

MASTERCLASS

GEN AI



Engenharia de Prompt

- Prompts
- Tipos de prompts
- Técnicas de Engenharia de Prompts
- Práticas de Engenharia de Prompts
- Paper: "Unleashing the potential ..."



LLM

- LLM
- Desafios & Pesquisa
- Alucinação
- Processamento de um prompt dentro LLM
- Produtores de LLM



Projeto

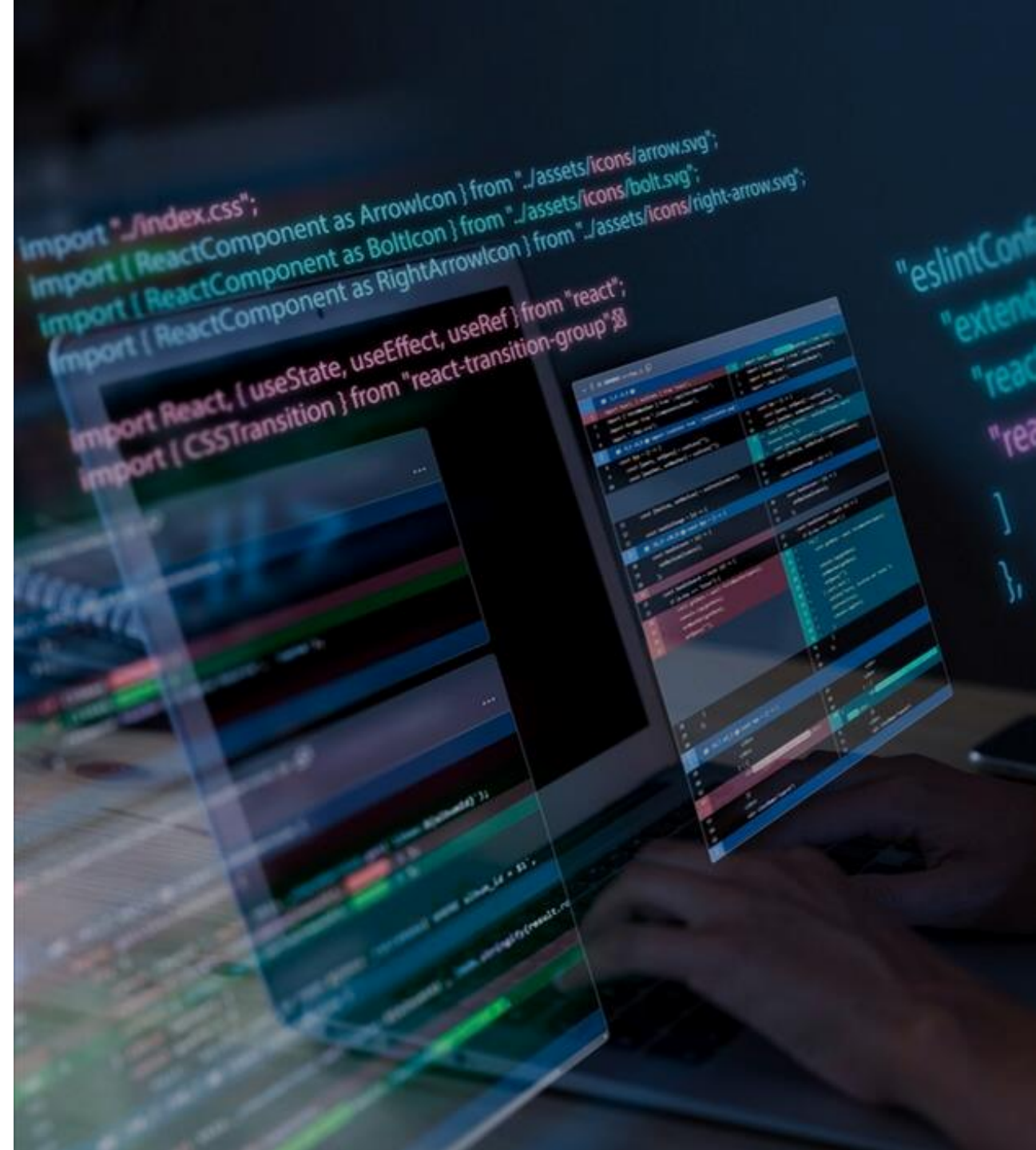
Loja física da **MAGALU**

Case de NLP:

- Análise de sentimento
- Análise de Tópicos
- Polarização de opinião
- Detecção de entidades

Analisar os comentários de 1 loja física da empresa, os dados são públicos do google maps.

Engenharia de Prompt





Prompt

Um prompt é o **texto de entrada** fornecido a um modelo de linguagem para gerar uma saída ou resposta. Ele serve como ponto de partida para a geração de texto e pode variar de uma simples palavra ou frase a um parágrafo completo, dependendo da tarefa.





Basicamente, o prompt é a informação que você fornece ao modelo de linguagem para orientá-lo na produção de uma resposta.

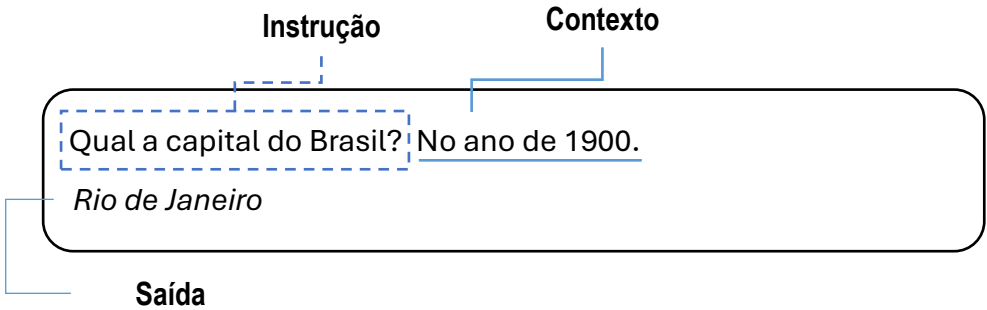


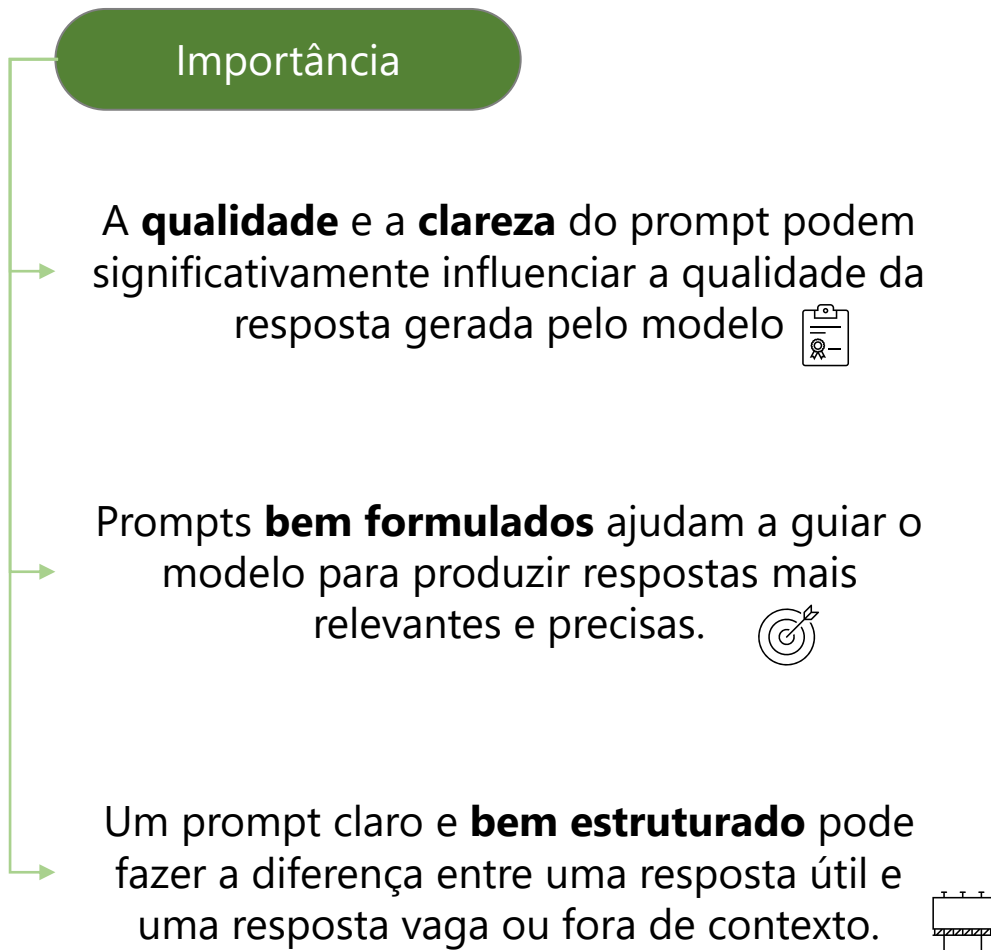
Quando começou

Década de 1960: Projeto ELIZA, um sistema de processamento de linguagem natural (PLN) que simulava um terapeuta. O ELIZA utilizava prompts simples para interagir com os usuários, demonstrando o potencial da IA na comunicação com a linguagem humana.

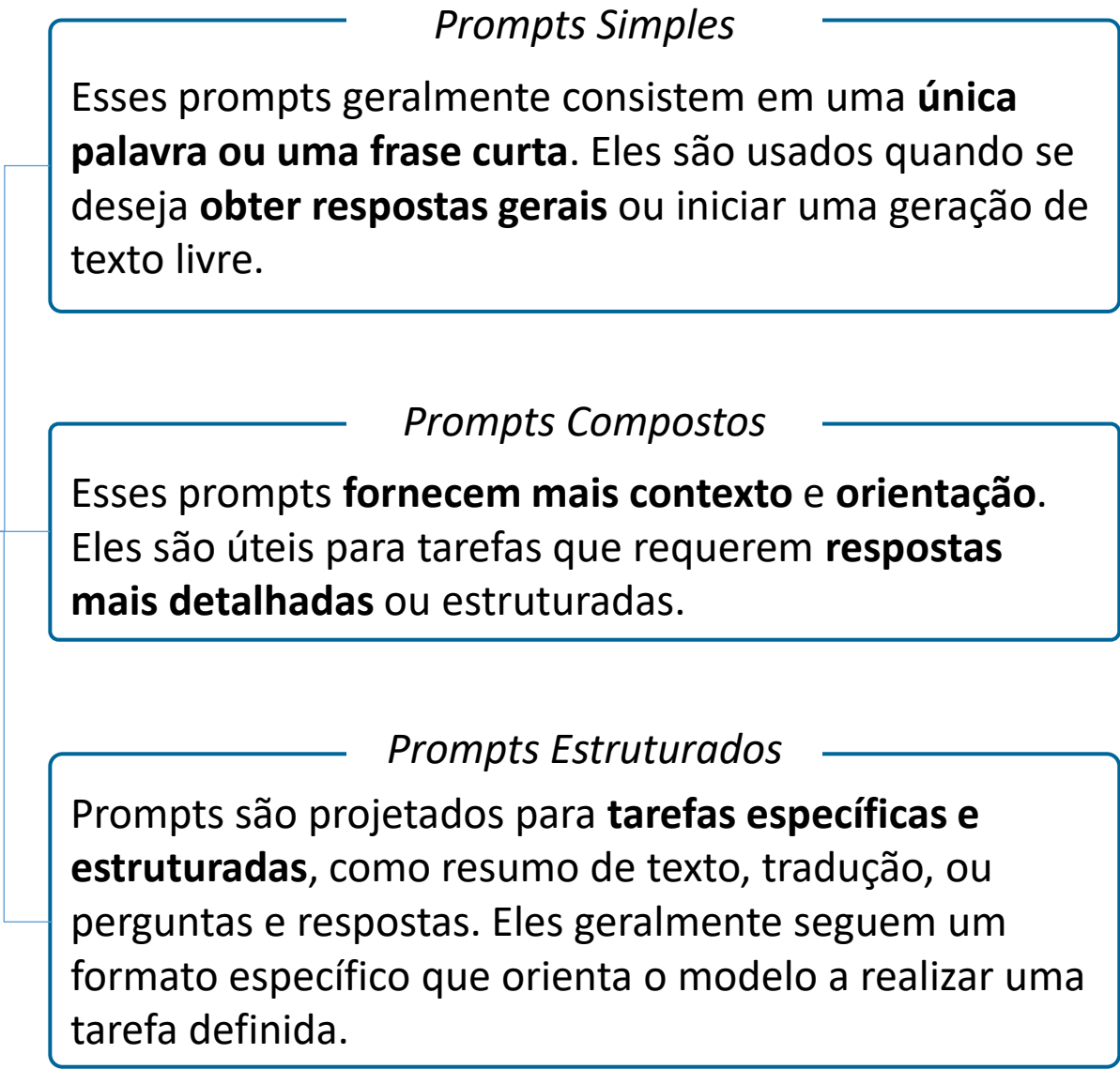
Elementos de um prompt

| | | |
|---|-------------------------|---|
|  | Instrução | qualquer tarefa específica que o usuário deseja que o modelo execute |
|  | Contexto | informações adicionais que o modelo pode usar para gerar a saída desejada |
|  | Dados de entrada | a entrada ou a pergunta cuja resposta é necessária |
|  | Saída | O formato do resultado |





Tipos de prompts





Engenharia
de Prompt

A **Engenharia de Prompt** é uma disciplina emergente que se concentra na **criação e otimização de prompts**, instruções que direcionam modelos de linguagem grandes (LLMs) como eu na geração de conteúdo, tradução de idiomas, escrita de diferentes tipos de textos criativos e até mesmo na execução de tarefas complexas.

A **Prompt Engineering** é um **campo relativamente novo**, mas já foi usado para gerar saídas e soluções para uma variedade de tarefas, incluindo:

- Escrevendo testes de Qualidade
- Gerando documentação
- Criando aplicações (linguagem de programação)
- Desenvolvendo modelos de aprendizado de máquina



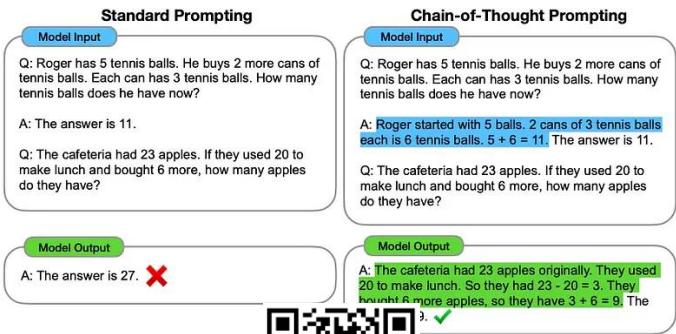
Engenheiro(a) de Prompt

Um Engenheiro de Prompt é um profissional especializado na **criação e otimização de prompts**, instruções que direcionam modelos de linguagem grandes (LLMs) na geração de conteúdo, tradução de idiomas, escrita de diferentes tipos de textos criativos e até mesmo na execução de tarefas complexas.

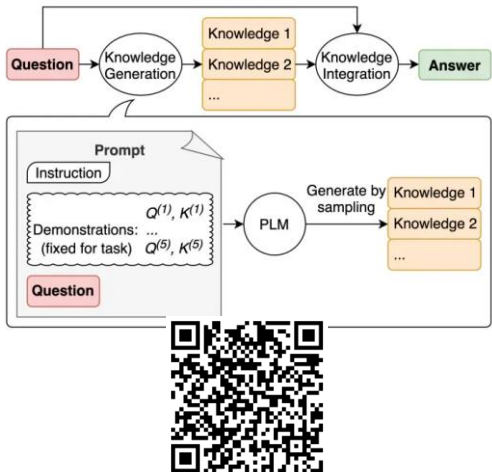
Solicitação de disparo zero



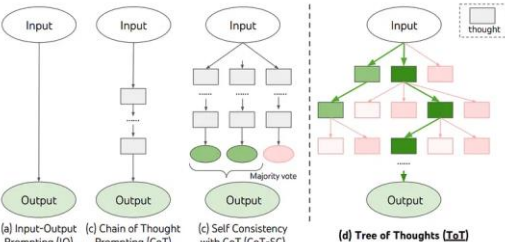
Solicitação de cadeia de pensamento (CoT)



Solicitação de Conhecimento Gerado



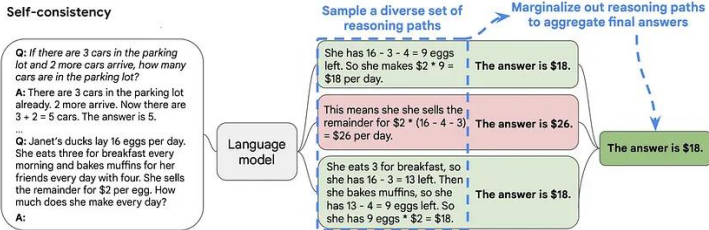
Solicitação de Árvore de Pensamentos (ToT)



Solicitação de poucos disparos



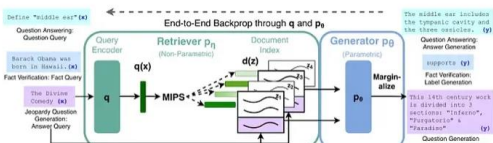
Solicitação de autoconsistência



Encadeamento de prompts



Geração Aumentada de Recuperação (RAG)



Solicitação de Disparo Zero

A **solicitação de disparo zero** refere-se ao cenário em que o modelo de linguagem recebe um prompt que não foi visto durante o treinamento e espera-se que execute alguma tarefa com base nesse prompt.

Essa técnica testa a capacidade do modelo de generalizar e entender novas instruções sem exemplos prévios. No entanto, quando essa abordagem não é eficaz, a técnica de solicitação de poucos disparos pode ser utilizada.

Prompt: "Traduza para o francês: 'Hello, how are you?'"

Resposta: "Bonjour, comment ça va?"


Solicitação de Poucos Disparos

Na **solicitação de poucos disparos**, o modelo recebe várias demonstrações ou exemplos no prompt e espera-se que aprenda a partir desses exemplos para produzir a saída desejada.

Esta técnica melhora a performance do modelo ao fornecer contexto e exemplos específicos que orientam sua resposta.

Prompt:

rust

 Copiar código

Exemplo 1: Traduza para o francês: 'Hello, how are you?' -> 'Bonjour, comment ça va?'

Exemplo 2: Traduza para o francês: 'Good morning' -> 'Bon matin'

Exemplo 3: Traduza para o francês: 'Thank you' -> 'Merci'

Traduza para o francês: 'See you later'

Resposta: "À plus tard"

Solicitação de Cadeia de Pensamento

A solicitação de **cadeia de pensamento** (Chain of Thought - CoT) permite que o modelo de linguagem forneça ao **usuário um raciocínio complexo através de etapas intermediárias**.

O CoT pode ser combinado com solicitações de poucos disparos para permitir que o modelo aprenda o raciocínio e gere o resultado necessário.

Esta técnica utiliza o método passo a passo do modelo de linguagem para gerar uma demonstração para a resposta.

Prompt:

yml

Copiar código

```
Pergunta: Se você tem 3 maçãs e compra 2 maçãs, quantas maçãs você tem no total?  
Passo 1: Você começa com 3 maçãs.  
Passo 2: Você compra mais 2 maçãs.  
Passo 3: Então, você tem 3 + 2 maçãs.  
Resposta: 5 maçãs.  
  
Pergunta: Se você tem 10 laranjas e dá 3 para um amigo, quantas laranjas você tem?  
Passo 1: Você começa com 10 laranjas.  
Passo 2: Você dá 3 laranjas para um amigo.  
Passo 3: Então, você tem 10 - 3 laranjas.  
Resposta:
```

Resposta: "7 laranjas."

Solicitação de Autoconsistência

A solicitação de autoconsistência envolve a **amostragem de várias soluções para problemas de raciocínio** usando CoT de poucas tentativas.

A ideia é permitir que o modelo de linguagem analise e **entenda o raciocínio por trás dos diferentes problemas apresentados no prompt**, aumentando assim a precisão da resposta e da explicação geradas.

```
Pergunta: Quantos anos terá uma pessoa que nasceu em 2000 no ano de 2050?  
Passo 1: O ano atual é 2050.  
Passo 2: A pessoa nasceu em 2000.  
Passo 3: Calcule a diferença de anos: 2050 - 2000.  
Resposta 1: 50 anos.
```

```
Pergunta: Quantos anos terá uma pessoa que nasceu em 2010 no ano de 2075?  
Passo 1: O ano atual é 2075.  
Passo 2: A pessoa nasceu em 2010.  
Passo 3: Calcule a diferença de anos: 2075 - 2010.  
Resposta 2: 65 anos.
```

```
Pergunta: Quantos anos terá uma pessoa que nasceu em 1995 no ano de 2045?  
Passo 1: O ano atual é 2045.  
Passo 2: A pessoa nasceu em 1995.  
Passo 3: Calcule a diferença de anos: 2045 - 1995.  
Resposta 3:
```

Respostas Amostradas:

1. "50 anos."
2. "50 anos."
3. "50 anos."

Resposta Final: "50 anos."

Solicitação de Conhecimento Gerado

Na solicitação de conhecimento gerado, o usuário pede ao modelo de linguagem para gerar **conhecimento sobre um determinado tópico** que será usado para gerar conteúdo adicional. Por exemplo, se você precisar escrever um blog sobre cricket, primeiro você pode pedir ao modelo para listar alguns fatos sobre o cricket.

Gerar conhecimento

Usar conhecimento

Passo 1: Gerar Conhecimento
Prompt:

```
Copiar código
```

Liste alguns fatos sobre o cricket.

Resposta:

1. O cricket é um esporte popular em países como Índia, Austrália e Inglaterra.
2. Um jogo de cricket pode durar de algumas horas a até cinco dias, dependendo do formato.
3. Existem três formatos principais de cricket: Test, One Day International (ODI), e Twenty20 (T20).
4. No cricket, dois times de 11 jogadores cada competem em um campo oval.
5. A principal competição internacional de cricket é a Copa do Mundo de Cricket, organizada pelo ICC (International Cricket Council).

Passo 2: Usar o Conhecimento Gerado
Prompt:

```
Copiar código
```

Usando os fatos listados, escreva um parágrafo introdutório para um blog sobre cricket.

Resposta:

```
Copiar código
```

O cricket é um esporte extremamente popular em países como Índia, Austrália e Inglaterra.

Encadeamento de Prompts

O encadeamento de prompts **divide um grande prompt em um conjunto de subprompts menores**. A resposta gerada para cada subprompt é então usada juntamente com o próximo prompt. Esta técnica ajuda a dividir problemas complexos em subproblemas menores que o modelo pode resolver mais facilmente.

Subprompts

Usar conhecimento

Prompt 1: "Responda às seguintes perguntas sobre animais:"

```
Copiar código
```

lua

Pergunta 1: Qual é o animal mais rápido do mundo?
Pergunta 2: Onde vivem os pinguins?
Pergunta 3: Como os morcegos conseguem voar à noite?

Resposta 1:

1. "O animal mais rápido do mundo é o falcão peregrino, que pode atingir velocidades de até 320 km/h em voo."
2. "Os pinguins vivem principalmente na Antártida, mas também podem ser encontrados em áreas costeiras de outros continentes, como América do Sul e África."
3. "Os morcegos têm um sistema de ecolocalização que lhes permite voar à noite, emitindo sons de alta frequência e interpretando os ecos para localizar obstáculos e presas."

Prompt 2: "Com base nas respostas anteriores, escreva um parágrafo explicando a adaptação dos animais ao seu ambiente."

```
Copiar código
```

lua

Resposta:

Os animais desenvolveram uma variedade de adaptações incríveis para sobreviver em seus amb

Solicitação de Árvore de Pensamentos

A solicitação de árvore de pensamentos (Tree of Thoughts - ToT) constrói **sobre a estrutura da Árvore de Pensamentos** e utiliza a **estrutura de cadeia de pensamento bem estabelecida**.

Esta técnica permite que o modelo exiba habilidades de raciocínio superiores, criando múltiplos blocos de conhecimento que ajudam na resolução de problemas complexos.

Prompt:

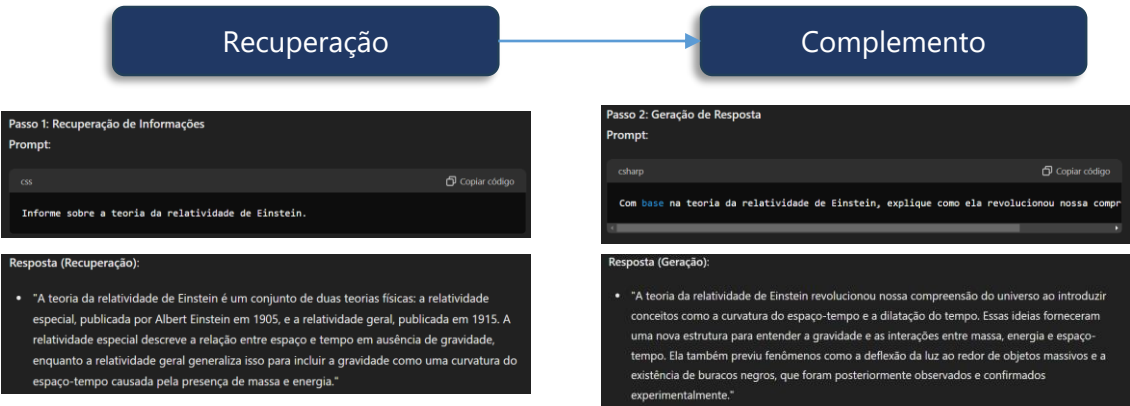
yaml

Copiar código

Pergunta inicial: Explique o ciclo da água.
Bloco 1: Descreva a evaporação.
Bloco 2: Explique a condensação.
Bloco 3: Detalhe a precipitação.
Bloco 4: Discuta a coleta.

Geração Aumentada de Recuperação

A geração aumentada de recuperação (Retrieval-Augmented Generation - RAG) utiliza um modelo de linguagem **apoiado por uma grande quantidade de dados**. Ao fornecer contexto relevante ao modelo, permite-se que ele recupere o conhecimento necessário do conjunto de dados. No caso do RAG, o modelo é complementado com um grande corpus de dados e um sistema recuperador que obtém a resposta necessária para a consulta.





Prompts inequívocos

Defina claramente a **resposta desejada no prompt** para evitar interpretações errôneas pela IA. Por exemplo, se você estiver solicitando um resumo de um romance, indique claramente que está procurando um resumo, não uma análise detalhada. Isso ajuda a IA a se concentrar apenas em sua solicitação e fornecer uma resposta alinhada ao seu objetivo.

Prompt Inadequado

Por favor, escreva sobre o artigo científico sobre a nova vacina contra a COVID-19.

Problema: Este prompt pode resultar em uma resposta que inclui uma análise detalhada, um resumo, ou até mesmo uma discussão sobre a relevância do artigo.

Prompt Adequado

Por favor, forneça um resumo de uma frase do artigo científico sobre a nova vacina contra a COVID-19, destacando a descoberta principal.



Contexto adequado

Forneça um contexto adequado no prompt e inclua os **requisitos do resultado na digitação do prompt**, limitando-os a um formato específico.

Por exemplo, digamos que você queira uma lista dos filmes mais populares da década de 1990 em uma tabela. Para obter o resultado exato, indique explicitamente quantos filmes deseja que sejam listados e solicite a formatação da tabela.

Prompt Inadequado

Liste os filmes mais populares da década de 1990.

Problema: Este prompt é vago e pode levar a uma resposta não formatada e sem limites claros sobre a quantidade de filmes.

Prompt Adequado

Liste os 10 filmes mais populares da década de 1990 em uma tabela. A tabela deve incluir as colunas: "Título", "Ano de Lançamento" e "Diretor".



Equilíbrio Informações x Resultado

Equilibre **simplicidade** e **complexidade** em seu prompt para **evitar respostas vagas**, dissociadas ou inesperadas. Um prompt muito simples pode não ter contexto, enquanto um **muito complexo pode confundir a IA**. Isso é importante principalmente para tópicos complexos ou para a linguagem específica de um domínio, que podem ser menos familiares para a IA.

Muito Simples

Explique a Teoria da Relatividade

Muito Complexo

Explique a Teoria da Relatividade, incluindo todos os aspectos da física quântica, a história da teoria, todas as fórmulas matemáticas envolvidas e as implicações filosóficas.

Adequado

Explique a Teoria da Relatividade de Einstein e suas implicações básicas para a física moderna.



Experimente e refine o prompt

A engenharia de prompt é um **processo iterativo**. É essencial **experimentar ideias diferentes e testar os prompts de IA para ver os resultados**. Você pode precisar de várias tentativas para otimizar a precisão e a relevância. Testes e iterações contínuos reduzem o tamanho do prompt e ajudam o modelo a gerar melhores resultados. Não há regras fixas sobre como a IA gera informações, portanto, flexibilidade e adaptabilidade são essenciais.



Teste e Iteração

É comum que os primeiros prompts não gerem os resultados ideais



Redução de Tamanho

Iterações ajudam a encontrar a maneira mais concisa e eficaz de formular prompts, evitando redundâncias e excessos



Flexibilidade e Adaptabilidade

Dada a natureza dinâmica da IA, é importante ser flexível e adaptar os prompts conforme necessário.

Prompt Design and Engineering: Introduction and Advanced Methods

janeiro de 2024



Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review

Junho de 2024



Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review

Solicitação de Função



A **solicitação de função** é um método fundamental na engenharia de prompts. Envolve atribuir ao modelo um **papel específico**, como o de um **assistente prestativo** ou um **especialista experiente**. Esta abordagem pode ser particularmente eficaz para orientar as respostas do modelo e garantir que elas estejam alinhadas com o resultado desejado.

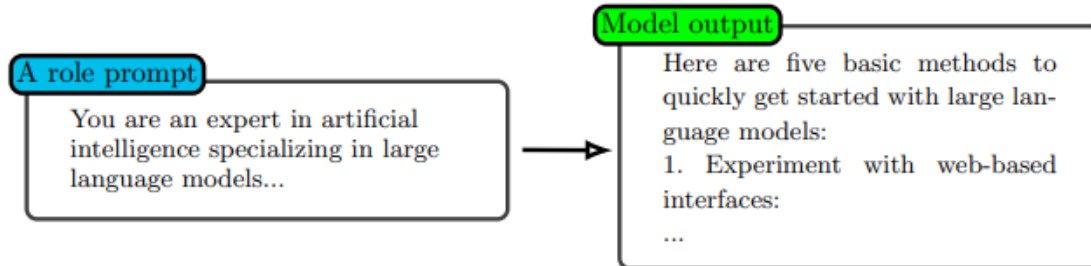


Fig. 3 Role prompting example.

Aspas Triplas para Separar

Na engenharia de prompts, o uso de aspas triplas (""") é uma técnica eficaz para separar diferentes partes de um prompt ou encapsular strings multilinhas. Essa prática é especialmente útil quando se lida com prompts complexos que possuem vários componentes ou quando o próprio prompt contém aspas

Exemplo:

```
python Copiar código

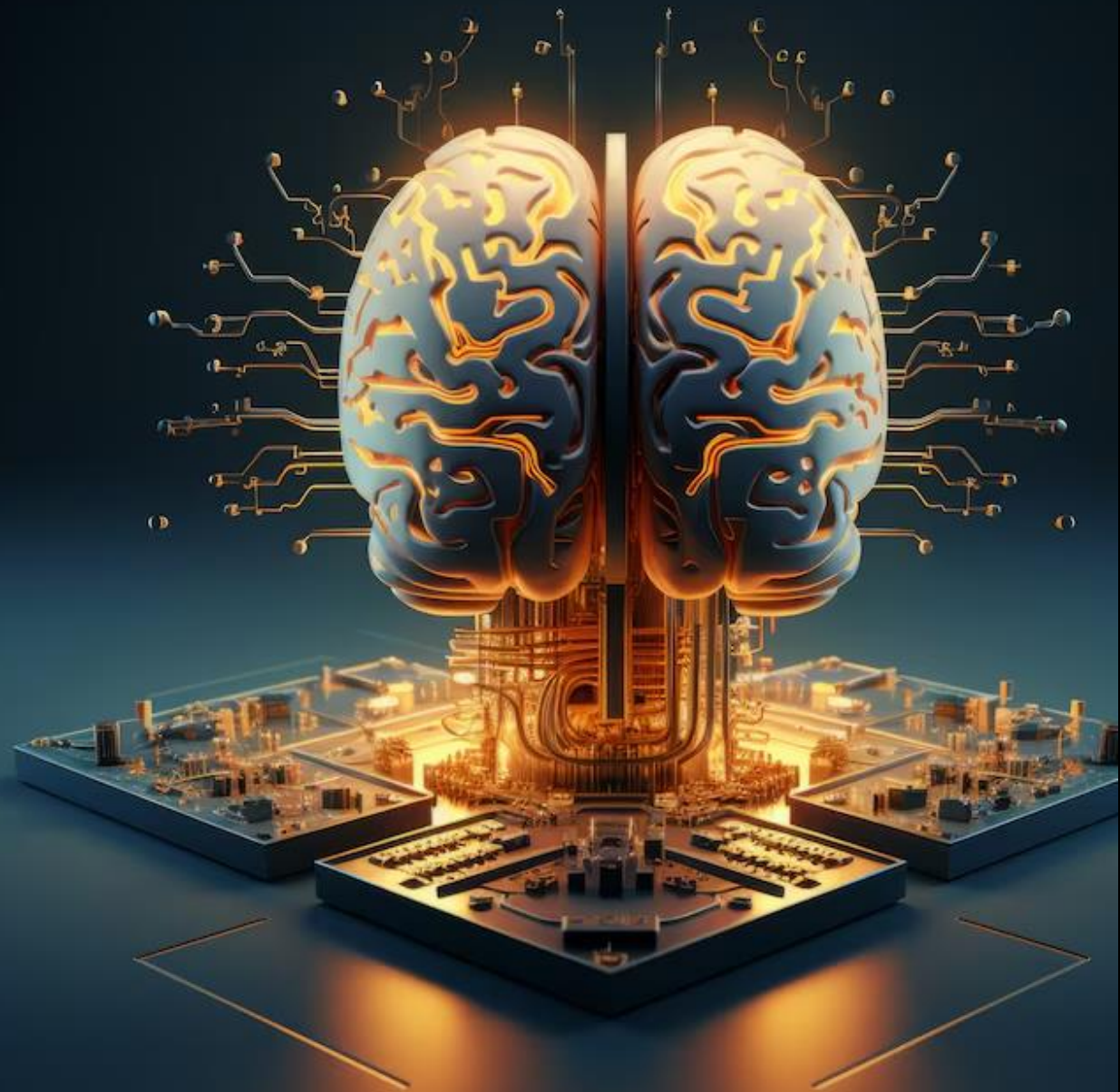
prompt = """
Contexto: A fotossíntese é o processo pelo qual as plantas convertem luz solar em

Instrução: Explique as etapas da fotossíntese.
"""
```

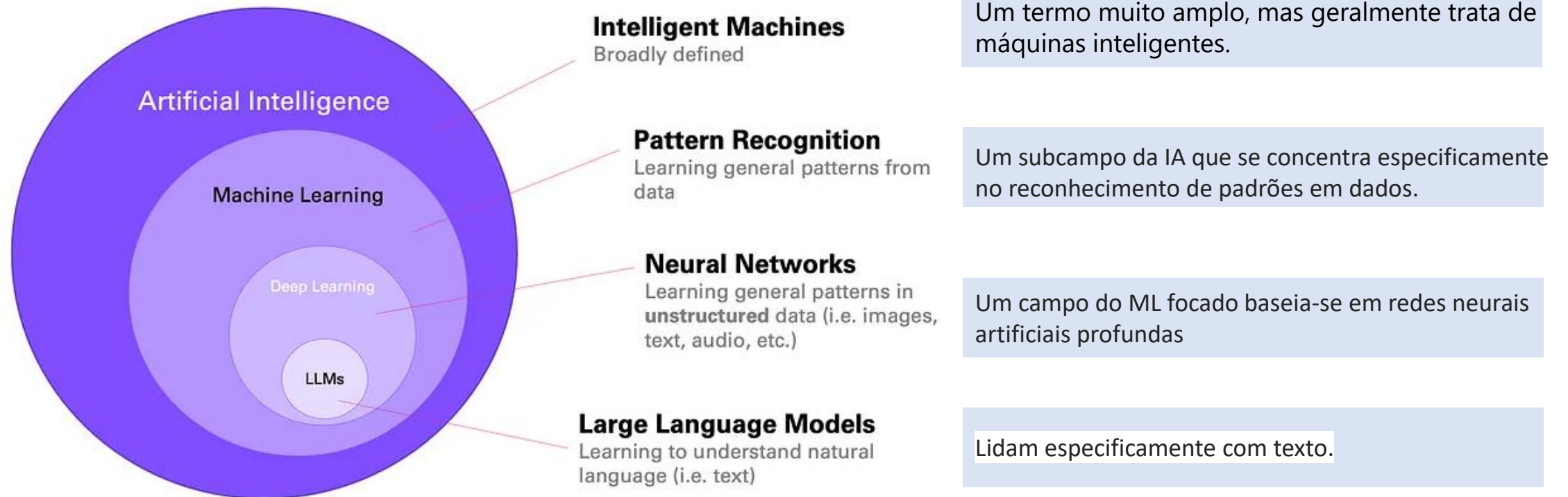


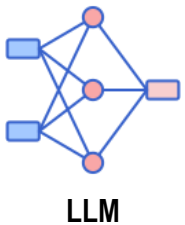

LLM

Grandes Modelos de Linguagem



O campo da Inteligência Artificial em camadas.





Grandes Modelos de Linguagem (Large Language Models, LLMs) são algoritmos de **aprendizado de máquina** que utilizam **redes neurais profundas** para processar e gerar texto natural.

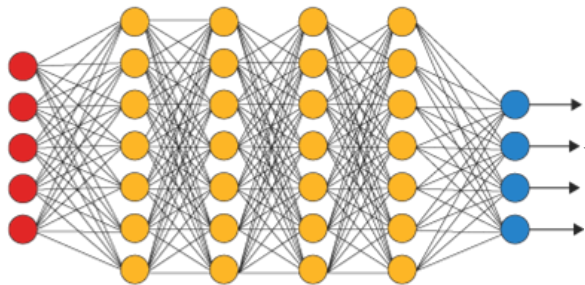


Ilustração de uma arquitetura de rede neural profunda

Os LLMs são capazes de realizar uma **variedade de tarefas de processamento de linguagem natural (NLP)**, como; tradução, resumo de texto, resposta a perguntas e geração de texto, com um alto nível de precisão e fluidez.

Como funciona uma LLM

O modelo aprende a reconhecer as **relações entre as palavras** e como elas são usadas em **diferentes contextos**. Isso permite que ele **gere texto** que seja gramaticalmente correto e que faça sentido semântico.

3 pontos sobre treinamento



Complexo

O processo de treinamento de um LLM é **bastante complexo** e requer recursos computacionais consideráveis.



Dados

O modelo precisa ser exposto a uma **enorme quantidade de dados** para que possa aprender a gerar texto de qualidade.

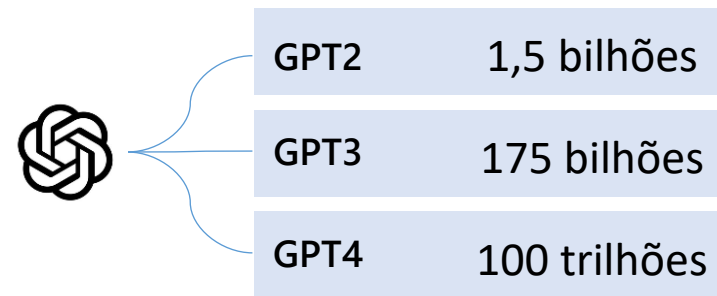


Tempo

O treinamento pode **levar dias, semanas ou até meses**, dependendo do tamanho do modelo e da quantidade de dados disponível.

Grandes Modelos de Linguagem ?

Neste caso refere-se simplesmente ao número de neurônios, também **chamados de parâmetros**, na rede neural.



Language modeling

Qual é a próxima palavra em uma determinada sequência de palavras?

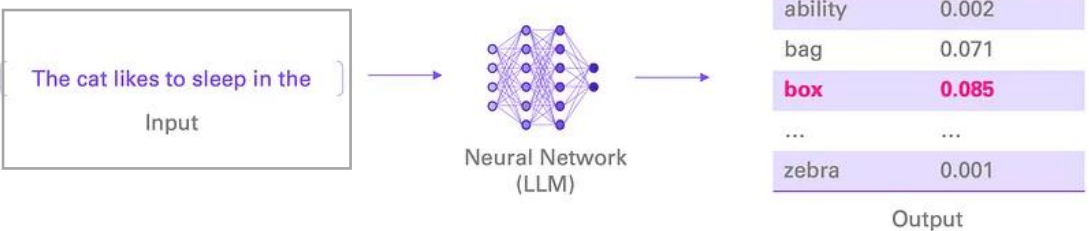
Imagine the following task: Predict the next word in a sequence

[The cat likes to sleep in the _____] → What word comes next?

Can we frame this as a ML problem? Yes, it's a classification task.

Now we have (say) ~50,000 classes (i.e. words)

Nessa situação para o modelo preencher a frase ele terá mais de 50 mil palavras (classes)



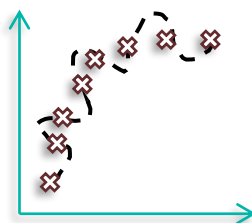
Caso queria criar sua LLM



Desafios

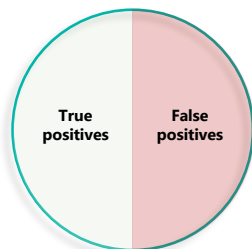
OVERFITTING

Ocorre quando um LLM **aprende os dados de treinamento muito bem**, mas não é capaz de **generalizar para novos dados**. Isso pode ser causado por um número excessivo de parâmetros ou por parâmetros que não estão bem ajustados.



INTERPRETABILIDADE

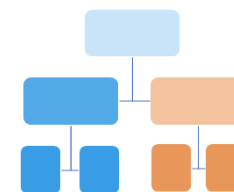
É difícil entender como os parâmetros de um LLM afetam o texto que ele gera. Isso dificulta a depuração de LLMs e a identificação de erros.



Pesquisa

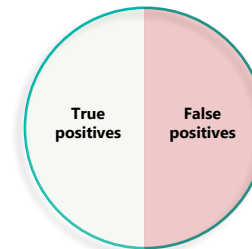
PARÂMETROS

Redução de parâmetros: Pesquisadores estão trabalhando em técnicas para reduzir o número de parâmetros em LLMs sem sacrificar o desempenho.



INTERPRETABILIDADE

Pesquisadores também estão trabalhando em técnicas para tornar os LLMs mais interpretáveis.





No contexto dos Grandes Modelos de Linguagem (LLMs), **alucinação refere-se à geração de informações imprecisas**, irrelevantes ou factualmente incorretas por parte do modelo. Isso ocorre quando o modelo "inventa" dados ou produz respostas que não são baseadas em evidências reais ou contextos apropriados.

Treinamento em Dados Imperfeitos

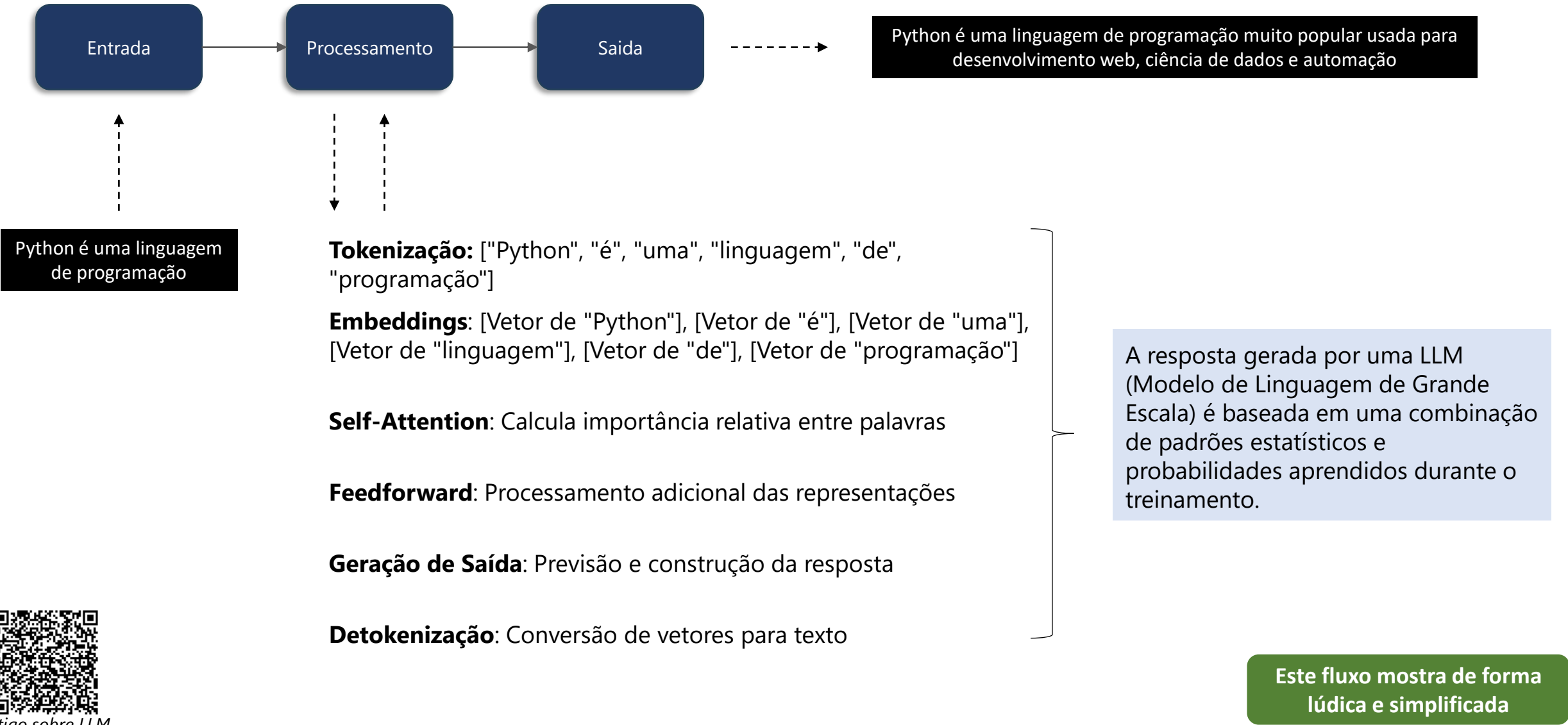
Os LLMs são treinados em **grandes volumes de dados textuais da internet**, que podem conter informações errôneas, contraditórias ou desatualizadas. A absorção desses dados pode levar a respostas alucinadas.

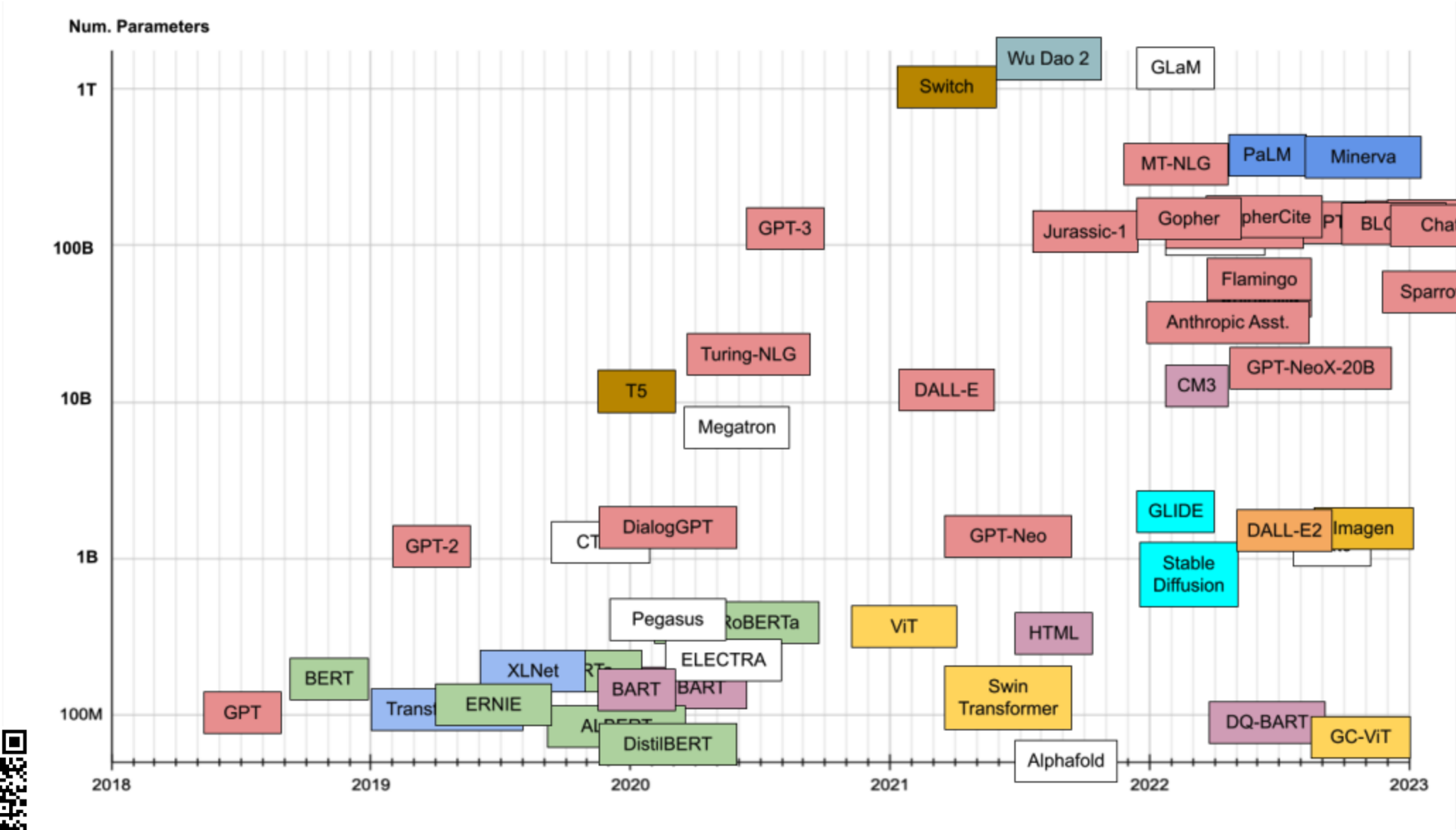
Limitações na Compreensão de Contexto

Apesar de sua capacidade avançada de processamento de linguagem, os LLMs podem falhar em compreender completamente o contexto ou a **nuance de certas perguntas**, levando a respostas imprecisas.

Propagação de Erros

Quando uma alucinação ocorre, ela pode ser exacerbada em subsequentes interações, especialmente em diálogos longos, onde o modelo tenta manter a coerência com suas respostas anteriores.







Projeto





Magazine Luiza

Perguntar à comunidade

Resumo de avaliações

?

5

4

3

2

1

4,0

2.820 avaliações

F

"Loja boa com uma boa **variedade** de **produtos** e o preço está na medida do mercado."

"No **setor** de **entregas** de mercadorias, os **profissionais** são rápidos e eficientes."

D

"Vendedores mal educados com **clientes** que efetuam **compras** via **site**"

Dados públicos

Uso para estudo

Acadêmico



Gmaps: O cliente expressa a opinião dele de forma pública.



Marcus Silva
Local Guide · 122 avaliações · 18 fotos



→ ★★★★★ 6 dias atrás NOVA

A vendedora Eduarda, super profissional, educada, e sabe o que faz, tirou nossas dúvidas sobre o item em questão da compra. Recomendamos, procurem a esta profissional. A empresa merece a profissional que tem! Parabéns Magazine Luiza



Jessica Santos
1 avaliação



★★★★★ 2 meses atrás

Quero deixar aqui a minha total indignação com essa loja lixo que é a magazine. Fiz o cartão da loja não recebi e muito menos desbloqueei o cartão estão me cobrando o valor do seguro que eu não pedi e a anuidade do mesmo. sendo que o Cartão já foi até cancelado e mesmo assim está vindo cobrança já fui diversas vezes na loja já liguei inúmeras vezes na central e nada e resolvido colocaram o meu nome no SPC Serasa falando que fizeram um extorno na minha conta que já provei que na minha conta não tem extorno nenhum estão se negando a entrega o contrato do cartão e do seguro que segundo o funcionário foi erro do sistema



Daniela Batista
7 avaliações



★★★★★ 4 meses atrás

Atendimento pessimo do atendente Leandro da Retirada de compra online. Fui para retirar uma escova de apenas 120 reais, valor baixo, apresentei meu documento e o mesmo informou que a carteira de trabalho não era válida, sendo que entrei em contato com o Magalu aonde informaram que poderia retirar com a mesma. Devido a esse tipo de atendimento que ninguém deveria retornar ao local, fora que o mesmo é super arrogante.



Daniela Batista
7 avaliações



★☆☆☆☆ 4 meses atrás

Atendimento **pessimo do atendente Leandro da Retirada de compra online.**

Fui para **retirar** uma escova de apenas 120 reais, valor baixo , apresentei meu documento e o mesmo informou que a carteira de trabalho nao era valido, sendo que entrei em **contato com o Magalu** aonde informaram que poderia retirar com a mesma.

Devido a esse tipo de atendimento que **ninguém deveria retornar ao local, fora que o mesmo e super arrogante.**

Podemos identificar

Sentimento do Cliente

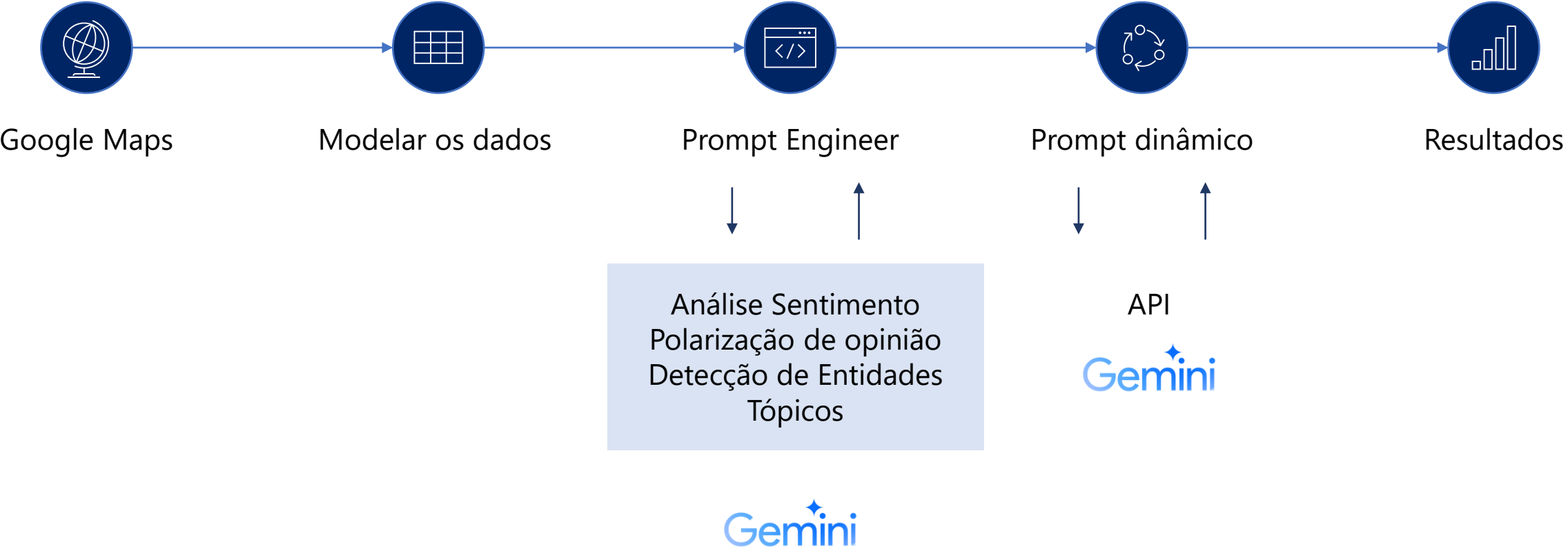
Entidades

Tópico

Polarização Opinião

Processamento
de linguagem
natural (PLN)

Etapas do projeto



MÃO NA
MASSA