



RAG e Few Shot

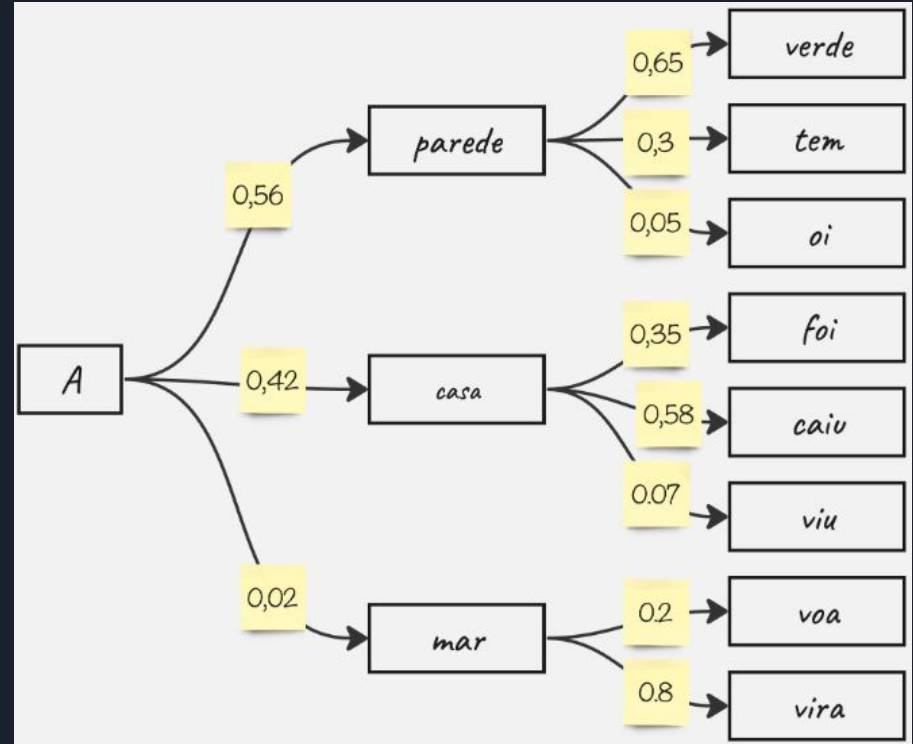
Técnicas de recuperação de informação e
engenharia de prompt

Integrantes:
Deivison Oliveira da Silva
Leonardo Côrtes Filho
Mateus de Almeida Souza

Visão geral

Inteligência Artificial em modelos de linguagem

- Prever os conjuntos de palavras
- Aprender sintaxe e semântica





Entender seus problemas

01

Large Language Model (LLM) como o Chat GPT são modelos generalistas e, justamente por serem treinados com uma ampla gama de dados, são muito bons em compreender sintaxe e semântica, porém o processo de treinamento dura meses

02

Claramente, o usuário irá fazer perguntas que não estão contidas na base de dados que o modelo foi treinado e, a depender de como foi treinado, o modelo não vai fornecer uma resposta coerente, ou não vai fornecer uma resposta

sucintamente, resuma o caso do influenciador brasileiro hytalo santo. Não pesquise na internet, use apenas o seu conhecimento.

Até a data do meu último treinamento, não há informações amplamente conhecidas ou registradas sobre um influenciador brasileiro chamado Hytalo Santo envolvido em um caso de grande repercussão pública. Pode se tratar de uma figura emergente ou de um caso recente que ainda não está consolidado nos dados que possuo.

Se quiser, posso buscar informações atualizadas para você. Deseja que eu faça isso?



sucintamente, qual a data de corte do seu último treinamento?

A data de corte do meu último treinamento é **junho de 2024**. Isso significa que meu conhecimento é baseado em informações disponíveis até esse mês, e não inclui eventos ou dados que surgiram após esse período — a menos que eu busque informações atualizadas na internet para você.

Retrieval augmented generation (RAG)

Retreinar o modelo toda vez que o usuário faz uma pergunta que a LLM não tem resposta é inviável (demora meses), portanto a saída é trazer informações para o modelo. A técnica mais usada para isso é o RAG

sucintamente, resuma para mim o caso do influenciador brasileiro Hytalo Santos em até 100 palavras. Inclua datas.

Aqui está um resumo sucinto do caso do influenciador Hytalo Santos em até 100 palavras, com as datas importantes:

Resumo (máximo 100 palavras):

Hytalo Santos, influenciador paraibano investigado desde novembro de 2024 pelo Ministério Público da Paraíba por exposição inadequada de adolescentes, foi alvo em 6 de agosto de 2025 de um vídeo denúncia do youtuber Felca sobre "adultização" de menores. Isso levou à suspensão de suas redes sociais em 8 e 12 de agosto e à desmonetização de seus conteúdos envolvendo menores [Wikipédia](#) [UOL](#) [UOL Notícias](#) [VEJA](#) [Agência Brasil](#) . Em 15 de agosto de 2025, Hytalo e seu marido foram presos em Carapicuíba (SP) por suspeitas de exploração sexual infantil e tráfico humano [Wikipédia](#) [Agência Brasil](#) [CNN Brasil](#) .

Se desejar, posso ampliar com mais detalhes ou contexto — é só avisar!

       Fontes

como você tem acesso atualizado a essas informações se a data de corte de seu treinamento é em 2024? Qual o nome formal da técnica que você usa? Responda sucintamente

Tenho acesso a essas informações porque, além do treinamento até 2024, uso ferramentas de busca em tempo real para consultar a web. O nome formal da técnica é "retrieval-augmented generation" (RAG).



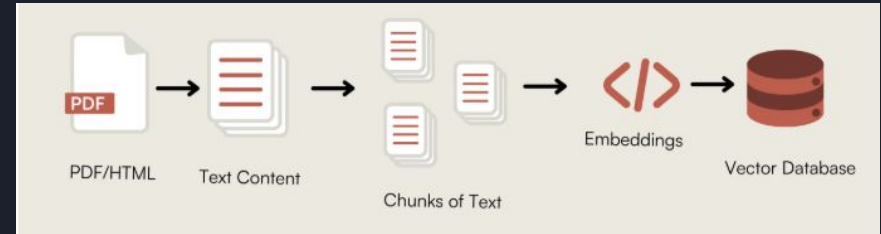
Retrieval augmented generation (RAG)

RAG é uma técnica de busca projetada para ser semântica, ou seja, ela não busca resultados apenas com base na semelhança de palavras entre os itens, mas também por meio da ideia que os itens querem transmitir. Logo, é relativamente capaz de lidar com sinônimos possibilitando retornar informações relevantes e contextualizadas. Enriquecendo MUITO as respostas.

- 01** Buscar por n informações
- 02** Transformar cada informação num vetor numérico (embedding)
- 03** Aplicar técnicas de semelhança de matrizes, por exemplo, similaridade de cosseno
- 04** Retornar os resultados mais bem avaliados como contexto para LLM

RAG e Chunk

Cada informação que o RAG retorna para para o modelo de linguagem, gasta recursos computacionais para serem processadas. Por isso, o RAG divide o trecho original em chunks (em partes) e, ao invés de retornar o trecho inteiro, retorna apenas os chunks



RAG e Chunk

Um chunk pode ser:

- Uma frase
- Um parágrafo
- Uma sequência de caracteres
- Qualquer forma de particionar o texto original

One of the most important things I didn't understand about the world when I was a child is the degree to which the returns for performance are superlinear.

Teachers and coaches implicitly told us the returns were linear. "You get out," I heard a thousand times, "what you put in." They meant well, but this is rarely true. If your product is only half as good as your competitor's, you don't get half as many customers. You get no customers, and you go out of business.

It's obviously true that the returns for performance are superlinear in business. Some think this is a flaw of capitalism, and that if we changed the rules it would stop being true. But superlinear returns for performance are a feature of the world, not an artifact of rules we've invented. We see the same pattern in fame, power, military victories, knowledge, and even benefit to humanity. In all of these,

Upload .txt

Splitter: Character Splitter

Chunk Size: 25

Chunk Overlap: 0

Total Characters: 905

Number of chunks: 37

Average chunk size: 24.5

One of the most important things I didn't understand about the world when I was a child is the degree to which the returns for performance are superlinear.

Teachers and coaches implicitly told us the returns were linear. "You get out," I heard a thousand times, "what you put in." They meant well, but this is rarely true. If your product is only half as good as your competitor's, you don't get half as many customers. You get no customers, and you go out of business.

It's obviously true that the returns for performance are superlinear in business. Some think this is a flaw of capitalism, and that if we changed the rules it would stop being true. But superlinear returns for performance are a feature of the world, not an artifact of rules we've invented. We see the same pattern in fame, power, military victories, knowledge, and even benefit to humanity. In all of these, the rich get richer. [1]

RAG e Chunk Overlap

Overlap = sobrepor

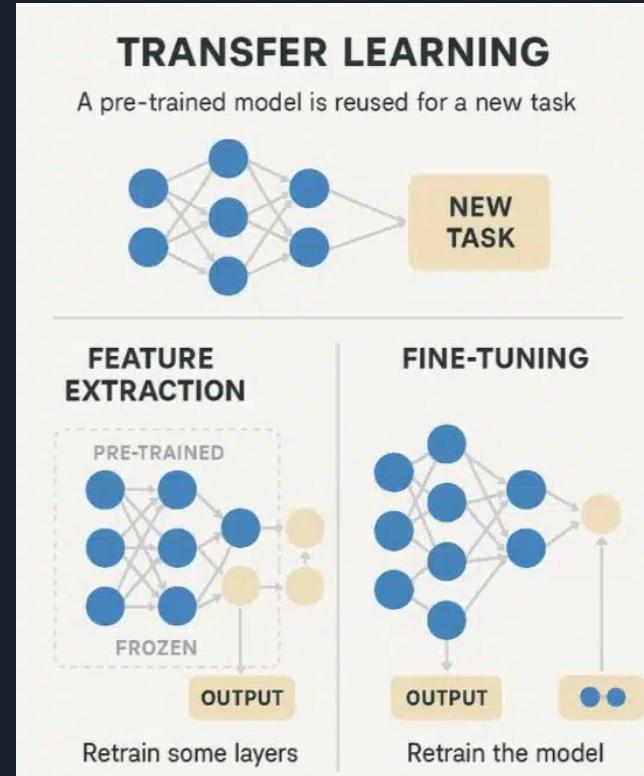
Chunk overlap = tamanho do trecho que dois chunks adjacentes irão compartilhar

A finalidade do chunk overlap é evitar a perda de contexto que pode ocorrer quando o texto é dividido em partes menores (chunks). Ao sobrepor uma pequena parte do final de um chunk com o início do próximo, o sistema evita cortar o texto abruptamente e preserva o contexto que une chunks adjacentes



Transfer Learning

- Definição
 - Extração de características
 - Fine Tuning
- Vantagens
 - Custos computacionais
 - Tamanho do conjunto dos dados
- Desafios
 - Ambas as tarefas são semelhantes





Fine-Tuning

O Fine-Tuning é destinado a modelos treinados ou pré-treinados. Ele modifica os parâmetros da rede neural pois ele treina o modelo novamente

- Por conta do modelo já está treinado (ou pré-treinado) o treinamento é mais rápido

Enquanto no RAG o modelo de linguagem TEMPORARIAMENTE importa dados de fontes externas para enriquecer a resposta fornecida ao usuário, o Fine-Tuning PERMANENTEMENTE agrega conhecimento ao modelo de linguagem pois ele ACRESCENTA informação aos parâmetros do modelo



Few shot

Few-shot é uma técnica em que se fornecem alguns exemplos de entrada e saída dentro do prompt para guiar o modelo a executar a tarefa desejada. Diferente do treinamento tradicional, o modelo não é ajustado nos pesos, ele apenas aproveita os exemplos cedidos pelo usuário como contexto temporário.

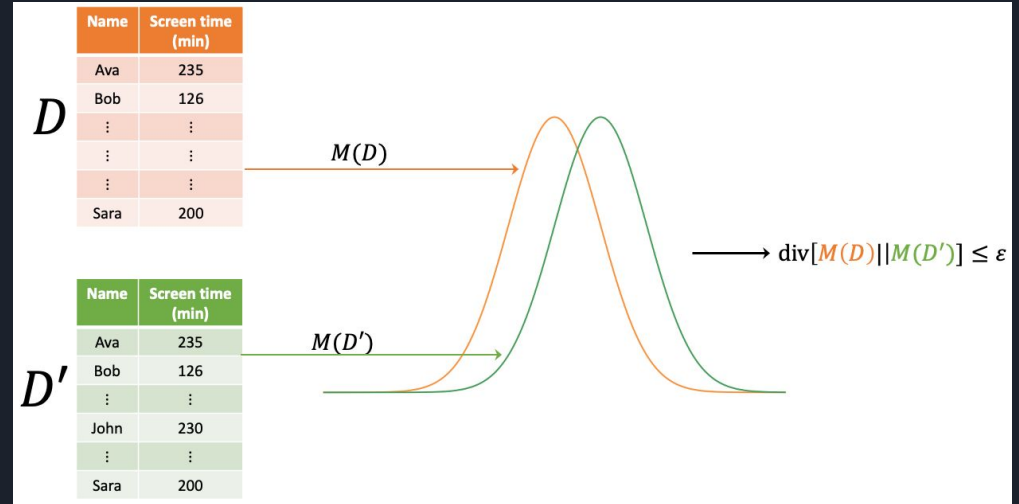


Few shot

Por exemplo, imagine que o modelo não saiba o que seja uma pitiaia. Usando few shot, um método de ajudar o modelo a corretamente identificar imagens de pitiaia é passar exemplos de imagens da fruta para o modelo, dessa forma, o modelo vai entender os padrões que identificam uma pitiaia. Posteriormente, você passa uma foto qualquer para o modelo e pergunta se, na imagem, há uma pitiaia.

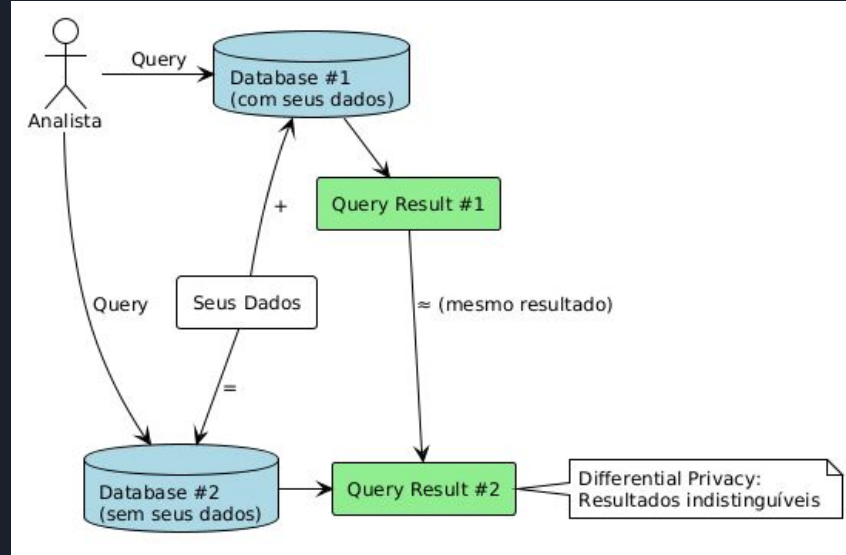
Differential Privacy

- $\Pr[M(D) \in O] \leq \exp(\epsilon) * \Pr[M(D') \in O] + \delta$
- Privacidade com garantias matemáticas
- Proteção contra inferências sobre indivíduos específicos nos dados



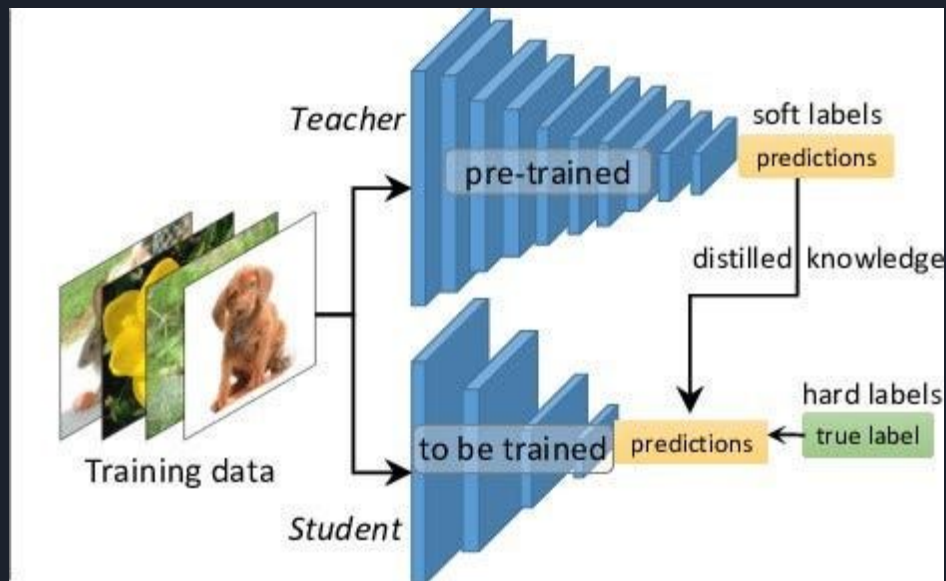
Differential Privacy

- Ruído controlado adicionado aos resultados das consultas
- Balanceamento entre utilidade e privacidade
- Calibração adequada



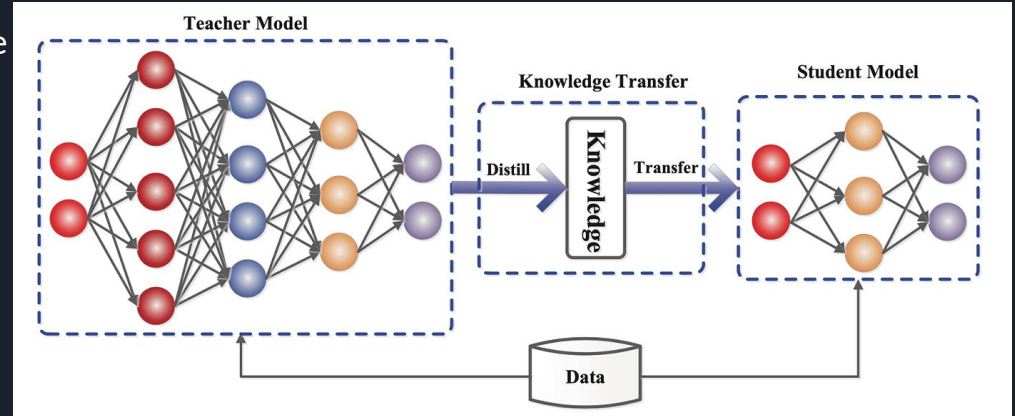
Knowledge Distillation de Modelos de Linguagem

- Transferência de conhecimento entre modelos
- Aprendizado através de "soft targets"
- Preservação da capacidade do modelo original



Knowledge Distillation de Modelos de Linguagem

- Destilação de representações internas e atenção
- Redução drástica de recursos computacionais
- Temperature scaling



Knowledge Distillation de Modelos de Linguagem

- Task-agnostic distillation
- Task-specific distillation
- Progressive distillation
- Multi-teacher distillation

