O que é Modelo de Linguagem (LLM - Large Language Model)?

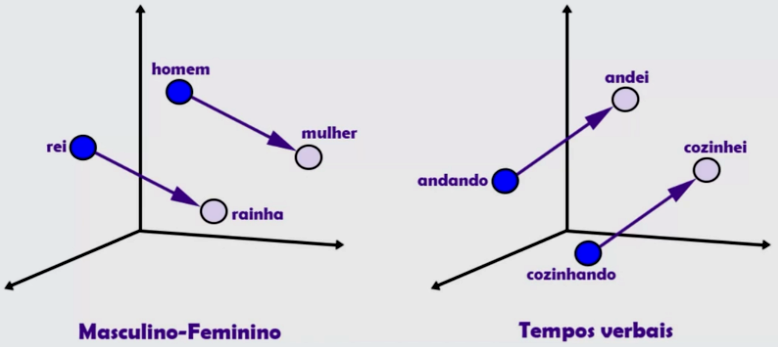
Aprender padrões da linguagem humana e relações entre as palavras.

*Word Embeddings*

Representações de palavras que compreendem a proximidade semântica entre elas.

As palavras são representadas por números.

E com isso ele faz vários cálculos d palavras aproximadas e por exemplo tenta prever quais são as próximas palavras.



Exemplo: vamos interagir com o chat GPT para a previsão da próxima palavra...

Vamos simular como funciona o ChatGPT. Para cada frase que eu escrever no prompt, você deve listar as 5 palavas com maior probabilidade que você usaria para completá-la, junto com a probabilidade de cada uma delas. Combinado? Apenas as palavras e probabilidades, sem mais nada. Me diga se você entendeu.

Será que o modelo de linguagem sempre escolhe a palavra com maior probabilidade?

Não!

Existe um conceito de temperatura, pra não deixar o modelo robótico demais.

Quanto mais alta a temperatura mais variadas são as respostas. Um temperatura baixa a tendência é ele sempre escolher a resposta mais provável.

*Tokens*

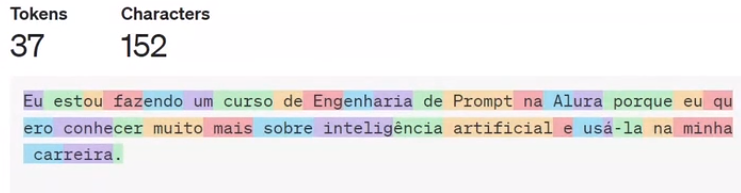
São unidades básica em modelos de linguagem.

Representam palavras ou subpalavras.

Exemplos:

infeliz = in + feliz = são dois tokens

https://platform.openai.com/tokenizer



--

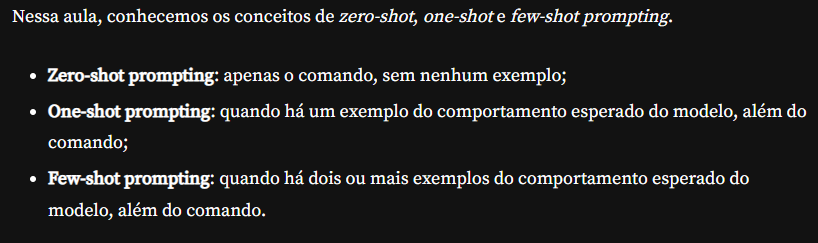
**Princípios da Engenharia de Prompt**

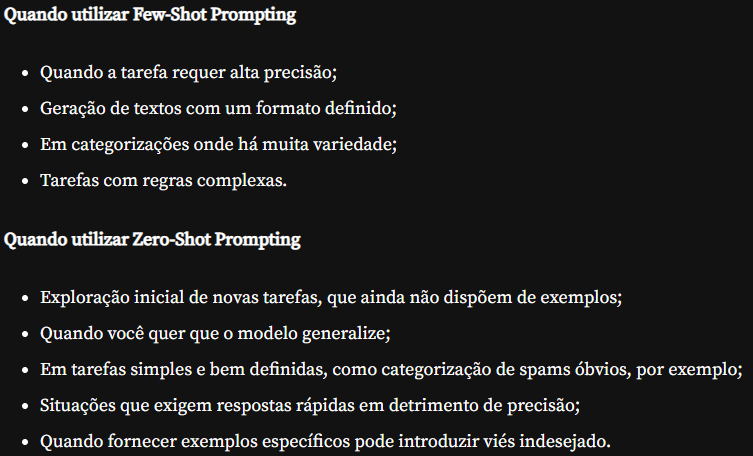
Como podemos engenhar ou projetar o nosso prompt da melhor maneira possível para obter o resultado almejado?

1. Ter clareza ao dar as instruções;
2. Dividir tarefas complexas em subtarefas menores;
3. Pedir para o modelo explicar seus passos antes de dar a resposta;
4. Pedir para o modelo dar justificativas de suas respostas;
5. Gerar várias respostas diferentes e pedir para o modelo escolher a melhor.

--

**Entendendo as técnicas**

****



--

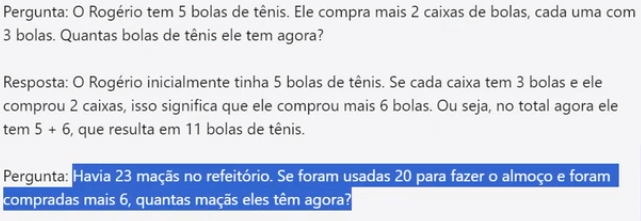
Entenda a técnica

Nessa aula, conhecemos a técnica Chain-of-thought, que pode ser traduzida como **cadeia de pensamentos**.

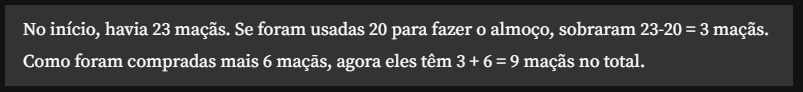
A cadeia de pensamentos se faz necessária em situações que demandam um pensamento lógico complexo ou abstrato, característicos de algumas áreas do conhecimento humano.

A técnica Chain of Thought nasceu com base na few-shot prompting. Ou seja, a cadeia de pensamentos era apresentada ao modelo na resposta de cada exemplo fornecido, e, ao se basear nos exemplos, o modelo replicava a lógica necessária para solução de problemas mais simbólicos.

*Neste exemplo:*



*Entenda que antes de fazer a pergunta, nós estamos ensinando, o que a gente quer. Resposta:*



*Isso se chama cadeia de pensamento, em inglês, chain-of-thought prompt*.

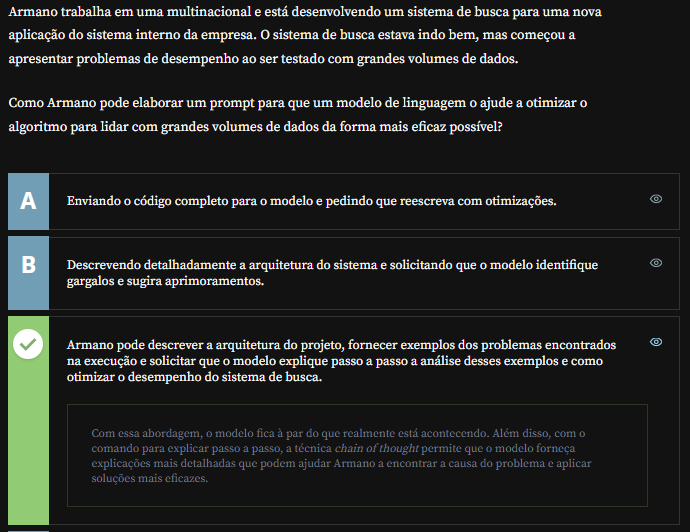
Essa técnica é chamada de Chain of Thought (Cadeia de Pensamento). Inclusive, existe um artigo publicado pelas equipes do Google Research e do Google Brain, atualmente conhecido como DeepMind. Você pode acessá-lo pelo seguinte link: [Chain-of-Thought Prompting Elicits Reasoning in Large Language Models](https://arxiv.org/abs/2201.11903" \t "_blank).

https://arxiv.org/abs/2201.11903

ensine a ela ter o seu tom de voz.

Não use o LLM como se fosse um martelo para resolver todos os meus problemas.

outro exemplo:



-- -- -- --

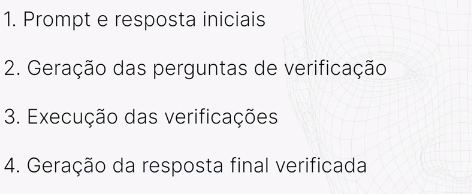
Least-to-Most prompting

quebrar em problemas

--

chain-of-verification

para verificar se a informação que o modelos te deu está correta, ou não.



--

Self-Consistency

Jargão: pode ser que as vezes a resposta seja alucinada, esteja incorreta.

Vc faz varias e vareias vezes a mesma pergunta