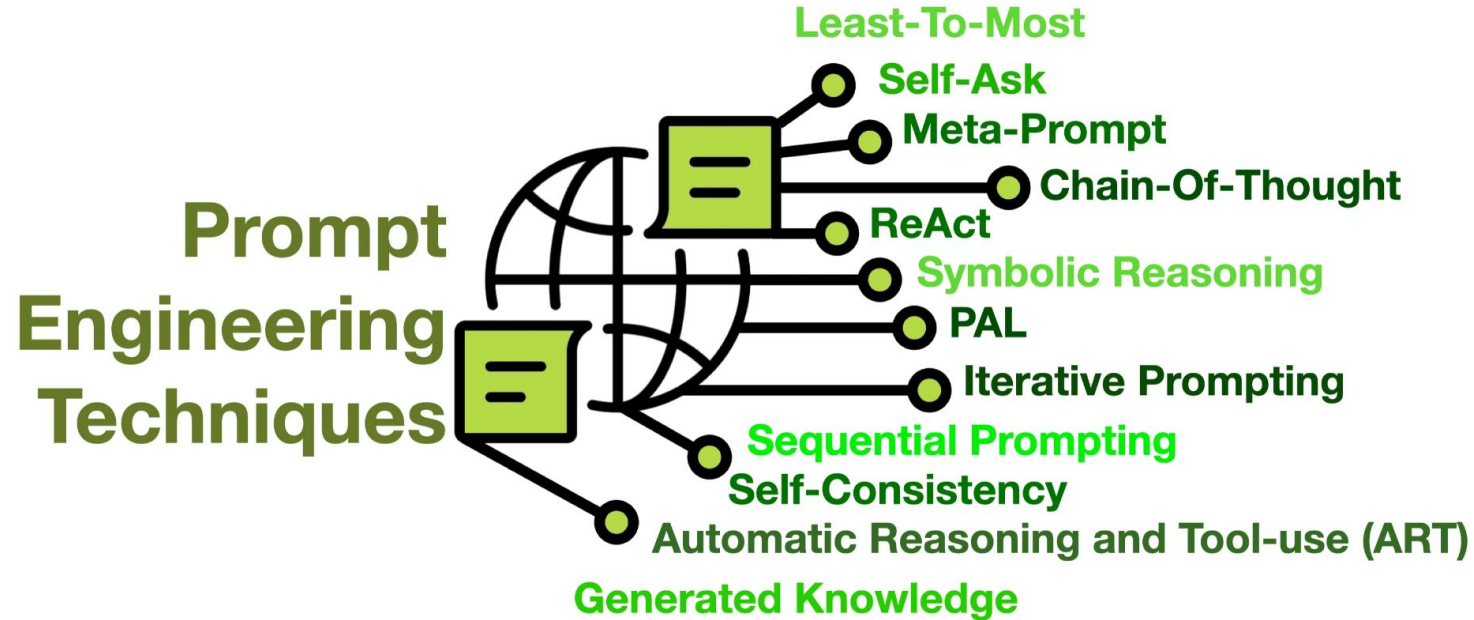


ETAPA 5

Técnicas de engenharia de Prompt

12 Prompt Engineering Techniques





Técnicas de Engenharia de Prompt

Least-to-Most Prompting: segue o princípio da divisão de problemas complexos em subtarefas, facilitando que o modelo resolva etapas sequencialmente. Essa técnica oferece maior controle e clareza no processo, ideal para problemas multi-etapa. No entanto, exige planejamento cuidadoso para sequências corretas.

- **Vantagens:** Melhora a precisão, facilita a resolução de tarefas complexas e permite flexibilidade entre LLMs para diferentes partes do problema.
- **Limitações:** Pode ser menos eficiente em questões que não requerem múltiplas etapas, bem como dependência e de prompts anteriores

Tarefa principal: escrever um código em Python para expor uma API com um serviço de LLM.

- **Prompt 1:** escreva um código “hello world” de uma aplicação Flask simples;
- **Prompt 2:** escreva o código da aplicação Flask para receber uma string chamada prompt.
- **Prompt 3:** adicione à aplicação um código simples para receber a string prompt e utilizar com a biblioteca `google.generative_ai` e LLM Gemini-Flash.
- **Prompt 4:** escreva um código em separado para testar essa aplicação com o prompt = “Olá”.



Técnicas de Engenharia de Prompt

Self-Ask é uma abordagem onde o LLM, ao responder uma **pergunta complexa**, divide-a em **subperguntas** resolvidas sequencialmente. Isso permite quebrar a questão em partes menores e mais gerenciáveis, produzindo uma **resposta detalhada** e estruturada. Cada **subpergunta** pode ser **respondida** com **fontes confiáveis** de dados, deixando o raciocínio final ao LLM.

- **Vantagens:** capacidade de gerar respostas mais precisas e detalhadas e de melhorar o entendimento de questões ambíguas ou complexas.
- **Desafios:** o aumento no tempo de processamento e coerência em perguntas com múltiplas interpretações.

Prompt:

Pergunta Principal: "Quais foram os principais fatores que levaram à queda do Império Romano?"

Você precisa considerar os seguintes questionamentos antes de responder:

- "Para entender a queda do Império Romano, quais foram os principais desafios econômicos que enfrentaram?"
- "Quais fatores sociais e culturais influenciaram a desestabilização do império?"
- "Que eventos militares e invasões contribuíram diretamente para a queda?"
- "Como as divisões internas de poder e a corrupção afetaram o império?"

Técnicas de Engenharia de Prompt

A técnica **ReAct** combina **raciocínio** e **ação** em prompts, permitindo que LLMs respondam a **perguntas complexas**, **decompondo-as** em etapas de **raciocínio** e **busca ativa** de informações.

- **Integração de Ações e Respostas:** ao combinar raciocínio com ações (como consultas de dados ou cálculos), o LLM pode ajustar suas respostas com base em novos dados ou processos iterativos.
- **Redução de Erros:** Como o modelo verifica etapas ao longo do processo, ReAct pode minimizar erros em problemas complexos, garantindo que a resposta final seja bem fundamentada e construída de forma lógica.
- **Maior Tempo de Resposta:** envolve múltiplas etapas de raciocínio e ação, aumentando o tempo de processamento.
- **Consumo de Recursos:** envolve uma série de prompts e respostas intermediárias para completar a tarefa.

(1) Hotspot QA

Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

(1d) ReAct (Reason + Act)

Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with.

Act 1: `Search[Apple Remote]`

Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple ... originally designed to control the `Front Row media center` program ...

Thought 2: Apple Remote was originally designed to control the Front Row media center program. I need to search `Front Row` next and find what other device can control it.

Act 2: `Search[Front Row]`

Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports', 'Front Row (software)', ...]

Thought 3: Front Row is not found. I need to search `Front Row (software)`.

Act 3: `Search[Front Row (software)]`

Obs 3: Front Row is a discontinued media center software ...

Thought 4: `Front Row (software)` is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.

Act 4: `Finish[keyboard function keys]`



ReAct

```
%%capture
# update or install the necessary libraries
!pip install --upgrade openai
!pip install --upgrade langchain
!pip install --upgrade python-dotenv
!pip install google-search-results

# import libraries
import openai
import os
from langchain.llms import OpenAI
from langchain.agents import load_tools
from langchain.agents import initialize_agent
from dotenv import load_dotenv
load_dotenv()

# load API keys; you will need to obtain these if you haven't yet
os.environ["OPENAI_API_KEY"] = os.getenv("OPENAI_API_KEY")
os.environ["SERPER_API_KEY"] = os.getenv("SERPER_API_KEY")
```

```
llm = OpenAI(model_name="text-davinci-003", temperature=0)
tools = load_tools(["google-serper", "llm-math"], llm=llm)
agent = initialize_agent(tools, llm, agent="zero-shot-react-description", verbose=True)
```

```
agent.run("Who is Olivia Wilde's boyfriend? What is his current age raised to the 0.23 power?")
```

```
> Entering new AgentExecutor chain...
  I need to find out who Olivia Wilde's boyfriend is and then calculate his age raised to the 0.23 power.
Action: Search
Action Input: "Olivia Wilde boyfriend"
Observation: Olivia Wilde started dating Harry Styles after ending her years-long engagement to Jason Sudeikis – see their relationship timeline.
Thought: I need to find out Harry Styles' age.
Action: Search
Action Input: "Harry Styles age"
Observation: 29 years
Thought: I need to calculate 29 raised to the 0.23 power.
Action: Calculator
Action Input: 29^0.23
Observation: Answer: 2.169459462491557

Thought: I now know the final answer.
Final Answer: Harry Styles, Olivia Wilde's boyfriend, is 29 years old and his age raised to the 0.23 power is 2.169459462491557.

> Finished chain.
```

```
"Harry Styles, Olivia Wilde's boyfriend, is 29 years old and his age raised to the 0.23 power is 2.169459462491557."
```

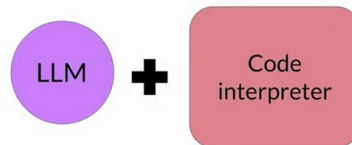
Técnicas de Engenharia de Prompt

Symbolic Reasoning e **PAL (Program-Aided Language Model)** aprimoram a capacidade dos modelos de linguagem em **resolver problemas** complexos que envolvem **lógica formal**, **cálculos simbólicos** e manipulação de variáveis.

Symbolic Reasoning: utiliza **representações simbólicas** para resolver **equações** e aplicar **regras lógicas**. Problemas onde o raciocínio passo a passo é fundamental, como manipulação algébrica e deduções matemáticas.

PAL: vai um passo além, permitindo que o **LLM gere e execute código** (Python) para resolver problemas. Em vez de apenas descrever a solução, o **modelo cria programas** que executam cálculos ou **validam resultados**. Isso é útil para tarefas de **análise de dados**, cálculos científicos, e outros problemas computacionais.

Program-aided language (PAL) models



Source: Gao et al. 2022, "PAL: Program-aided Language Models"

Chain-of-Thought (Wei et al., 2022)

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 tennis balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The bakers at the Beverly Hills Bakery baked 200 loaves of bread on Monday morning. They sold 93 loaves in the morning and 39 loaves in the afternoon. A grocery store returned 6 unsold loaves. How many loaves of bread did they have left?

Model Output

A: The bakers started with 200 loaves. They sold 93 in the morning and 39 in the afternoon. So they sold $93 + 39 = 132$ loaves. The grocery store returned 6 loaves. So they had $200 - 132 + 6 = 62$ loaves left. The answer is 62.

✗

Program-aided Language models (this work)

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 tennis balls.
`tennis_balls = 5`
2 cans of 3 tennis balls each is
`bought_balls = 2 * 3`
tennis balls. The answer is
`answer = tennis_balls + bought_balls`

Q: The bakers at the Beverly Hills Bakery baked 200 loaves of bread on Monday morning. They sold 93 loaves in the morning and 39 loaves in the afternoon. A grocery store returned 6 unsold loaves. How many loaves of bread did they have left?

Model Output

A: The bakers started with 200 loaves
`loaves_baked = 200`
They sold 93 in the morning and 39 in the afternoon
`loaves_sold_morning = 93`
`loaves_sold_afternoon = 39`
The grocery store returned 6 loaves.
`loaves_returned = 6`
The answer is
`answer = loaves_baked - loaves_sold_morning - loaves_sold_afternoon + loaves_returned`
`>>> print(answer)`
`62`

✓



Modelos de classificação utilizando LLMs através de few-shot learning

Few-shot learning: o modelo de linguagem é capaz de **classificar** novos dados a partir de **poucos exemplos anotados**. Essa técnica é especialmente útil quando há **limitação de dados** para o treinamento do modelo, pois permite que ele **aprenda padrões** a partir de um número mínimo de exemplos.

- Selecionar exemplos representativos da tarefa para guiar o LLM.
- Formular o prompt com clareza, incluindo as instruções e exemplos.
- Testar e ajustar os exemplos até que o modelo apresente respostas confiáveis e consistentes.

Prompt:

Gere cinco sugestões de títulos para um artigo sobre o tema de inteligência artificial aplicado à saúde. Siga o formato dos exemplos abaixo:

- Exemplo 1: 'Como a Inteligência Artificial Está Transformando a Medicina'
- Exemplo 2: 'Inteligência Artificial e Saúde: Inovações e Desafios'
- Exemplo 3: 'O Papel da IA no Diagnóstico de Doenças'

Gere títulos criativos e informativos

Prompt:

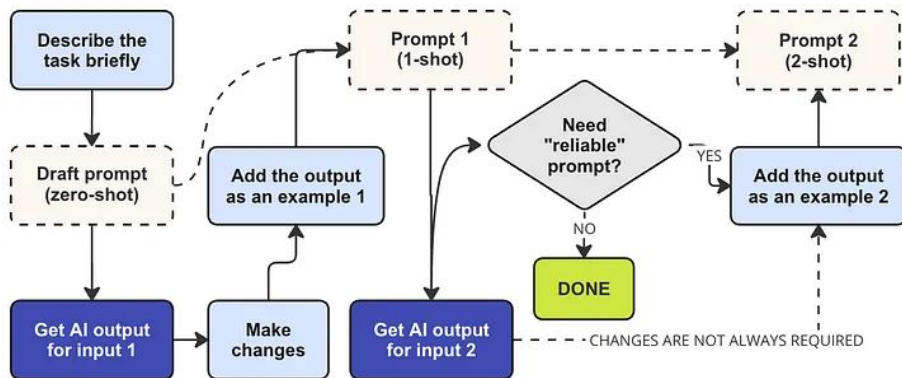
Gere uma lista de tarefas para iniciar um projeto de desenvolvimento de software, semelhante aos exemplos abaixo:

- Exemplo 1: 'Definir requisitos funcionais e não funcionais do sistema'
- Exemplo 2: 'Criar um esboço do fluxo de trabalho do usuário'
- Exemplo 3: 'Elaborar protótipos iniciais da interface'

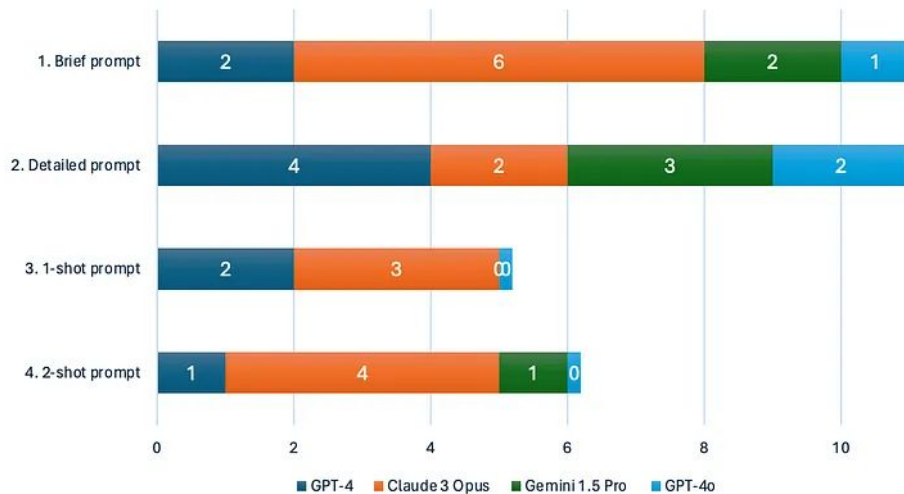
Siga o estilo direto e orientado para a ação

Modelos de classificação utilizando LLMs através de few-shot learning

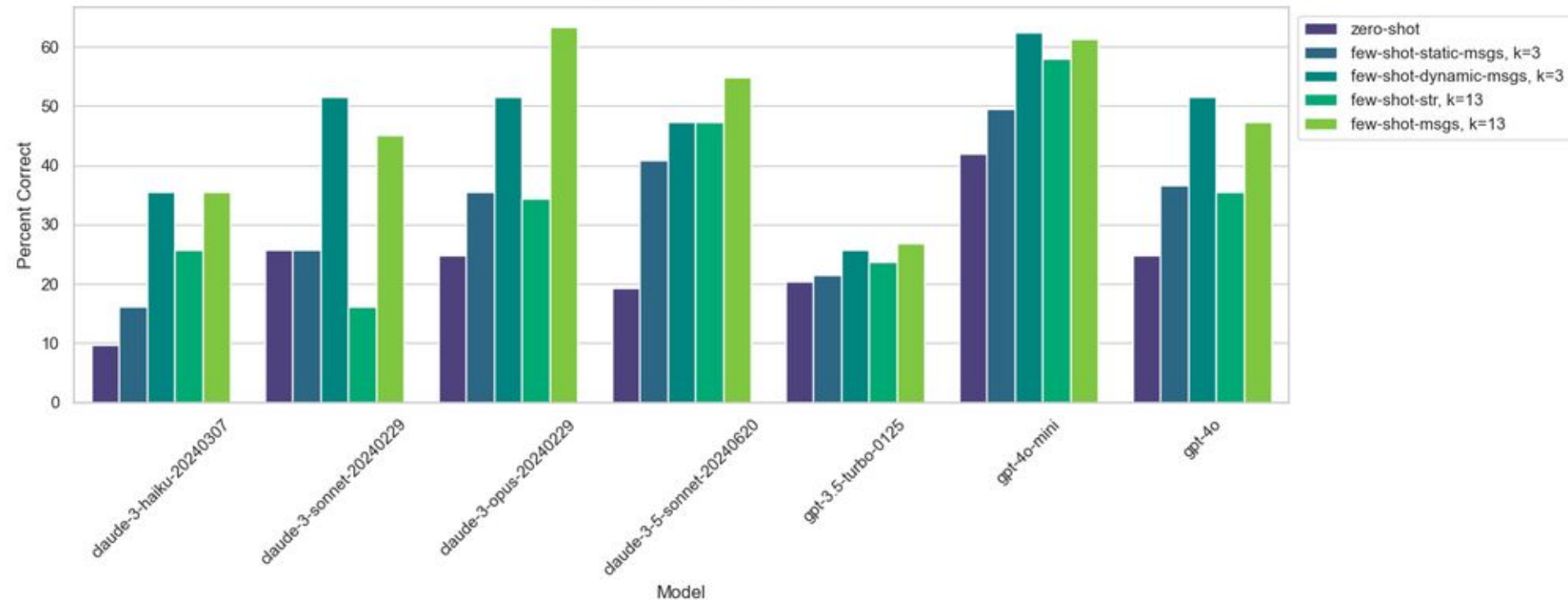
Quick creation of few-shot prompts



Total number of defects in (not-first) outputs



Extraction Task Results





Meta prompting para otimizar o desempenho de LLMs

Meta prompting: técnica de design de prompts onde um modelo de linguagem é instruído a criar seus próprios prompts. Meta prompting permite que o modelo escreva instruções para si mesmo, tornando-o capaz de resolver problemas de múltiplas etapas sem intervenção humana constante.

- **Agilidade:** Reduz o tempo necessário para criar prompts personalizados
- **Autonomia:** Habilita o modelo a gerar instruções para novas tarefas sem supervisão.
- **Dependência de Revisão:** O output do modelo ainda requer ajustes e validação humana.
- **Complexidade Inicial:** pode exigir um planejamento cuidadoso e conhecimento avançado de prompt.

Prompt:

Crie um prompt que irá gerar uma lista de perguntas para uma entrevista de emprego para a posição de *Cientista de Dados*. As perguntas devem ser divididas em três seções: habilidades técnicas, habilidades interpessoais e resolução de problemas. As perguntas devem permitir avaliar a experiência e o raciocínio do candidato.

Prompt:

Escreva um prompt que gere uma estrutura detalhada para um artigo de blog sobre *Inteligência Artificial*. A estrutura deve incluir uma introdução que apresenta o tema, seções intermediárias com sub-tópicos relevantes e uma conclusão. Cada seção deve sugerir pontos principais a serem abordados, com um tom informativo e acessível para o público.



EXERCÍCIOS

Tarefa 1: Classificação de sentimento das falas (few-shot prompting)

Criar Agentes de IA que classificam as falas dos personagens como positivas, negativas ou neutras.

Tarefa 2: Meta-prompting para criar Atores dos Simpsons

Usar meta-prompting para aprimorar o prompt de criação dos atores dos personagens.

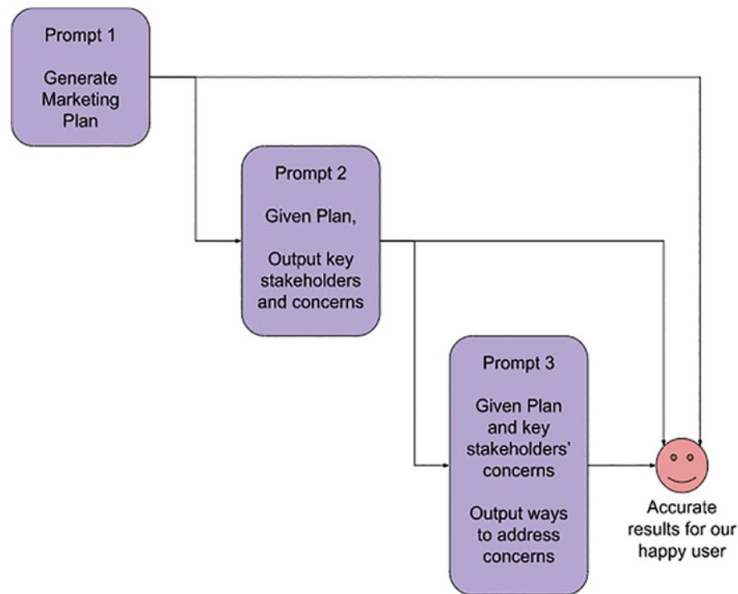


Técnicas de Engenharia de Prompt

Chain-of-Thought Prompting: Estimula o modelo a **pensar** em **sequência**, detalhando o raciocínio antes da resposta final. Útil para processos analíticos complexos, facilita respostas mais robustas, mas pode gerar conteúdo extenso em questões simples.

- **Vantagens:** Aumenta a clareza e lógica de respostas.
- **Limitações:** Excesso de detalhe pode ser redundante em perguntas diretas.

Exemplo: Enumere as etapas de um plano de marketing de um lançamento musical para o carnaval.





Técnicas de Engenharia de Prompt

Self-Consistency: utiliza **múltiplas** tentativas de **raciocínio** para uma mesma **pergunta**, selecionando a **resposta** mais **consistente** entre elas. Se baseia no princípio de que avaliar muitas respostas candidatas aumenta a precisão da resposta final.

- **Vantagens:** melhora a confiabilidade de LLM através da “média” das respostas a um mesmo prompt, robusto a outliers na resposta e medição de confiança.
- **Limitações:** além do custo e do tempo de analisar respostas longas, vieses de consenso e não recomendado para tarefas criativas.

Self-Consistency Prompt

Criar um prompt específico

Solve the following math problem step by step: A train travels at a speed of 60 km/h for 2 hours, then at 80 km/h for 1 hour. What is the average speed of the train for the entire journey? Provide your answer in km/h, rounded to two decimal places.

Gerar diversas respostas para o mesmo prompt

Comparar e examinar as respostas

Agregar os resultados para resposta final



Técnicas de Engenharia de Prompt

Generated Knowledge: técnica que envolve pedir ao LLM que gere **informações úteis** sobre um **tema** antes de **responder à pergunta** ou completar uma tarefa complexa. Essa base de conhecimento aumenta a precisão e profundidade das respostas (prompt único ou duplo).

- **Vantagens:** respostas detalhadas e bem-informadas, útil para temas complexos.
- **Limitações:** mais lento, pois envolve múltiplas chamadas ao LLM.

Prompt:

"Liste 3 impactos do aquecimento global no meio ambiente."

Resposta:

- Aumento das temperaturas médias globais.
- Derretimento de geleiras e elevação do nível do mar.
- Alteração dos padrões climáticos e aumento de eventos climáticos extremos.

Prompt:

"Usando os fatos sobre o aumento das temperaturas, derretimento de geleiras e alterações climáticas, escreva um parágrafo sobre os impactos do aquecimento global no meio ambiente."



Técnicas de Engenharia de Prompt

Automatic Reasoning and Tool-Use (ART) :
combina **raciocínio** em **múltiplas etapas** e uso
de **ferramentas externas** para resolver tarefas
complexas com precisão.

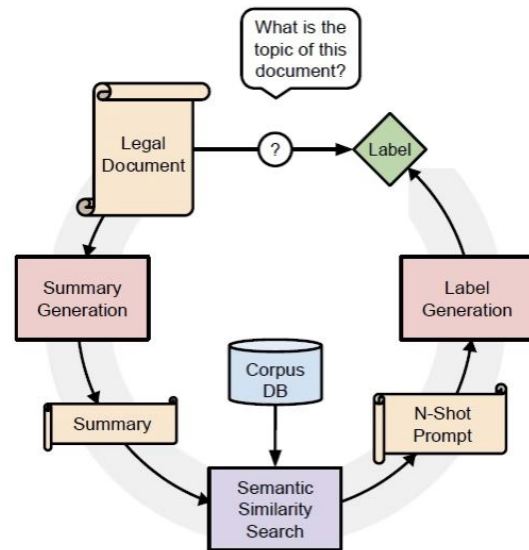
Combina **CoT** com ferramentas externas,
permitindo ao LLM pausar o raciocínio para
realizar consultas ou cálculos específicos para
melhorar a resposta.

- **Vantagens:** precisão, confiabilidade e extensível.
- **Limitações:** demanda computacional alta, necessidade sistemas configurados para suportar chamadas de API.

Feature	CoT	Auto CoT	Tool-former	ART
Multi-step reasoning	✓	✓		✓
Limited supervision		✓	✓	✓
Tool use			✓	✓
Extendable libraries				✓
Cross-task transfer		✓	✓	✓
Human feedback	✓			✓

Tarefas complexas com prompt chaining

O **prompt chaining** é uma técnica poderosa para resolver problemas complexos, dividindo-os em **sub-tarefas gerenciáveis**. Em vez de tentar resolver o problema de uma só vez, o prompt chaining permite **encadear uma série de prompts**, onde cada etapa contribui para a solução final. Essa abordagem é especialmente útil em questões que exigem várias etapas, como **resumo de textos extensos**, **análise de múltiplas variáveis** ou a **geração de relatórios detalhados a partir de dados**.





Tarefas complexas com prompt chaining

Vantagens incluem a capacidade de lidar com tarefas complexas **sem sobrecarregar** o modelo, enquanto **desafios** podem envolver a **manutenção** de **coerência** e contexto entre os prompts, além do **tempo** de **processamento** adicional para múltiplas etapas.

Pontos importantes:

- **Divisão em sub-tarefas:** facilita a resolução de problemas complexos.
- **Encadeamento de prompts:** permite refinar e expandir a solução.
- **Desafios:** manter contexto e otimizar o tempo de resposta

Prompt 1: "Leia o seguinte artigo e forneça um resumo básico com as ideias principais e os tópicos abordados: [trecho do artigo]."

Prompt 2: "Baseado no resumo inicial, forneça uma análise detalhada de cada seção do artigo, incluindo introdução, metodologia, resultados e conclusão."

Prompt 3: "Com base no resumo e na análise detalhada das seções, organize um relatório estruturado com títulos para cada seção e uma conclusão crítica sobre o impacto do estudo."



Tarefas complexas com prompt chaining

Prompt 1: "Liste os requisitos essenciais para desenvolver um sistema de gerenciamento de inventário para pequenas empresas."

Prompt 2: "Com base nos requisitos listados, organize-os em módulos principais, como 'Cadastro de Produtos', 'Controle de Estoque' e 'Relatórios de Vendas'."

Prompt 3: "Para cada módulo, defina tarefas específicas para a implementação, incluindo etapas de desenvolvimento e testes necessários para cada funcionalidade."

Prompt 1: "Identifique os principais tópicos de um curso de Biologia para o ensino médio que devem ser cobertos em um guia de estudo."

Prompt 2: "Para cada tópico listado, forneça um resumo curto que explique os conceitos principais e quaisquer fórmulas ou diagramas importantes."

Prompt 3: "Baseado nos tópicos e resumos, crie três perguntas de prática para cada tópico, variando entre questões conceituais e questões de múltipla escolha."



EXERCÍCIOS

Tarefa 1: Análise Exploratória da Estatística dos Episódios

Utilizar técnica de prompt chaining para implementar uma análise exploratória dos dados de audiência, personagens e reviews dos episódios e temporadas.

Tarefa 2: Chain of Thought para criação do dashboard de interação com os atores

Usar o LLM para gerar um código Python para implementar um dashboard streamlit com a interação entre o usuários e os atores.

Tarefa 3: Generated Knowledge para gerar Insights

Usar o LLM como um cientista de dados para gerar insights de acordo com a estatística e análises realizadas na base de dados.