Normalização de Camada e Conexões Residuais**: Ajudam a estabilizar e acelerar o treinamento.

5. Pilha de Decoders (*decoder stack*)

*(usado em modelos como GPT ou T5)*Cada camada do *decoder* inclui:

- **Atenção Automática Multi-Cabeça com Máscara (***Masked Multi-Head Self-Attention***)**: Além do que foi descrito no item Atenção Automática do *encoder*, nessa camada o termo *masked* indica que, durante o treinamento, o modelo é impedido de ver *tokens* futuros na sequência. Isso é feito aplicando uma máscara que zera as atenções para posições à frente do token atual. Isso se faz necessário para que o modelo consiga prever o próximo *token* com base apenas no que já foi visto.

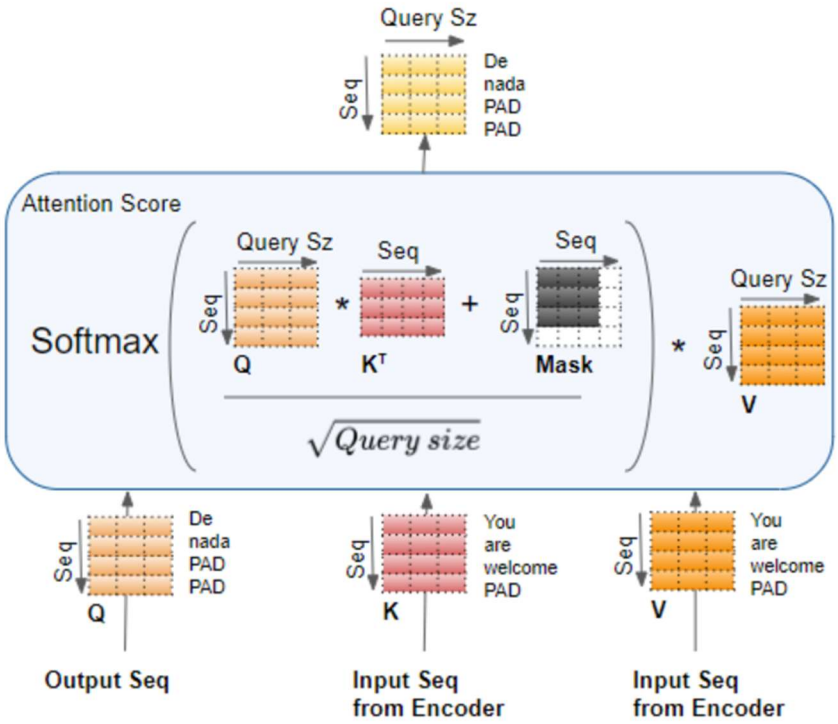

Figura 5 - Ilustração de aplicação de Masking em Encoder-Decoder Attention. Fonte: (Doshi, 2021)

- **Rede Neural Feedforward**: Igual à usada no *encoder*.

6. Projeção de Saída (*Output Projection*)

- **O que acontece**: a saída final do *decoder* é projetada em um vetor do tamanho do vocabulário, usando uma camada linear com a distribuição de probabilidade sobre todos os possíveis próximos *tokens*.

- **Por que é importante:** porque o modelo então utiliza essa distribuição para selecionar a próxima palavra ou símbolo mais provável na sequência.

7. Geração de Saída (*Output Generation***)**

- **O que acontece**: O modelo seleciona o próximo token mais provável.
- **Por que é importante**: É assim que o modelo gera texto coerente e apropriado ao contexto.

3.2 O que é RAG?

De forma objetiva e bem simplificada, RAG pode ser definida como uma técnica de IA que combina recuperação de informações com modelos de geração de texto (Gen IA). Nesse contexto, o "gerador de conteúdo" utiliza um elemento denominado *retriever* que busca informações relevantes e específicas ao tema, e os disponibiliza ao modelo de linguagem (*generator)*, que irá formular as respostas.

Mais detalhadamente, o RAG se inicia exatamente aonde a GenAI termina ao prover as informações que o modelo de LLMs falha em responder de maneira acurada. Assim como descrito no excelente trabalho de (Patrick Lewis, 2021), "*o modelo RAG foi projetado para LLMs*". Nesse artigo o conceito de RAG é introduzido justamente para preencher as lacunas deixadas pelos LLMs, ao abordar a questão "*...esses modelos têm desvantagens: eles não podem expandir ou revisar sua memória facilmente, não conseguem fornecer insights claros sobre suas previsões e podem produzir 'alucinações'*". A proposta dos autores se baseia em "*...modelos híbridos que combinam memória paramétrica com memórias não paramétricas (ou seja, baseadas em recuperação) podem resolver alguns desses problemas, pois o conhecimento pode ser diretamente revisado e expandido, e o conhecimento acessado pode ser inspecionado e interpretado*". De forma análoga poderíamos comprar esse processo a uma tarefa de um pesquisador que escreve sobre um tema em particular. Assim como qualquer outra pessoa, a habilidade de ler, escrever e resumir de forma genérica é comum e bem desenvolvida, porém o conhecimento específico do assunto necessário para que a elaboração da pesquisa resulte no desenvolvimento de um trabalho acurado, depende de conhecimento específico. Nesse cenário, poderíamos comparar os leigos às LLMs, que são ótimos em gerar conteúdo mas não conseguem identificar se este é acurado ou não, e a falta de conhecimento necessário poderá trazer conclusões absurdas (alucinações). Já o conhecimento do pesquisador pode ser interpretado como a capacidade de coletar e organizar as informações específicas e necessárias para elaboração de um trabalho com conclusão objetiva e aderente as questões pesquisadas.

A figura 6 explica o funcionamento básico proposto pelos pesquisadores da Meta criadores do artigo inaugural do modelo da RAG (Patrick Lewis, 2021), "*em que foram combinados um recuperador pré-treinado (encoder de consultas + índice de documentos) com um modelo seq2seq pré-treinado (Gerador) e realizou-se um ajuste fino de ponta a ponta. Para a consulta x, foi usada a Maximum Inner Product Search (MIPS)*[7] *para encontrar os documentos top-K* z_i*. Para a previsão final y, tratamos z como uma variável latente e marginalizamos sobre as previsões seq2seq dadas diferentes documentos*".

7 Definição MIPS na Wikipedia https://en.wikipedia.org/wiki/Maximum_inner-product_search;

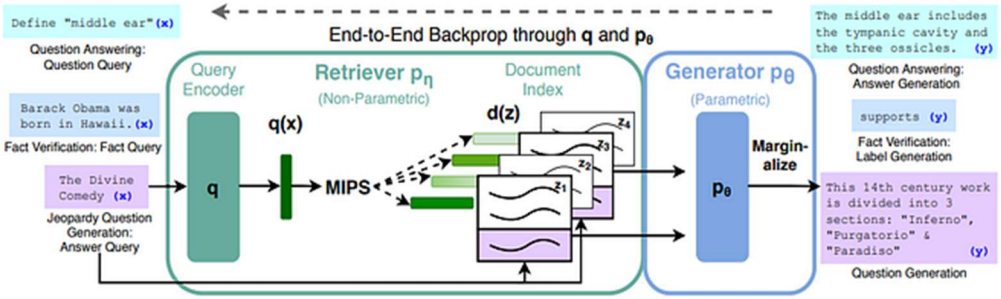

Figura 6 - O esquema apresentado pelos pesquisadores da Meta em 2021 (Patrick Lewis, 2021). Fonte: https://arxiv.org/pdf/2005.11401

Pode -se interpretar de maneira simplificada esse diagrama como sendo um processo onde são carregados dados em um LLM oriundos de múltiplas fontes, podendo ser dados estruturados, semiestruturados ou não estruturados. Na sequência um processo de indexação é aplicado sobre essas informações para que o processo de busca possa ser viabilizado. A partir de então, as consultas dos usuários podem utilizar os índices, para o contexto mais relevante. Por fim, esse contexto gerado e a consulta do usuário são enviados para o LLM que processa o conteúdo e gera a resposta, geralmente através de texto em prompt.

3.3 Tipos de RAG

O modelo inicialmente proposto por (Patrick Lewis, 2021), conhecido como Naïve RAG que se baseava em três componentes distintos: *indexing* (indexação), *retrieval* (recuperação) e *generation* (geração) mostrou-se ineficiente em algum casos mais complexos. Conforme descrito em (Yunfan Gao, 2024) em artigo denominado *Modular RAG: Transforming RAG Systems into LEGO-like Reconfigurable Frameworks* "*Os principais desafios do Naive RAG incluem:*

1. ***Compreensão Superficial das Consultas**: A similaridade semântica entre uma consulta e um fragmento de documento nem sempre é altamente consistente. Confiar apenas em cálculos de similaridade para a recuperação carece de uma exploração aprofundada da relação entre a consulta e o documento.*

2. ***Redundância e Ruído na Recuperação**: Alimentar todos os fragmentos recuperados diretamente nos LLMs nem sempre é benéfico. Pesquisas indicam que um excesso de informações redundantes e ruidosas pode interferir na identificação de informações chave pelos LLMs, aumentando assim o risco de gerar respostas errôneas e alucinatórias."*

O estudo traz três abordagens de modelos que dependem basicamente da complexidade do projeto a ser implementado:

- **Naïve RAG:** Esse tipo de implementação de RAG não envolve indexações e incorporações de dados complexas. Nesse cenário palavras-chave simples acessam de maneira eficiente um conjunto específico de dados
- **Advanced RAG:** Esse cenário envolve elementos mais complexos como busca vetorial e recuperação baseada em índice. Nesse método podem existir múltiplos tipos de fontes de dados em formatos estruturados ou não estruturados.

- **Modular RAG:** Engloba os cenários de Naïve e Advanced e outros com grande complexidade, mesmo que precisem incluir algoritmos de Machine Learning para atingir os resultados necessários

A seguir um esquema comparativo entre os tipos de RAG

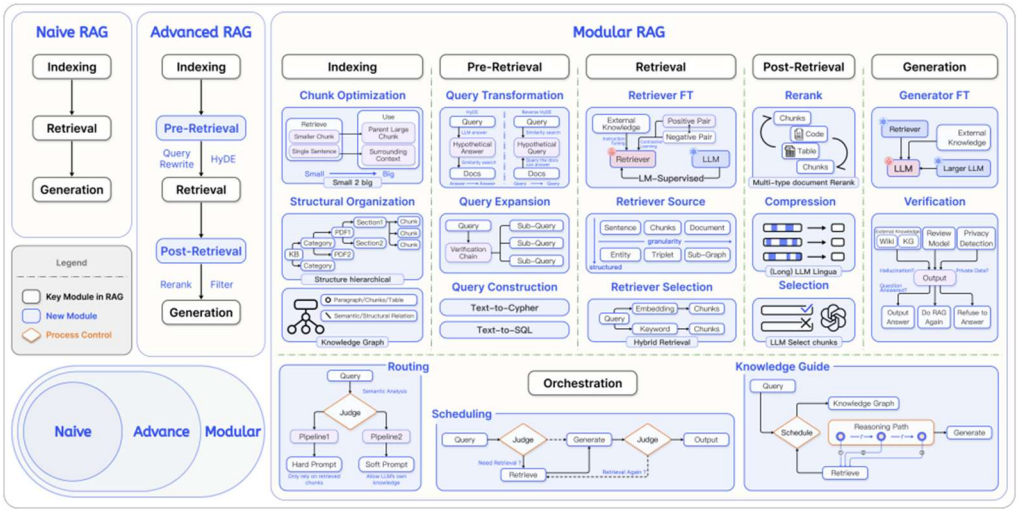

Figura 7 - Comparativo entre Naïve, Advanced e Modular RAG. Fonte: retirado do artigo (Yunfan Gao, 2024)

Na sequência é apresentado um caso de Naïve e Advanced RAG quando submetidos com perguntas complexas. Nota-se que ambos enfrentam limitações e têm dificuldade em fornecer respostas satisfatórias.

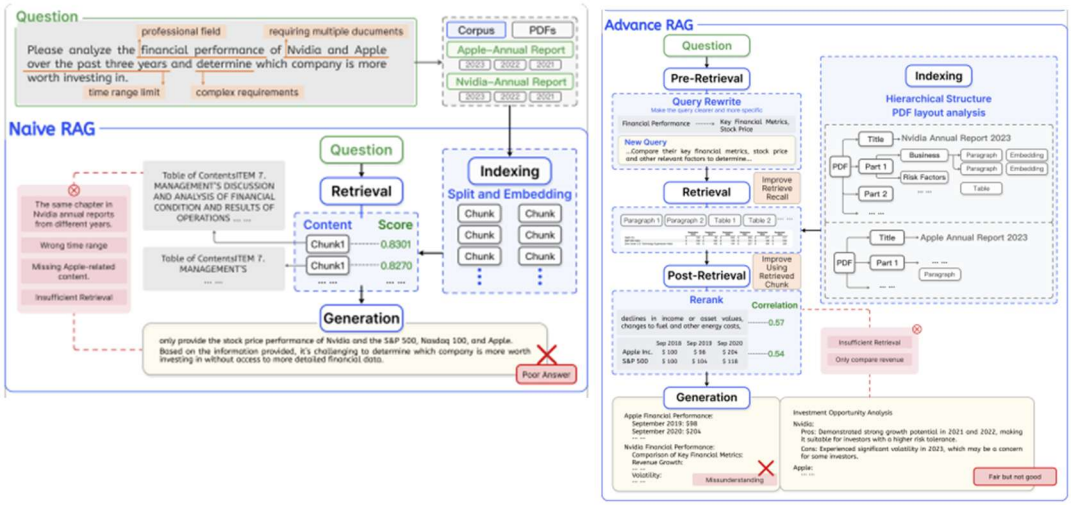

Figura 8 – Exemplo de resposta de dois modelos RAG, um Naïve e outros Advanced a uma pergunta complexa. Fonte: retirado do artigo (Yunfan Gao, 2024)

A partir da mesma "pergunta exemplo", utiliza-se o modelo Modular em que o processo não está mais restrito a uma sequência linear, mas sim controlado por múltiplos componentes de controle para recuperação e geração.

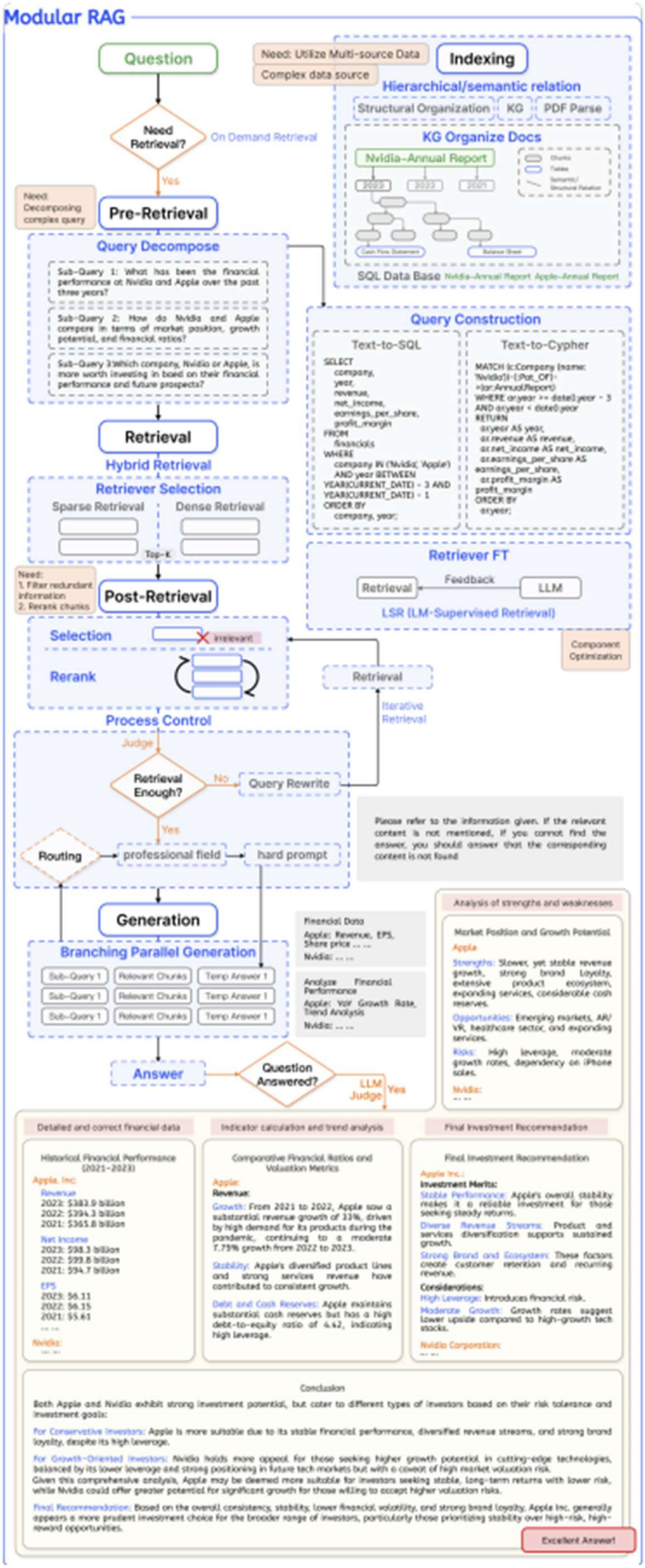

Figura 9 - Resposta usando o modelo Modular - Fonte: : retirado do artigo (Yunfan Gao, 2024)

3.4 Componentes do RAG

Independentemente do tipo de implementação que esteja definido, o modelo de recuperação e geração estará dividido em 4 principais domínios, que podem ser classificados da seguinte maneira:

- **Dados (*data*):** refere-se aos dados e sua origem. Devem ser confiáveis e suficientes e atender questões de privacidade, segurança e de direitos autorias
- **Armazenamento (*storage*):** compreensão de como os dados serão armazenados antes e após o processamento, assim como o volume
- **Recuperação (*retreival*):** como os dados serão recuperados para ser enviados para o modelo generativo em conjunto com o modelo (Naïve, Avançado ou Modular) a ser adotado.
- **Geração (*generation*):** estabelece o modelo de GenAI encaixa no modelo RAG escolhido.

Outro aspecto fundamental, antes da decisão de qual modelo será utilizado é importante estabelecer a proporção entre informação paramétrica (refere-se ao conhecimento que o modelo de linguagem já aprendeu durante seu treinamento – pré-treinada - implícita) e não-paramétrica (refere-se aos dados frescos e específicos que são trazidos pelos documentos atuais que o sistema de recuperação encontra - explícita).

Um fluxo genérico que resume os tópicos abordados até agora, e que descreve o papel de cada componente do RAG é apresentado a seguir (Rothman, 2024):

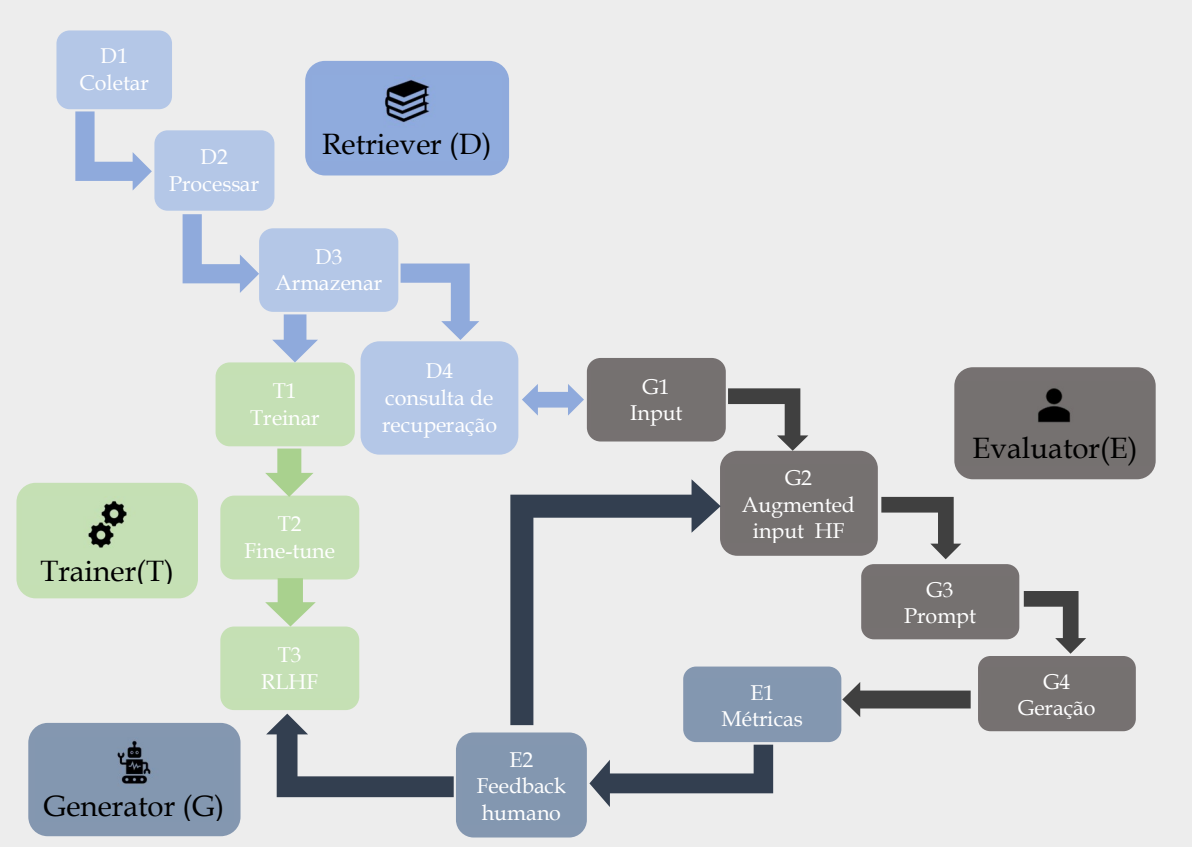

Figura 10- The RAG Framework and ecosystem. Fonte (Rothman, 2024)

- **Retriever (D)** é responsável pela coleta dos dados, armazenamento, processamento e pela recuperação, independente se esses são estruturados ou não.
- **Generator (G)** é responsável pelo input do usuário, pela a engenharia de prompt e a própria geração
- **Evaluator (E)** é responsável feedback, avaliação humano e pelas métricas de desempenho do modelo
- **Trainer (T)** é responsável pelo modelo inicial pré-treinado e pelo seu ajuste-fino

3.5 *Agentic* RAG

Na seção 2.1 deste trabalho (A IA Generativa - GenAI - em ambientes corporativos) destacamos que os agentes autônomos de IA estão sendo observados como a tendência de aplicação de ferramentas de IA em ambientes corporativos para os próximos anos. De fato, a chamada *Agentic RAG* é considerada um tipo de IA Autônoma e pode ser caracterizada, conforme o artigo pulicado na revista *Cognition, Technology & Work* (Allyson I. Hauptman, 2024), "*a IA Autônoma é caracterizada por agentes de inteligência artificial que são capazes de executar tarefas do início ao fim com mínima intervenção ou controle humano direto. Esses agentes operam com altos níveis de autonomia, o que significa que eles podem:*

- *Perceber o ambiente*
- *Tomar decisões*
- *Executar ações*
- *Adaptar-se a mudanças*
- *Aprender com os resultados*

Além disso, esses agentes deixam de ser apenas ferramentas e passam a atuar como membros ativos de equipes humanas, colaborando em ambientes complexos e intensivos em dados."

Uma outra definição encontrada está no estudo *Artificial Intelligence: Definition and Background* (Haroon Sheikh, 2023), que define de forma resumida como sendo "*sistemas que exibem comportamento inteligente ao analisar seu ambiente e tomar ações – com algum grau de autonomia – para alcançar objetivos específico*". Portanto, podemos entender melhor a autonomia desses sistemas tendo como ponto de partida a comparação entre as IAs generativas e os Agentes autônomos, no nosso caso específico o *Agentic RAG*.

De fato, sob uma perspectiva comparativa podemos entender a *GenAI* como um sistema fundamentalmente reativo, isto é, eles esperam que o usuário envie um *prompt* e a tarefa deles é produzir conteúdo baseado no que foi provido no *prompt*. As LLMs então usam os padrões que aprenderam e este aprendizado está fundamentado em conhecer as relações estatística entre palavras, pixels e ondas acústicas em bases de dados enormes. A partir daí geram textos, imagens e vídeos.

Agentic AI ao contrário, não é reativa, é proativa. Isso porque, assim como a *GenAI* o sistema recebe um *prompt* do usuário, mas ao contrário dela, este é usado para atingir metas através de uma série de ações. Faz isso seguindo uma espécie de ciclo de vida, como na figura a seguir:

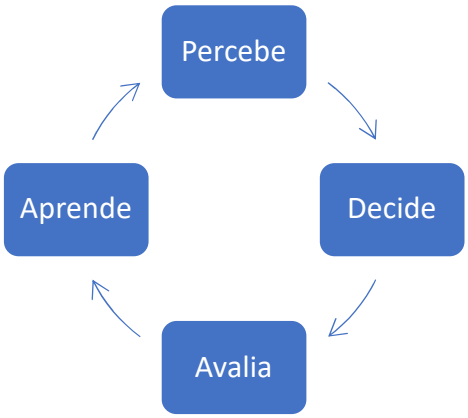

Figura 11 - Ciclo de vida da Agentic IA (IBM, 2025)

- **Percebe**: O agente coleta dados do ambiente por meio de sensores, APIs, bancos de dados ou entrada do usuário.
- **Decide**: Com base nos dados percebidos, o agente analisa a situação e escolhe uma ação ou plano apropriado.
- **Avalia**: Após executar baseando-se na decisão tomada, o agente avalia o resultado para verificar se a ação foi eficaz.
- **Aprende**: O agente atualiza seus modelos ou estratégias com base na avaliação, melhorando seu desempenho futuro.

É válido notar que ambas as abordagens (*GenAI* e *Agentic RAG*) tem fundamentalmente as LLMs como base.

Olhando para o futuro parece que as arquiteturas mais poderosas baseadas em IA, provavelmente não serão puramente generativas ou *agentics*, isto é, somente baseadas em agentes, serão uma mistura de ambas. Uma definição interessante seria, como proposto em (IBM, 2025), que estes sistemas poderão ser considerados colaboradores inteligentes.

A seguir, está proposto a implementação de um sistema baseado em RAG como tema de projeto desse trabalho, porém sem a inclusão de agentes. Isso porque até o momento da conclusão desse trabalho a empresa não dispunha ainda de forma bem definida, de uma solução de arquitetura para se trabalhar com agentes. Portanto, iremos tratar de um RAG que tem como base específica uma série de dados em arquivos pdf comparando-o com LLMs genéricas e disponíveis no mercado. No futuro, quando houver a definição dos padrões corporativos para adoção de agentes autônomos, pretende-se evoluir o projeto coma utilização de *Agentic* RAG. Espera-se que tais agentes possam ser usados para detalhar informações em bases de dados estruturadas ou buscar informações até mesmo fora da corporação.

4 O Projeto

4.1 Objetivo

O objetivo geral desse trabalho é utilizar tecnologia de *GenAI* em um contexto específico (apoio à tomada de decisão) nas empresas como ferramenta confiável, de fácil acesso,

ágil e segura. Com isso esperamos não somente proporcionar ao decisor uma alavanca tecnológica que faça parte do rol de ferramentas de seu dia-a-dia, mas também um instrumento de reconhecido valor em seu processo decisório. Para isso iremos sugerir inicialmente a implantação de um RAG em um contexto específico e mais simplificado, para que depois, possa evoluir para temas mais complexos também da área orçamentária.

O objetivo específico é criar um RAG capaz de responder textualmente, via *chat* a perguntas relacionadas especificamente à *CAPEX* (Capital Expediture)[8] e suas subdivisões em agrupamentos de contas orçamentárias específicas de uma empresa de Óleo & Gás. Esse *chatbot* deverá ser capaz de responder em linguagem natural e textual a perguntas feitas pelo usuário, que receberá as repostas baseadas em informações históricas provenientes de várias fontes de dados curadas e internas à empresa (estruturadas ou semiestruturadas).

4.2 Justificativa

Apesar de parecer contraditório, a grande difusão de ferramentas de análise de dados e o incremento significativo da capacidade de computação não parece ter resolvido um dos grandes problemas nas grandes corporações, que é a confiabilidade e acessibilidade de dados históricos. É comum ainda hoje que algumas demandas oriundas da alta administração não tenham resposta rápida, e as vezes as respostas se alongam além do tempo disponível para tomada de decisão. As origens desses problemas são muitas e vão desde dados armazenados em várias bases diferentes, mal governados, inacessíveis, perdidos nas evoluções e mudanças de plataformas, até a sobrecarga de informações. Com enumeras fontes de dados, os decisores precisam "garimpar" os dados que necessitam em muitos sistemas e por várias vezes se sentem inseguros em utilizar as informações disponíveis. Um comportamento muito comum é solicitar dados de fontes diferentes para confrontá-los e, quando não diferenças relevantes, se basear neles para decidir.

Dessa forma esse trabalho propõe diminuir a distância entre o dado e o decisor, não só através do aumento da confiança da alta administração nos canais digitais da empresa, como também através da geração de valor efetiva no processo decisório utilizando-se *framewroks* de uso livre e bem difundido na comunidade de desenvolvedores. Com isso vamos demonstrar que a chamada "Transformação Digital" tem grande potencial de impactar positivamente, não somente na eficiência dos processos rotineiros de uma empresa, como também na eficácia dos processos decisórios.

4.3 Descrição

Os padrões internos da corporação que descrevem como deve ser feito a construção de aplicações de IA, especificamente de RAG, definem desde a construção do experimento, a fim de servir como uma prova de conceito, até o momento de construir uma aplicação robusta em ambiente produtivo com as boas práticas e ferramentas corporativas.

Atualmente a empresa dispõe de uma arquitetura corporativa bem definida e padronizada. Existem três caminhos principais a serem seguidos para a criação da aplicação esta enumerada a seguir:

---

8 CAPEX – Capital expenditure - Wikipedia - https://en.wikipedia.org/wiki/Capital_expenditure

- Aplicação web corporativa baseada em prompt
- Experimentação local
- Ambiente produtivo

O escopo desse trabalho é o segundo da lista, experimentação local, para posterior prova de conceito para ser ou não disponibilizado em ambiente produtivo. Para fins os fins de experimentação os principais componentes a serem utilizados em um projeto RAG são:

- **Modelos de linguagem/LLMs**: Modelos presentes no Hub de Modelos (ex: GPT-4o, Claude Sonnet 3.5, Llama 3.3 etc);
- **Motor de Busca (banco local - Chroma, FAISS, etc)**:
- **Linguagem de programação recomendada**: Python;
- **Framework de orquestração**: LangChain/LangGraph (Python);
- **Frontend**: Streamlit (Python) ou React (JavaScript);

Para o projeto iremos utilizar uma aplicação que interage com usuário através de *prompts* disposto em interface web. Na construção a aplicação nos baseamos no projeto do professor orientador Pedro Gomes (https://github.com/phbgomes22/GISIA)[9], no qual adaptamos para atingir os resultados desse trabalho. Os componentes utilizados estão completamente aderentes ao proposto pela arquitetura interna da empresa e a experimentação foi feita localmente e está disponível no *Git* do autor (https://github.com/alexfettermann/MDT/tree/main/TCC).

O funcionamento básico da ferramenta consiste em implementar as 3 fases do RAG (Retrieval, Augmetation e Generation) para obter respostas mais relevantes e apuradas do que uma LLM que tenha sido pré-treinada com informações genéricas. A ferramenta extrai dinamicamente conhecimento em tempo real de fontes externas, como bancos de dados, APIs, documentos corporativos ou da web. Esse mecanismo de buscar as respostas em uma base de dados específica melhora significativamente a precisão, relevância e adaptabilidade da IA, já que utiliza informações contextuais relevantes que são recuperadas com base no input e fornecidas ao modelo como dados complementares e que contém as informações específicas do contexto. A figura 11 a seguir descreve a arquitetura básica com as etapas enumeradas e explicadas na sequência.

9 Até a publicação deste trabalho o brench mais atualizado utilizado como referência foi o IARIS - https://github.com/phbgomes22/GISIA/tree/iaris),

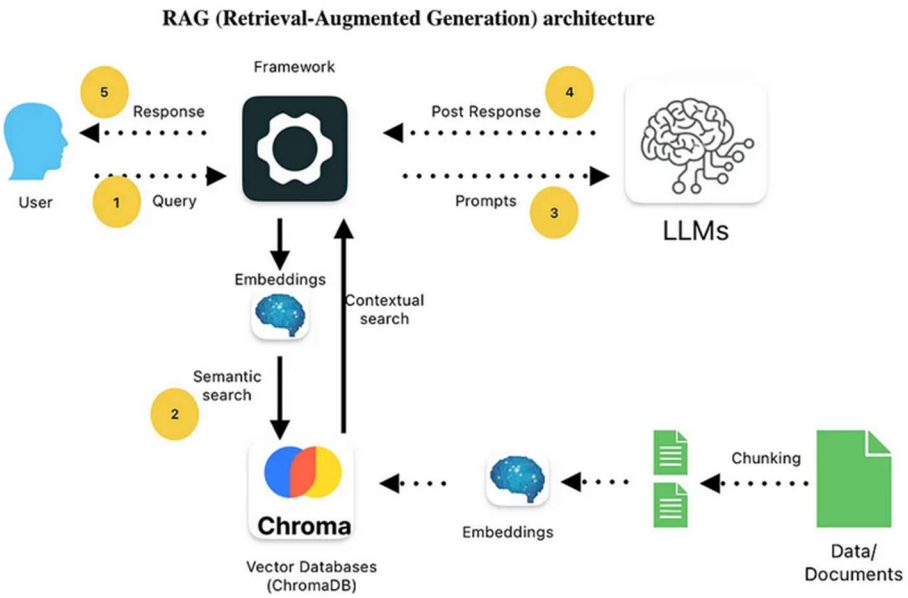

Figura 12 - Arquitetura básica do RAG implementado. Retirado de (@IBM, 2025)

1. Usuário envia uma pergunta em texto através da interface gráfica
2. O componente central da arquitetura do projeto, denominado na figura de *framewrok* é o *LangChain*[10] que tem como função principal o encadeamento de operações ou etapas de processamentos. Nessa etapa é responsável por buscar informações de contexto em uma base de dados vetorial local e especializada no assunto. Ao fazer isso, garante que a resposta seja baseada em dados atualizados e específicos da base local, o que gera confiabilidade. Essa ação, o envio do prompt com um contexto especializado, e oriundo de uma base curada pelo própria corporação, (podemos considerar como o "pulo-do-gato" do RAG) evita uma resposta desatualizadas do treinamento da LLM pré-treinada.
3. Aqui o *framework* compõe o contexto recebido no passo anterior em um *template* de *prompt* e envia para um modelo de linguagem, no projeto usamos o GTP 4o através de sua API, OpenAI API[11]. Note que esse passo atua como uma espécie de refinador
4. O LLM, (GPT 4o) envia a resposta para o *LangChain*
5. O *LangChain* então envia a resposta para o browser e faz isso utilizando uma biblioteca denominada *Streamlit*[12] que é muito utilizada para desenvolvimento de interfaces web

O processo de *embeding* representado no canto inferior direito da figura 11 não ocorre em nenhum dos momentos descritos acima. Este processo é fundamental para a criação do banco de dados vetorial e deve estar disponível antes da interação do usuário com o sistema. Para obter uma explicação detalhada do que é esse processo (um dos pilares da IA moderna) é sugerida uma visita aos seguintes conteúdos: (Harsoor, 2024) e (Google Developers). No entanto, de maneira resumida, podemos dizer que o processo de criação

10 Site oficial do LangChain - https://python.langchain.com/docs/introduction/

11 Quick start da OpenAI API no site oficial - https://platform.openai.com/docs/quickstart?api-mode=responses

12 Site oficial do Streamlit - https://streamlit.io/

de entradas na base dados vetorial, (no nosso caso o *Chroma*[13]) a partir de documentos (no nosso caso PDFs) é geralmente conhecido como *document embeding*. O processo consiste em converter o conteúdo dos documentos em vetores numéricos que traduzem, ou melhor, capturam o sentido semântico do texto nesses documentos. Esses vetores então ficam armazenados no banco de dados vetorial a fim de prover buscas semânticas eficientes por similaridade de conteúdo (Databricks). O resultado dessas buscas é que compõe o contexto que é enviado juntamente com o prompt, para o LLM (item 3 da figura 11).

Já o processo denominado *chunking* também faz parte da geração das entradas no banco de dados vetorial. A função do processo otimizar o sentido semântico dos itens dos documentos. Para isso, não basta simplesmente fatiar grandes documentos, mas separá-los em blocos que mantêm sentido próprio e coerente, como parágrafos.

Os documentos que são origem de base de dados vetorial do projeto são dois resumos de dois planos de negócio da Petrobras (2024-2028 e 2025-2029) disponíveis em dois arquivos pdf públicos. Obviamente, em um ambiente empresarial esses documentos podem ser substituídos por outros internos com maior nível de sigilo, o que possibilita aumentar o nível de privacidade e rastreabilidade do conteúdo usado como fonte da informação, comparativamente com LLMs pré-treinadas.

A seguir estão listados os principais componentes e recursos utilizados no projeto:

- **Modelos de linguagem/LLMs**: GPT-4o acessado via API
- **Motor de Busca:** Chroma 1.0.13
- **Linguagem de programação**: Python 3.13.5;
- **Framework de orquestração**: LangChain (Python);
- **Frontend**: Streamlit 1.46.0 (Python);

5 Coleta de Dados

5.1 Avaliação da performance do RAG frente modelos LLM genéricos

O ponto crucial desse trabalho é investigar de forma objetiva o quão eficaz um RAG é em relação aos LLMs genéricos e disponíveis no mercado. Ao falarmos em eficácia não estamos medindo a dimensão financeira de manutenção desses modelos, mas sim da capacidade de responder de maneira objetiva e correta a perguntas feitas à ferramenta de IA.

5.2 Metodologia

Para a avaliação foram criadas 18 (dezoito) perguntas, sendo 9 (nove) delas de conteúdo público e divulgado na internet e outras 9 (nove) cujo conteúdo é privado e não disponível publicamente na internet, portanto somente acessíveis pela base acessada pelo modelo RAG implantado no projeto. Além disso, a cada uma das perguntas

13 Site oficial de documentação do Chroma - https://docs.trychroma.com/docs/overview/introduction

também foi atribuída um nível de complexidade: Alta, Média e Baixa. A avaliação quanto a cada uma delas obedeceu aos seguintes critérios, simples:

- **Baixa:** Pergunta com resposta direta em um texto disponível no documento público e/ou no privado; Por exemplo: *Qual o investimento total da Petrobras no plano de negócios 2024-2028?*
- **Média:** Pergunta com resposta indireta em um texto disponível no documento público e/ou no privado. Nesse caso, uma pergunta em que um termo deveria ser "entendido" pelo modelo para se chegar a resposta. O exemplo usado foi o termo CAPEX para se referir a investimento. Por exemplo, ao invés de perguntar: *Qual o investimento total do Segmento E&P no plano 2024-2025 para o segmento E&P?*, substituímos por, *Qual o CAPEX total do Segmento E&P no plano 2024-2025 para o segmento E&P?*;
- **Alta:** Pergunta que necessita de uma análise de mais de um documento e que a resposta dependa de uma análise simples, mas que necessite da capacidade de agrupar termos correlatos e compará-los através de divisões simples após uma ordenação dos resultados. Por exemplo: *Qual o segmento de negócio da Petrobras teve a maior variação de CAPEX entre os dois últimos planos de negócio aprovados?;*

Para fins de classificação de perguntas, criou-se então uma série de nomenclaturas que podem ser distribuídas em uma tabela simples, como a seguir:

	Baixa	**Média**	**Alta**
Pública	$PBB_{1,2,3}$	$PBM_{1,2,3}$	$PBA_{1,2,3}$
Privada	$PVB_{1,2,3}$	$PVM_{1,2,3}$	$PVA_{1,2,3}$

Tabela 2 - Classificação das perguntas quanto à complexidade e privacidade

Essa classificação irá nos permitir avaliar a qualidade das respostas de cada um dos modelos avaliados quanto à complexidade e a privacidade. A avaliação foi simples, como sendo resposta satisfatória (S) ou Insatisfatória(I) [14]. Entende-se por satisfatória uma resposta objetiva e correta. Portanto, resposta como "Não sei" ou erradas foram consideradas como insatisfatórias. Obviamente, para os testes, as respostas corretas já eram conhecidas de antemão. Outro premissa fundamental é que para as perguntas privadas, apenas os pdf que foram carregados no modelo RAG continham as respostas.

5.3 Modelos avaliados

As 18 (dezoito) perguntas elaboradas foram disparadas contra 3 (três) modelos:

- O RAG implementado como descrito na seção Projeto desse trabalho.
- Chat GPT 4o mini, gratuito e disponível no site da Open AI ou API
- Chat GPT 4o, gratuito porém de uso limitado em número de *tokens* por dia, disponível no site da Open AI ou API.

14 Uma evolução do estudo seria classificar as respostas insatisfatórias de pelo menos duas maneiras: alucinação ou negativa, como por exemplo, Não sei.

Por motivos de simplificação não foram comparados fornecedores diferentes e a escolha da OpenAI se deveu não somente pela disponibilidade do modelo atualmente implantado na empresa, mas também pela facilidade de utilizar o modelo através de uma API simples. Para maiores informações de que como gerar uma chave de acesso para API do OpenAI, uma referência simples que pode ser consultada é (GeekToGeek)

Além da OpenAI existem vários outros fornecedores de modelos. A seguir estão dispostos em uma tabela os principais deles:

Modelo	**Fornecedor**	**Acesso via**
GPT-4o	OpenAI	Chatbot e API
o3 e 1	OpenAI	Chatbot e API
Gemini	Google	Chatbot e API
Gemma	Google	Open
Llama	Meta	Chatbot e aberto
R1	DeepSeek	Chatbot, API, e aberto
V3	DeepSeek	Chatbot, API, e aberto
Claude	Anthropic	Chatbot e API
Command	Cohere	API
Nova	Amazon	API
Mistral	Mistral AI	API
Qwen	Alibaba Cloud	Chatbot, API, e aberto
Phi	Microsoft	Open
Grok	xAI	Chatbot e aberto

Tabela 3 - Principais fornecedores de LLMs atualmente

5.4 Resultados obtidos

A seguir apresentamos os resultados obtidos que estão dispostos em uma tabela cujas linhas representam a categoria das perguntas (foram 3 perguntas por categoria), as colunas contém os modelos e as células o resultado de cada uma das respostas (tabela 4): S (satisfatório) ou I (Insatisfatório). As perguntas de caráter público também seguem listadas após a tabela de resultados, em uma tabela própria (tabela 5).

	GPT 4o mini	**GPT 4o**	**RAG**
PBB	I,I,I	I,S,I	S,S,S
PBM	I,I,I	I,I,I	S,S,S
PBA	I,I,I	I,I,I	I,I,I
PVB	I,I,I	I,I,I	S,S,S
PVM	I,I,I	I,I,I	I,S,S
PVA	I,I,I	I,I,I	I,I,I

Tabela 4 - Resultados das respostas dos modelos testados

Pergunta	Complexidade
Qual o valor total do investimento no plano 25-29 da Petrobras?	Baixa
Quantos sistemas de produção estão previstos para ser implantados no horizonte do plano?	Baixa
Qual é projeção pretendida da produção total de barris equivalentes de óleo e gás por dia?	Baixa
Qual o CAPEX total no plano 25-29 para o segmento E&P da Petrobras?	Média
Qual o CAPEX destinado para atividades exploratórias?	Média
Quantos poços estão previstos a ser perfurados no quinquênio no último plano de negócios aprovados pelo conselho Qual foi o segmento com maior variação, (positiva ou negativa) entre os palnos de negócio 2024-2028 e 2025-2029, aprovados pela Petrobras?	Média
Qual o tempo médio em dias de sonda para perfuração e completação de poços na camada pré-sal nos campos operados pela Petrobras?	Alta
Qual a participação do pré-sal em pontos percentuais na produção da Petrobras, no plano?	Alta
Qual o aumento do fluxo de caixa livre esperado no resultado da Petrobras por aumento de US$ 1 no preço médio do Brent, tendo como premissas o plano de negócios 2025-2029?	Alta

Tabela 5 - Lista de perguntas de conteúdo público

A seguir alguns exemplos de iteração do usuário com os prompts de usuário e de sistema destacados, assim como a resposta e a avaliação da resposta com a ferramenta desenvolvida pelo projeto:

- **Prompt do usuário:** *Qual o valor total do investimento no plano 25-29 da Petrobras?*
- **Prompt do sistema:** "*Olá! Fico feliz em receber sua pergunta sobre o investimento no plano 2025-2029 da Petrobras. Vamos esclarecer isso!*

  Neste plano, a Petrobras anunciou um investimento total de ***US$ 111 bilhões****. Esse capital será direcionado a diversas áreas estratégicas, focando no aumento da produção de petróleo e gás, além de investimentos em projetos sustentáveis e inovação tecnológica.*

  Detalhamento e Orientações

  Investimentos em Exploração e Produção: *Parte significativa está destinada a novos projetos de exploração em campos pré-sal, que são essenciais para aumentar a produção e garantir a competitividade da empresa.*

  Projetos Sustentáveis: *O plano também inclui iniciativas para reduzir a emissão de carbono e aumentar a eficiência energética, alinhando-se às demandas por um futuro mais sustentável.*

  Inovação: *Um foco em pesquisa e desenvolvimento, visando tecnologias que reduzam o impacto ambiental e melhorem a eficiência dos processos.*

  Resumo Conciso

A Petrobras investirá US$ 111 bilhões no plano 2025-2029, com foco em exploração, inovações e projetos sustentáveis.

Fonte: *Petrobras. "Petrobras lança Plano de Negócios 2025-2029 com investimentos de US$ 111 bilhões." Postado em 21/11/2024.*

Se precisar de mais detalhes sobre como esses investimentos podem impactar o setor ou outros aspectos desse plano, estou à disposição! O que mais você gostaria de saber?

- **Resultado:** Foi considerado Satisfatório pois a resposta está correta e totalmente alinhada com o conteúdo da fonte.

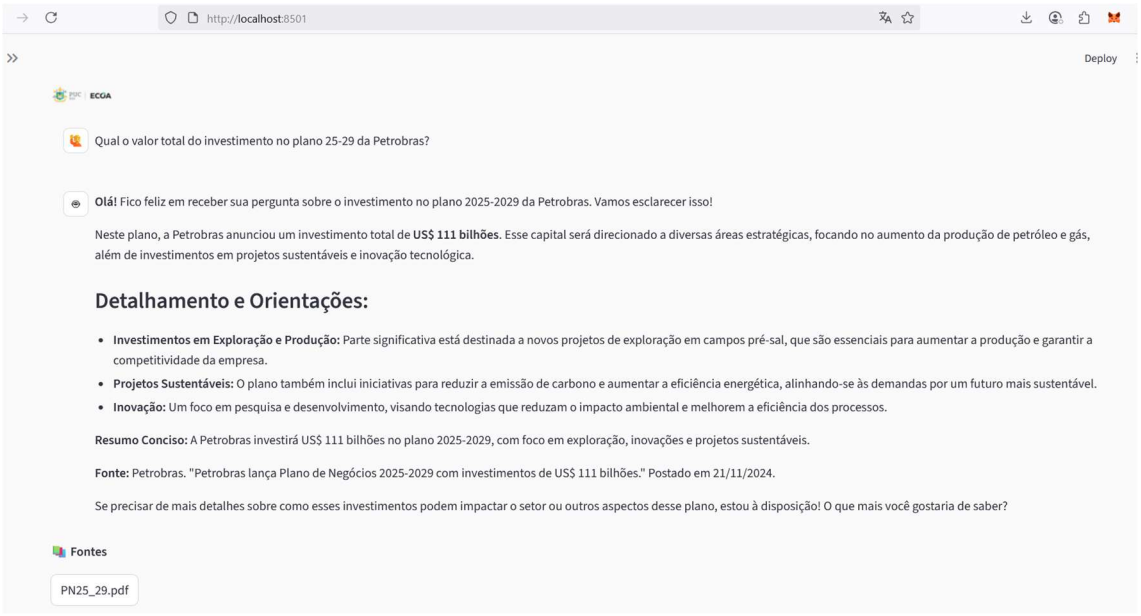

Figura 13 - Tela do sistema com iteração de PBB_1

- **Prompt do usuário:** *Qual o CAPEX destinado para atividades exploratórias?*
- **Prompt do sistema:** *Olá! Fico feliz em ajudar com sua pergunta sobre o CAPEX destinado a atividades exploratórias. Vamos esclarecer esse assunto!*

  Explicação Clara e Adaptada

  O CAPEX (Capital Expenditure) para atividades de exploração e produção, especificamente em segmentos como o do petróleo, pode ser um indicativo importante de como uma empresa se posiciona para atender à demanda e à competitividade. No caso que você mencionou, os investimentos totais para o quinquênio previsto somam aproximadamente US$ 77,3 bilhões. Desses, cerca de 60% são destinados para ativos do pré-sal, que é uma prioridade estratégica.

  Detalhamento com Exemplos e Orientações

  Aqui estão alguns pontos relevantes sobre esse CAPEX e suas implicações:

  Foco no Pré-sal: *O investimento no pré-sal visa maximizar a produção de petróleo de alta qualidade com custos menores. A área do pré-sal é conhecida por suas reservas significativas e por possibilitar a extração com menor impacto ambiental.*

  Emissões Reduzidas: *A estratégia de investimentos também busca cumprir metas de sustentabilidade, reduzindo as emissões de gases de efeito estufa durante a exploração.*

Comparação de Planos: *O plano atual representa um aumento de 5% em relação ao planejamento anterior, o que pode indicar uma busca por inovação ou adaptação às novas exigências do mercado.*

Exemplos de Uso de CAPEX: *Empresas bem-sucedidas usam CAPEX de forma estratégica, como na implementação de tecnologias que melhoram a eficiência operacional e reduzem custos a longo prazo.*

Resumo Conciso

O CAPEX total previsto é de US$ 77,3 bilhões, com 60% alocado no pré-sal, focando em produção de petróleo de qualidade, menores custos e redução de emissões.

Fonte

A informação é baseada na análise de investimentos prevista para o setor de exploração e produção, onde cerca de 60% do CAPEX é direcionado para ativos do pré-sal.

Estou à disposição para discutir mais sobre essa estratégia ou se você tem alguma outra dúvida relacionada ao seu contexto. Que tal falarmos sobre como isso se integra na sua gestão ou nos objetivos de impacto da sua organização?"

- **Resultado:** Foi considerado Satisfatório pois a resposta está correta e totalmente alinhada com o conteúdo da fonte.

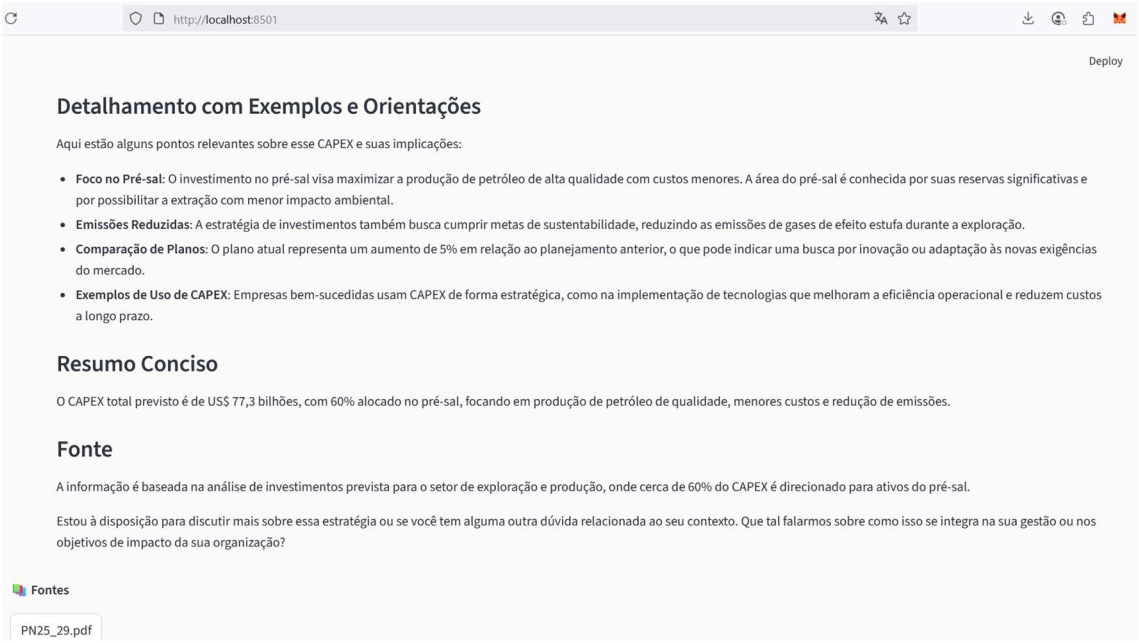

Figura 14 - Tela do sistema com iteração de PBM_2

- **Prompt do usuário:** *Qual a participação do pré-sal em pontos percentuais na produção da Petrobras, no plano?*
- **Prompt do sistema:** *Oi! Entendo que você esteja buscando informações sobre a participação do pré-sal na produção da Petrobras. Infelizmente, não tenho acesso a documentos específicos ou informações recentes que não estejam em meu treinamento até outubro de 2023. No entanto, posso oferecer uma visão geral sobre a importância do pré-sal para a Petrobras e suas implicações socioambientais.*

Participação do Pré-Sal na Produção da Petrobras

Relevância do Pré-Sal: *O pré-sal é uma das áreas mais ricas em reservas de petróleo no Brasil, e sua produção tem crescido significativamente nos últimos anos. Isso é crucial, pois a Petrobras tem investido pesado nesse segmento para aumentar sua produção e atender ao mercado interno e internacional.*

Dados Gerais: *A participação do pré-sal na produção total da Petrobras vinha se aproximando de 70%, conforme relatórios anteriores, refletindo a importância estratégica dessas reservas para a empresa. Os números exatos podem variar ao longo do tempo e dependem de uma série de fatores, incluindo investimentos em tecnologia e a capacidade de perfuração.*

Impacto Socioambiental: *Esse aumento na produção também levanta questões importantes sobre o impacto socioambiental. Por exemplo:*

Emissões de CO2: *A produção de petróleo tem uma contribuição significativa para as emissões de gases de efeito estufa. É fundamental que a Petrobras alinhase sua operação a metas de sustentabilidade.*

Investimentos em Tecnologias Limpas: *A empresa tem buscado investir em tecnologias que minimizem os impactos e ampliem a sustentabilidade.*

Resumo Conciso

O pré-sal tem uma participação significativa na produção da Petrobras, chegando a quase 70%. Essa área é crucial para o crescimento da empresa, mas também exige atenção às questões socioambientais, como emissão de gases.

Fonte

Dados gerais sobre a participação do pré-sal podem ser encontrados em relatórios anuais da Petrobras e em análises do setor de petróleo e gás.

Estou à disposição para ajudar com outras questões relacionadas a impacto socioambiental ou a gestão da sua organização. Alguma outra dúvida ou área específica que você gostaria de explorar?

- **Resultado:** Foi considerado Insatisfatório pois não foi encontrada uma resposta objetiva para a pergunta. Mesmo que o sistema tenha respondido que não tinha uma resposta para esse questionamento (o que está correto), sob o ponto de vista do usuário essa resposta é insatisfatória, e não está errada. De fato, nenhum dois dois *pdfs* tem essa informação, mesmo que indiretamente. Portanto uma resposta realmente não era esperada.

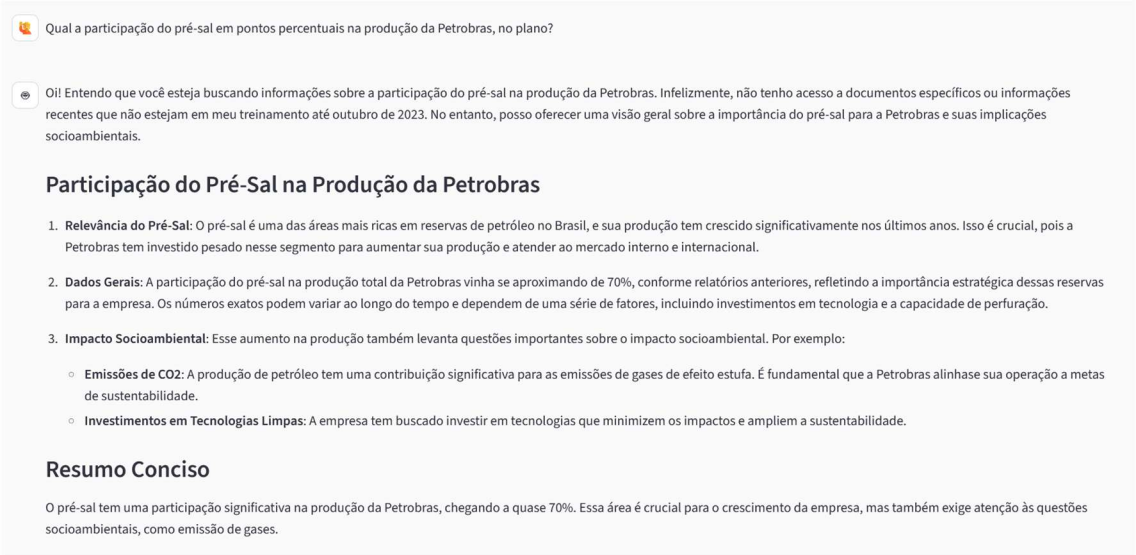

Figura 15 - Tela do sistema com iteração de PBA_1

Os dois *pdfs* utilizados como podem ser encontrados em:

- **Plano de negócios 2025-2030:** *Petrobras esclarece notícia sobre Plano de Negócios 2025-2029* - https://agencia.petrobras.com.br/w/negocio/petrobras-esclarece-noticia-sobre-plano-de-negocios-2025-2029
- **Plano Estratégico 2024-2028+:** *Novo Plano Estratégico prevê criação de 280 mil empregos por ano* - https://agencia.petrobras.com.br/w/negocio/novo-plano-estrategico-preve-criacao-de-280-mil-empregos-por-ano

As mesmas iterações utilizando o GPT 4o com dados públicos treinados sem os *pdfs* Todas as respostas foram insatisfatórias.

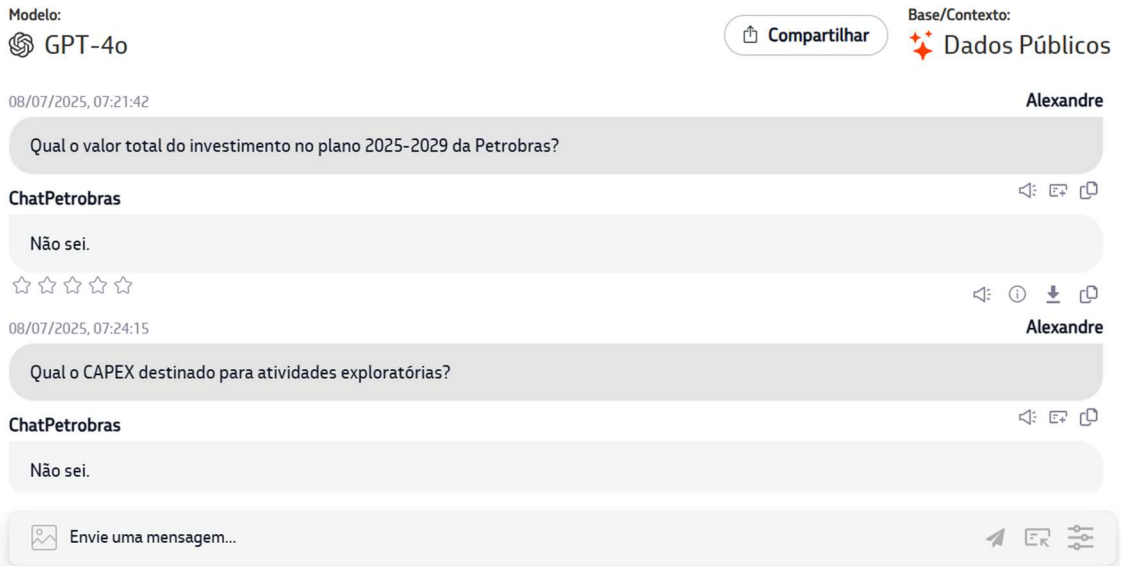

Figura 16 - Resposta do GPT 4o para PBB_1 e PBM_2

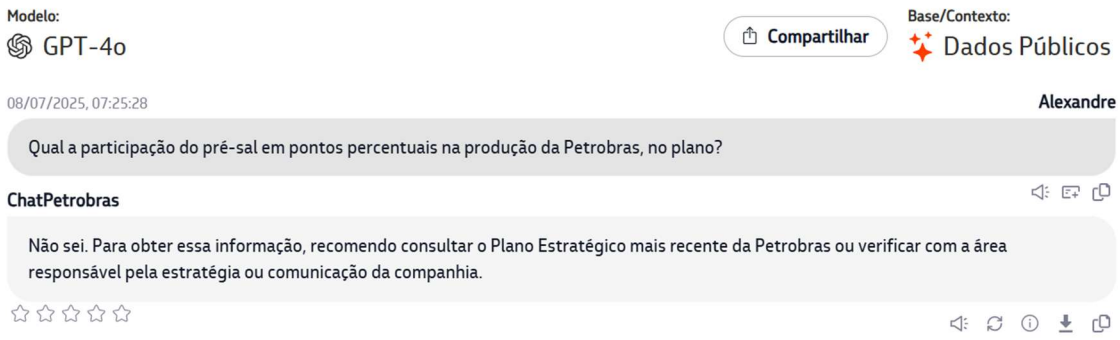

Figura 17 - Resposta do GPT 4o para PBA_1

As respostas como GTP-4o mini foram idênticas ao do GPT 4o.

5.5 Análise dos resultados

A partir das respostas obtidas e expostas na tabela 4, pode-se concluir que o RAG foi o único modelo capaz de responder de maneira satisfatória às perguntas cujas informações dependiam de dados privados. Mesmo as mais simples. Vale ressaltar aqui que o modelo GPT 4o para as perguntas que não conseguia resposta ainda tentava elaborar um plano de busca para a informação para o usuário, mas ainda assim não pareceu convincente nas respostas.

Um outro aspecto que merece destaque foi o comportamento do sistema nas perguntas que foram consideradas de alta complexidade, (PBA e PVA). Nessas duas categorias, entendeu-se que para se obter a resposta seria necessário a busca de informações em outras base de informação já que nenhum dos *pdfs*, mesmo os privados, dispunham, mesmo que indiretamente (através de cálculos) desses dados. A dita alta complexidade, então poderia ser reduzida se dispuséssemos de algum *pdf* com esses dados, ou se o sistema tivesse acesso de maneira mais ampla, a uma base de dados estruturada com essas informações.

É justamente nesse ponto que se percebe a possibilidade de, no futuro, se explorar o uso de agentes que sejam capazes de buscar em outros repositórios de dados tais informações. Essa possibilidade traz uma excelente oportunidade para a aplicação tendo em vista que a empresa dispõe de bases históricas robustas e confiáveis com esses dados. De fato, ao se conectar a repositórios especializados em disciplinas que estão envolvidas na geração dessas informações, como Poços, Submarina, Engenharia, Segurança e etc..., estaremos habilitando o sistema a responder detalhadamente às perguntas, tendo como subsídio bases de dados curadas.

Portanto, me parece que naturalmente o RAG é um caminho fundamental e natural para a maturidade na utilização de IA nas corporações, mas não é o fim. A evolução natural dessa trilha passa pela aplicação de agentes autônomos (*Agentic RAG*) .

6 CONSIDERAÇÕES FINAIS

Este trabalho propõe a adoção da tecnologia de Inteligência Artificial conhecida como RAG (Retrieval-Augmented Generation) como uma alavanca tecnológica relevante para a geração de valor nas organizações. A proposta é discutir não apenas as oportunidades, mas também os desafios que precisam ser superados para que esse valor seja efetivamente percebido pelos tomadores de decisão corporativos.

Na introdução, destaca-se que a fase inicial (o chamado "boom" da IA nas empresas) parece estar chegando ao fim. Superamos o momento de experimentação, no qual diversas possibilidades foram testadas com foco principal na eficiência de processos. Ferramentas de atendimento via chat, por exemplo, tornaram-se comuns, assim como o uso do termo "robotização" para descrever processos antes manuais que passaram a ser automatizados. Embora essas soluções pareçam recentes, tratam-se de ideias antigas que agora se tornaram mais acessíveis e difundidas, evidenciando o valor que carregam.

Oportunidades surgem justamente nesse ponto. Vivemos em um mundo hiper conectado e intensivo na geração de dados, onde a implantação de modelos de IA tornou-se significativamente mais importante e viável— não apenas pela evolução de hardware e software, mas também pela necessidade de se interpretar e dar sentido aos dados. Por isso os ativos digitais tornaram-se, talvez, os ativos mais valiosos da atualidade, pois hoje somos capazes de capturar os dados em detalhes, armazená-los em grande escala e processá-los com eficiência. Isso os torna essenciais para a tomada de decisão em todos os níveis organizacionais.

Como discutido no referencial teórico, os LLMs (Large Language Models) são ferramentas poderosas no processamento de linguagem natural. A arquitetura de *transformers*, com sua capacidade de processamento paralelo, permitiu um salto expressivo na exploração dos aspectos semânticos dos dados — ou seja, seu significado em contexto. É como se estivéssemos humanizando os sistemas, conferindo-lhes a capacidade de compreender perguntas e respondê-las com base em vastos repositórios de conhecimento.

No ambiente corporativo, isso pode ser traduzido em ferramentas capazes de interpretar perguntas em linguagem natural e respondê-las em tempo real com confiança (substituindo longas queries em SQL ou planilhas complexas em Excel). A oportunidade está justamente aí: o RAG combina o poder de processamento dos LLMs com a capacidade de refinar o entendimento semântico por meio de bases vetoriais especializadas. O resultado é um sistema que entende, busca, "pensa" e responde de forma precisa e ágil.

Por outro lado, como em toda revolução tecnológica, surgem também desafios. A IA não é exceção. Problemas antigos são potencializados, e a confiança nos dados e a privacidade tornam-se ainda mais críticas. Questões como o elevado consumo de energia para processar modelos e não somente custos altos decorrentes desse consumo, mas também os impactos ambientais decorrentes, se tornaram questões importantes a serem resolvidas. Além das questões ambientais existem problemas éticos e legais, como por exemplo geração de resposta com vieses. Como os modelos são treinados com uma enorme quantidade de dados, isso os torna suscetíveis à reprodução de vieses presentes nesses dados. Assim, um dos principais desafios para a adoção do RAG nas empresas é a governança de dados. É fundamental estabelecer processos e métodos que garantam dados corretos, livres de vieses, atualizados e entregues com segurança a quem de direito.

Conclui-se que o RAG representa uma evolução significativa no uso da *GenAI* nas corporações. Trata-se de uma ferramenta poderosa, com grande potencial para agregar valor aos processos decisórios. Naturalmente, essa tecnologia continuará evoluindo —

com a incorporação de agentes autônomos e o refinamento de sua arquitetura por meio de outras técnicas de Machine Learning. No entanto, os desafios continuarão recaindo sobre a confiabilidade, privacidade dos dados e custo.

Referências

@IBM, W. S. (21 de 02 de 2025). *Building a Local RAG-Based Chatbot Using ChromaDB, LangChain, and Streamlit and Ollama.* Fonte: Medium: https://medium.com/@Shamimw/building-a-local-rag-based-chatbot-using-chromadb-langchain-and-streamlit-and-ollama

Allyson I. Hauptman, B. G. (24 de 05 de 2024). Understanding the influence of AI autonomy on AI explainability levels in human-AI teams using a mixed methods approach. *Cognition, Technology & Work* , pp. Volume 26, pages 435–455.

Ashish Vaswani, N. S. (12 de 06 de 2017). *Attention is All you need.* Fonte: ArchiveX.org: https://arxiv.org/abs/1706.03762

Belcic, I. (11 de 11 de 2024). *What is a generative model?* Fonte: ibm.com: 2024

Bergman, D. (s.d.). *O que é aprendizado autossupervisionado?* Fonte: IBM Think: https://www.ibm.com/br-pt/think/topics/self-supervised-learning

Databricks. (s.d.). *Vector Database.* Fonte: Databricks: https://www.databricks.com/glossary/vector-database

Desvelar23. (25 de 10 de 2023). *Como os vieses entram na IA generativa?* Fonte: www.desvelar.org: https://desvelar.org/2023/10/25/como-vieses-entram-na-ia-generativa/

Doshi, K. (17 de 01 de 2021). *Transformers Explained Visually (Part 3): Multi-head Attention, deep dive*. Fonte: Towards Data Science: https://towardsdatascience.com/transformers-explained-visually-part-3-multi-head-attention-deep-dive-1c1ff1024853/

GeekToGeek. (03 de 04 de 2025). *vanishing and exploding gradients problems in deep learning*. Fonte: GeekToGeek: https://www.geeksforgeeks.org/vanishing-and-exploding-gradients-problems-in-deep-learning/

GeekToGeek. (s.d.). *How to Get an Open AI Key.* Fonte: GeekToGeek: https://www.howtogeek.com/885918/how-to-get-an-openai-api-key/

Google Developers. (s.d.). *Machine learning crash course.* Fonte: Google Developers: https://developers.google.com/machine-learning/crash-course/embeddings?hl=pt-br

Haroon Sheikh, C. P. (31 de 01 de 2023). Haroon Sheikh, Corien Prins & Erik Schrijvers. *Mission AI*, pp. 15-41.

Harsoor, S. (28 de 11 de 2024). *Embeddings: A Deep Dive from Basics to Advanced Concepts.* Fonte: Medium: https://medium.com/@sharanharsoor/embeddings-a-deep-dive-from-basics-to-advanced-concepts

HEIMES, H., SHIRALI, A., WOODCOCK, E., & GOSWAMI, S. (23 de Outubro de 2024). *Gen AI in corporate functions: Looking beyond efficiency gains.* Acesso em 21 de fevereiro de 2025, disponível em McKinsey & Company: https://www.mckinsey.com/capabilities/operations/our-insights/gen-ai-in-corporate-functions-looking-beyond-efficiency-gains#/

IBM. (30 de Abril de 2025). Youtube. *Generative vs Agentic AI: Shaping the Future of AI Collaboration*.

IBMAIHALL. (01 de 09 de 2023). *What are AI hallucinations?* . Fonte: www.ibm.com: https://www.ibm.com/think/topics/ai-hallucinations

Ilya Sutskever,Oriol Vinyals,Quoc V. Le. (14 de dezembro de 2014). *Sequence to Sequence Learning*. Fonte: https://arxiv.org/pdf/1409.3215

Jamil, U. (02 de 06 de 2023). *transformer-from-scratch-notes.* Fonte: GitHub: https://github.com/hkproj/transformer-from-scratch-notes/blob/main/Diagrams_V2.pdf

Jeff Pollard, J. B. (05 de 12 de 2023). *Security And Privacy Concerns Are The Biggest Barriers To Adopting Generative AI*. Fonte: Forrester: https://www.forrester.com/report/security-and-privacy-concerns-are-the-biggest-barriers-to-adopting/RES180179

Jr, N. F. (28 de 09 de 2024). *Porque Embeddings são a Base Fundamental das LLMs?* Fonte: Medium: https://nelsonfrugeri-tech.medium.com/porque-embeddings-s%C3%A3o-a-base-fundamental-das-llms-98041fe7121a

Khan, M. A. (15 de 05 de 2025). *Text Embedding Generation with Transformers*. Fonte: Machine Learning Mastery: https://machinelearningmastery.com/text-embedding-generation-with-transformers/

Litan, A. (24 de 12 de 2024). *AI Trust and AI Risk*. Fonte: Gartner: https://www.gartner.com/en/articles/ai-trust-and-ai-risk

Michael Chui, B. H. (30 de 08 de 2023). *O estado da inteligência artificial em 2023: o ano do crescimento explosivo da IA Generativa*. Fonte: Mckinsey & Company: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year/pt-BR

Microsoft. (21 de 05 de 2025). *O que é a IA Generativa?* Fonte: Microsoft: https://www.microsoft.com/pt-br/ai/ai-101/what-is-generative-ai
Olah, C. (27 de 08 de 2015). *Undertanding LSTM Networks.* Fonte: Research Google: https://research.google/pubs/understanding-lstm-networks/

Patrick Lewis, E. P.-t. (2021, abril 12). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. Retrieved from Arxiv.org: https://arxiv.org/pdf/2005.11401

Pytorch. (2025 de 05 de 02). *Pytorch*. Fonte: Pytorch: https://pytorch.org/

Raschika, S. (14 de 01 de 2024). *Understanding and Coding Self-attention*. Fonte: Ahead of AI: https://magazine.sebastianraschka.com/p/understanding-and-coding-self-attention

Rothman, D. (2024). *RAG - Driven generative AI.* Birmingham: <packt>.

SHACT, L., KREIT, B., VERT, G., HOLDOWSKY, J., & BUCKLEY, N. (25 de outubro de 2024). *Four futures of generative AI in the enterprise: Scenario planning for strategic resilience and adaptability.* Acesso em 21 de Março de 2025, disponível em Deloitte Center for Integrated Research: https://www2.deloitte.com/us/en/insights/topics/digital-transformation/generative-ai-and-the-future-enterprise.html

Stamford, C. (17 de 02 de 2025). *artner Predicts 40% of AI Data Breaches Will Arise from Cross-Border GenAI Misuse by 2027*. Fonte: Gartner: https://www.gartner.com/en/newsroom/press-releases/2025-02-17-gartner-predicts-forty-percent-of-ai-data-breaches-will-arise-from-cross-border-genai-misuse-by-2027

STAMFORD, Conn. (2024, Agosto 21). *Hype Cycle for Artificial Intelligence.* Retrieved from https://www.gartner.com/en/newsroom/press-releases/2024-08-21-gartner-2024-hype-cycle-for-emerging-technologies-highlights-developer-productivity-total-experience-ai-and-security

Walsh, D. (2023 de 08 de 2023). *The legal issues presented by generative AI*. Fonte: MIT Management Sloan School: https://www.gartner.com/en/articles/ai-trust-and-ai-risk

Wikipedia. (2025 de 05 de 02). *Back Propagation Through Time*. Fonte: Wikipedia: https://en.wikipedia.org/wiki/Backpropagation_through_time

Wikipedia. (07 de 05 de 2025). *Inteligência Artificial Generativa*. Fonte: www.wikipedia.com.br: https://pt.wikipedia.org/wiki/Intelig%C3%AAncia_artificial_generativa

Wikipedia. (09 de 05 de 2025). *Loss Function*. Fonte: Wikipedia: https://en.wikipedia.org/wiki/Loss_function

Wikipedia. (12 de 05 de 2025). *Regra da cadeia*. Fonte: Regra da Cadeia: https://pt.wikipedia.org/wiki/Regra_da_cadeia

Wikipedia. (12 de 05 de 2025). *Unupervised Learning*. Fonte: Wikipedia: https://en.wikipedia.org/wiki/Unsupervised_learning

Yunfan Gao, Y. X. (26 de julho de 2024). *Modular RAG: Transforming RAG Systems into.* Fonte: Arxiv.org: https://arxiv.org/pdf/2407.21059