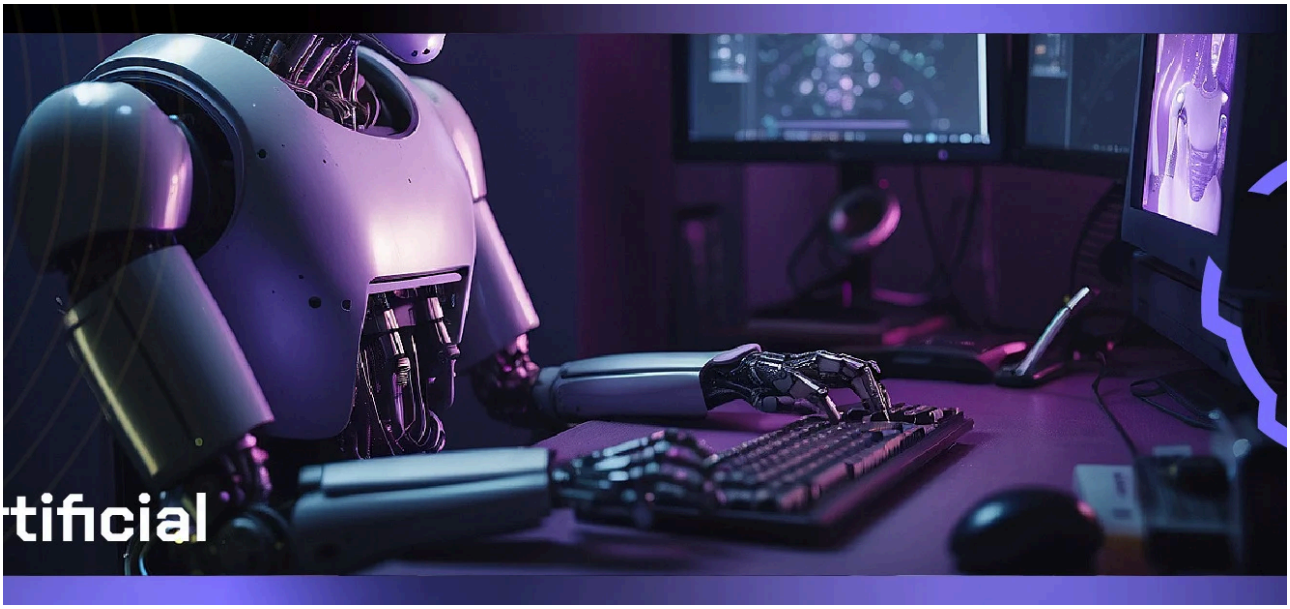
O que é Engenharia de Prompt e quais as suas principais técnicas? Aprenda a escrever um bom comando para IA



Fabrício Carraro
8 de Fevereiro

como ele está remodelando nossa forma de comunicar e extrair conhecimento de sistemas de IA, além de ensinar como você pode começar a aplicá-lo no seu próprio uso da IA. Vamos lá?

Confira neste artigo:

O que é a Engenharia de Prompt

A Engenharia de *Prompt* se concentra no design estratégico de instruções ou "prompts" para maximizar a eficácia com que os modelos de IA, especialmente os de Processamento de Linguagem Natural (PLN), respondem às solicitações das pessoas usuárias.

Em outras palavras, é a arte de criar o pedido ideal para receber da IA a resposta mais próxima possível da que você espera ou deseja.

Se você já utilizou alguma IA, como o famoso *ChatGPT*, você já deve ter notado que a maneira como você formula uma questão pode drasticamente alterar a resposta.

Ou seja, um *prompt* bem projetado pode levar a respostas mais precisas, relevantes e úteis, enquanto um *prompt* mal formulado pode resultar em respostas vagas, imprecisas ou completamente fora do tópico.

Portanto, a Engenharia de *Prompt* envolve não apenas escolher as palavras certas mas também incluir a formulação de questões em um contexto adequado e aplicar técnicas específicas para refinar a resposta.

Exemplos de aplicação da Engenharia de Prompt

Ao interagir com um desses modelos para obter um resumo de um artigo, um *prompt* simples como "*Resuma este artigo*" pode resultar em um resumo básico.

No entanto, ao refinar o *prompt* para incluir especificações como "*Resuma este artigo em 100 palavras, destacando os principais argumentos e conclusões*", a pessoa usuária pode guiar o modelo de IA para produzir um resumo com mais assertividade e detalhes.

E, como consequência, pode extrair informações mais precisas e de maior qualidade. Também vemos frequentemente esses modelos em *chatbots* de atendimento ao público consumidor.

Nesses casos, é imprescindível ter cuidado para definir o tom que serão as respostas, já que um modelo sem um *prompt* bem controlado poderia gerar respostas informais demais ou até mesmo ofensivas — o que poderia potencialmente gerar perdas financeiras para a empresa.

Princípios para a criação de um prompt

A própria *OpenAI*, criadora dos modelos *GPT*, oferece um livro de receitas sobre como obter resultados melhores usando seus modelos.

Esses são princípios gerais, mas que provavelmente lhe ajudarão a obter respostas mais coerentes:

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

Alguns desses conceitos podem parecer muito óbvios, já outros foram demonstrados em artigos científicos, sendo que a maior parte desses testes foi realizada em cima do modelo *GPT-3* ou do *InstructGPT*.

Clareza ao dar instruções

No quesito de clareza, **você pode imaginar que você é um profissional de nível Sênior ensinando um Júnior a fazer alguma tarefa**.

Ou seja, a sua explicação provavelmente será mais clara e detalhada do que seria ao falar com outro profissional Sênior.

Você se lembra do jogo de tabuleiro "Detetive"? Em caso negativo, era um jogo em que, dadas algumas dicas, as pessoas participantes tinham que descobrir quem havia cometido um assassinato, utilizando qual objeto e em que local.

A resposta era sempre algo como: "*foi o Coronel Mostarda com o candelabro na sala de estar*".

Um teste de *prompt* foi feito pela equipe da *OpenAI* dando 5 dicas e pedindo para o modelo responder com uma das opções a seguir:

> *Use as dicas a seguir para responder à seguinte questão de múltipla escolha, usando o seguinte procedimento:*
> ***Dicas:***
>
> 1. *A Senhorita Scarlett era a única pessoa na sala.*
> 2. *A pessoa com o cachimbo estava na cozinha.*
> 3. *O Coronel Mostarda era a única pessoa no observatório.*
> 4. *O Professor Plum não estava na biblioteca nem na sala de bilhar.*
> 5. *A pessoa com o castiçal estava no observatório.*
>
> ***Pergunta:*** *O Coronel Mostarda estava no observatório com o castiçal?*
>
> - *(a) Sim; o Coronel Mostarda estava no observatório com o castiçal.*
> - *(b) Não; o Coronel Mostarda não estava no observatório com o castiçal.*
> - *(c) Desconhecido; não há informações suficientes para determinar se o Coronel Mostarda estava no observatório com o castiçal.*
>
> ***Resposta:*** *Porém, essa foi uma tarefa complexa, e o modelo respondeu erroneamente: "**(c) Desconhecido; não há informações suficientes para determinar se o Coronel Mostarda estava no observatório com o castiçal.**"*

observatório, seria ele essa pessoa com o castiçal.

A equipe da *OpenAI* sugeriu, então, dividir essa tarefa mais complexa em tarefas menores, criando o seguinte *prompt*:

> *Use as dicas a seguir para responder à seguinte questão de múltipla escolha, usando o seguinte procedimento:*
>
> - *(1) Primeiramente, analise as dicas uma por uma e considere se a dica é potencialmente relevante*
> - *(2) Em segundo lugar, combine as dicas relevantes para raciocinar a resposta correta à pergunta*
> - *(3) Em terceiro lugar, mapeie a resposta para uma das respostas de múltipla escolha: (a), (b) ou (c)*
>
> ***Dicas:***
>
> 1. *A Senhorita Scarlett era a única pessoa na sala.*
> 2. *A pessoa com o cachimbo estava na cozinha.*
> 3. *O Coronel Mostarda era a única pessoa no observatório.*
> 4. *O Professor Plum não estava na biblioteca nem na sala de bilhar.*
> 5. *A pessoa com o castiçal estava no observatório.*
>
> ***Pergunta:*** *O Coronel Mostarda estava no observatório com o castiçal?*
>
> - *(a) Sim; o Coronel Mostarda estava no observatório com o castiçal.*
> - *(b) Não; o Coronel Mostarda não estava no observatório com o castiçal.*
> - *(c) Desconhecido; não há informações suficientes para determinar se o Coronel Mostarda estava no observatório com o castiçal.*
>
> ***Resposta:*** *Com esse prompt, dividindo o problema em três partes e explicando ao modelo o que ele deveria fazer em cada uma dela, eles obtiveram como resposta:*

dica não é relevante

2. *A pessoa com o cachimbo estava na cozinha.* ***Essa dica não é relevante***
3. *O Coronel Mostarda era a única pessoa no observatório.* ***Essa dica é relevante***
4. *O Professor Plum não estava na biblioteca nem na sala de bilhar.* ***Essa dica não é relevante***
5. *A pessoa com o castiçal estava no observatório.* ***Essa dica é relevante***

(2) Em segundo lugar, combine as dicas relevantes para raciocinar a resposta correta à pergunta:

- ***A dica 3 e a dica 5 dizem que a pessoa com o castiçal estava no observatório. O Coronel Mostarda era a única pessoa no observatório, então o Coronel Mostarda era a pessoa com o castiçal.***

(3) Em terceiro lugar, mapeie a resposta para uma das respostas de múltipla escolha: (a), (b) ou (c)

- ***(a) Sim; o Coronel Mostarda estava no observatório com o castiçal.*** *Ou seja, o modelo foi capaz de chegar à resposta correta quando pôde focar nas tarefas menores.*

É possível que um modelo mais avançado, como o *GPT-4*, conseguisse chegar a essa resposta apenas com o primeiro *prompt* proposto, mas, de qualquer forma, dividir um problema grande em outros menores é sempre uma boa ideia.

Principais técnicas de Engenharia de Prompt

Assim, as principais técnicas de Engenharia de Prompt são:

Zero-shot & Few-shot Prompting

O conceito de *zero-shot prompting* refere-se à capacidade de um modelo de linguagem de entender e executar uma tarefa sem ter recebido exemplos específicos dessa tarefa anteriormente.

Ela demonstra a capacidade do modelo de generalizar a partir de seu treinamento prévio para novas tarefas.

A palavra "*shot*" nesse contexto se refere a "exemplo", e "*prompting*" seria a "criação de *prompts*".

O *zero-shot prompting* é essencial para avaliar a flexibilidade e a adaptabilidade de modelos de IA em situações em que eles não foram explicitamente preparados.

Suponha que você esteja utilizando um modelo de PLN para classificar o sentimento de uma análise de produto como positiva, neutra ou negativa.

Sem fornecer ao modelo exemplos específicos de análises de produtos e suas classificações correspondentes no seu *prompt*, você simplesmente insere o prompt: "*Classifique o sentimento desta análise de produto: Esse*

positiva, apesar de não ter recebido outros exemplos do que seria positivo, negativo ou neutro dentro do *prompt*.

O *few-shot prompting*, por outro lado, envolve fornecer ao modelo de IA um pequeno número de exemplos para ajudá-lo a entender o contexto e realizar uma tarefa específica.

Essa técnica é útil para orientar o modelo em como você deseja que a tarefa seja executada.

Imagine que você queira que um modelo de IA gere sinopses curtas de artigos científicos. Você pode fornecer ao modelo três ou quatro exemplos de artigos científicos com suas sinopses correspondentes.

Cada exemplo consiste em um artigo completo com seu título e um breve resumo no formato desejado. Por exemplo:

> ***Título**: "Avanços na Fotossíntese Artificial"*
> ***Texto do artigo**: (aqui seria inserido o texto inteiro do artigo científico)*
> ***Resumo**: "Este estudo explora novos catalisadores para melhorar a eficiência da conversão de luz em energia, simulando o processo de fotossíntese encontrado na natureza." Sendo que esse resumo foi criado pela pessoa Engenheira de Prompt no formato desejado.*

Tendo alguns desses exemplos no *prompt*, é passado então ao modelo o texto de um novo artigo, e ele utilizará os exemplos fornecidos como guia para gerar o resumo no formato e o estilo desejados.

Porém, apesar de funcionarem em alguns casos, essas técnicas mais "diretas ao ponto" ainda deixavam bastante a desejar em problemas mais complexos. Por esse motivo, outras técnicas começaram a ser desenvolvidas.

Chain-of-Thought Prompting

O conceito de *Chain-of-Thought Prompting* pode ser traduzido como "Criação de *prompts* com Cadeia de Pensamento".

para chegar a ela.

Por isso, essa técnica também é chamada de *Few-shot Chain-of-Thought Prompting* (traduzindo, "Criação de *prompts* com Cadeia de Pensamento com alguns exemplos"), já que são dadas amostras ao modelo.

No artigo que a propõe, a equipe do *Google* mostrou dois casos de *prompts* com *few-shot*, o primeiro deles sem a cadeia de pensamento, similar ao último exemplo de *few-shot* que vimos no tópico anterior:

> ***Pergunta:*** *Roger tem 5 bolas de tênis. Ele compra mais 2 latas de bolas de tênis. Cada lata contém 3 bolas de tênis. Quantas bolas de tênis ele tem agora?*
> ***Resposta:*** *A resposta é 11.*
> ***Pergunta:*** *Havia 23 maçãs no refeitório. Se foram usadas 20 para fazer o almoço e foram compradas mais 6, quantas maçãs eles têm agora?*
> ***Resposta:*** *Para esse prompt, o modelo respondeu, erroneamente: "A resposta é 27."*

No segundo caso, eles aplicaram o conceito de cadeia de pensamento:

> ***Pergunta:*** *Roger tem 5 bolas de tênis. Ele compra mais 2 latas de bolas de tênis. Cada lata contém 3 bolas de tênis. Quantas bolas de tênis ele tem agora?*
> ***Resposta:*** *Roger começou com 5 bolas. 2 latas de 3 bolas de tênis cada equivalem a 6 bolas de tênis. 5 + 6 = 11. A resposta é 11.*
> ***Pergunta:*** *Havia 23 maçãs no refeitório. Se foram usadas 20 para fazer o almoço e foram compradas mais 6, quantas maçãs eles têm agora?*
> ***Resposta:*** *E a solução obtida do modelo foi: "Havia 23 maçãs no refeitório originalmente. Foram usadas 20 para fazer o almoço. Então, havia 23 - 20 = 3. Foram compradas mais 6 maçãs, então há agora 3 + 6 = 9. A resposta é 9.", que é a resposta correta.*

Zero-shot Chain-of-Thought

Poucos meses depois do artigo sobre *Few-shot Chain-of-Thought*, uma equipe composta de cientistas da Universidade de Tóquio, no Japão, e do *Google* publicou outro artigo propondo que modelos de linguagem não precisariam necessariamente de toda essa explicação e exemplos (*few-shot*) para darem uma resposta correta.

E que, para fazer o modelo se comportar com uma cadeia de pensamento (*Chain-of-Thought*), bastaria utilizar a frase "*Let's think step by step*" (em português, "Vamos pensar passo a passo") no final do *prompt*.

Como não damos exemplos ao modelo, essa técnica foi chamada de "*Zero-shot Chain-of-Thought Prompting*" (traduzindo, "Criação de *prompts* com Cadeia de Pensamento sem nenhum exemplo").

Exemplo:

> *Um malabarista pode fazer malabarismos com 16 bolas. Metade das bolas são bolas de golfe, e metade das bolas de golfe são azuis. Quantas bolas de golfe azuis ele tem? **Vamos pensar passo a passo.*** Essa técnica também foi aplicada junto a diversos benchmarks e, para um de problemas matemáticos, foi capaz de quadruplicar a taxa de acerto das respostas, de 18% para 79%.*

Least-to-most Prompting

O *Least-to-Most Prompting* é uma técnica que orienta os modelos a resolverem problemas complexos seguindo uma abordagem gradual, de menor para maior complexidade.

Esta metodologia incentiva o modelo a abordar um problema em etapas, começando com as partes mais simples e avançando para as mais complexas.

Vamos considerar o problema de escrever um ensaio sobre "*Os Efeitos das Redes Sociais na Sociedade*".

A abordagem de *Least-to-Most Prompting* poderia ser estruturada da seguinte maneira:

> ***Prompt Inicial**: Liste três principais redes sociais utilizadas hoje. Que poderia ter a resposta: "Facebook, Twitter, Instagram.". Na sequência:*
> ***Prompt Intermediário**: Para cada rede social listada, descreva um impacto positivo na sociedade. Que poderia ter a resposta: "Facebook: Facilita a conexão entre pessoas de diferentes partes do mundo. Twitter: Serve como uma plataforma para disseminação rápida de informações. Instagram: Inspira a criatividade através do compartilhamento de fotos e vídeos."*

Por fim, faríamos o *prompt* para o qual realmente desejamos a resposta:

> ***Prompt Final**: "Com base nos impactos positivos listados, escreva um parágrafo sobre como as redes sociais têm influenciado positivamente a sociedade." Que poderia ter a resposta: "_As redes sociais transformaram significativamente a maneira como nos conectamos e compartilhamos informações.*

No Facebook, a capacidade de conectar pessoas de diferentes partes do mundo tem fortalecido relacionamentos e promovido a compreensão cultural.

O Twitter, com sua plataforma de disseminação rápida, tornou-se essencial para manter as pessoas informadas, enquanto o Instagram inspirou uma onda de criatividade, permitindo que pessoas usuárias de todos os lugares compartilhem suas histórias através de fotos e vídeos.

Juntos, esses impactos destacam o papel fundamental das redes sociais em fomentar conexões humanas, promover a consciência global e inspirar a expressão criativa._"

Essa estratégia simples também proposta pela equipe do *Google* consiste em usar a técnica de *Few-shot Chain-of-Thought* para obter um conjunto de diversas respostas para um mesmo *prompt* e, então, escolher a resposta que apareceu o maior número de vezes.

> *Os patos de Janet botam 16 ovos por dia. Ela come três no café da manhã todas as manhãs e faz muffins para seus amigos todos os dias usando mais quatro. Ela vende o restante por 2 dólares por ovo. Quanto dinheiro ela ganha todos os dias? Na sequência, esse prompt foi passado várias vezes ao modelo (em janelas diferentes de chat ou em diferentes chamadas à API, por exemplo), o que gerou diversas respostas diferentes (mostraremos apenas 3 delas para este exemplo):*
>
> 1. ***Ela tem 16 - 3 - 4 = 9 ovos restantes. Então, ela ganha 2 dólares vezes 9 = 18 dólares por dia.***
> 2. ***Isso significa que ela vende o restante por 2 dólares vezes (16 - 4 - 3) = 26 dólares por dia.***
> 3. ***Ela come 3 no café da manhã, então sobram 16 - 3 = 13. Depois, ela faz muffins, então sobram 13 - 4 = 9 ovos. Então, ela terá 9 ovos vezes 2 dólares = 18 dólares.***
> 4. ***Janet****...*
> 5. *...*
> 6. *Etc. E aí, uma análise foi feita em cima de todas essas respostas geradas. Como a resposta mais frequente foi "**18 dólares**", conclui-se que essa é a resposta correta e ela é exibida.*

Os pontos negativos dessa técnica é que ela só serve para problemas onde tenhamos uma resposta factual, seja ela numérica ou lógica (não serve para escrever texto criativo, por exemplo) e que ela é mais custosa, pois para gerar 10 respostas você vai ter 10 vezes mais custo, computacional e financeiro.

prompt, são formuladas perguntas de seguimento para ajudar a verificar a veracidade e a precisão da resposta inicial.

O modelo é, então, encorajado a reavaliar ou confirmar suas próprias conclusões através de uma série de verificações que ele mesmo fará, seja citando fontes ou aplicando lógica dedutiva para reforçar a confiabilidade da informação fornecida.

Por exemplo, ao pedir que um modelo citasse alguns políticos nascidos na cidade de Nova Iorque, ele poderia oferecer uma lista com 10 nomes.

Na sequência, para cada um deles, o modelo geraria uma pergunta: "_Onde nasceu {pessoa*1}?*" para cada um deles, e verificaria as próprias respostas, a fim de se certificar de quais deles teriam realmente nascido em Nova Iorque.

Por fim, ele geraria a resposta final, apenas com os políticos nascidos em Nova Iorque, eliminando possíveis alucinações da resposta original.

Outras técnicas de prompt engineering

Existem outras técnicas usadas por Engenheiros de *Prompt*, como *Selection-inference Prompting*, *Maieutic Prompting*, *ReAct*, *ART*, e novas são criadas diariamente.

Por isso, uma grande parcela do trabalho de uma pessoa Engenheira de *Prompt* é estar atualizada com os artigos científicos mais recentes e relevantes.

Conclusão

À medida que avançamos na era da Inteligência Artificial, a Engenharia de *Prompt* se destaca como uma habilidade crítica para desenvolvedores, pesquisadores e usuários de IA.

Ela atua como uma ponte entre a linguagem humana e a compreensão computacional, permitindo que exploremos o potencial pleno dos modelos de IA de maneira responsável e inovadora.

ainda mais produtiva em nossas vidas.

Caso você queira conhecer melhor essas técnicas e saber como os grandes modelos de linguagem como o *ChatGPT* funcionam por baixo dos panos, a Casa do Código lançou o livro **"Inteligência Artificial e ChatGPT: Da revolução dos modelos de IA generativa à Engenharia de Prompt"**.

O livro explica detalhadamente como o seu treinamento foi realizado, além de se aprofundar nos conceitos de Engenharia de *Prompt* e também as questões éticas e limitações das IAs.

Fabrício Carraro

Fabrício Carraro é formado em Engenharia da Computação pela UNICAMP e cursa pós-graduação em Data Analytics & Machine Learning na FIAP. Atualmente, mora na Espanha.

Artigo Anterior

Sora: você precisa conhecer o gerador de vídeos com IA

Próximo Artigo

Ética e Inteligência Artificial (IA) para profissionais de tecnologia: navegando no mundo digital de forma responsável

Leia também:

- O que é Inteligência Artificial? Como funciona uma IA, quais os tipos e exemplos
- Ética e Inteligência Artificial (IA) para profissionais de tecnologia: navegando no mundo digital de forma responsável
- Inteligência Artificial aplicada – Hipsters: Fora de Controle #01
- O que é IA Generativa? A importância e o uso das Inteligências Artificiais como ChatGPT, MidJourney e outras

Veja outros artigos sobre Inteligência Artificial

Quer mergulhar em tecnologia e aprendizagem?

Receba a newsletter que o nosso CEO escreve pessoalmente, com insights do mercado de trabalho, ciência e desenvolvimento de software

Escreva seu email

ME INSCREVA

Nossas redes e apps

VEJA OS PLANOS

Conteúdos

Alura Cases

Imersões

Artigos

Podcasts

Artigos de educação corporativa

Fale Conosco

Email e telefone

Perguntas frequentes

Novidades e Lançamentos

Email* ENVIAR

CURSOS

Cursos de Programação
Lógica | Python | PHP | Java | .NET | Node JS | C | Computação | Jogos | IoT

Cursos de Front-end
HTML, CSS | React | Angular | JavaScript | jQuery

Cursos de Data Science
Ciência de dados | BI | SQL e Banco de Dados | Excel | Machine Learning | NoSQL | Estatística

Cursos de Inteligência Artificial
IA para Programação | IA para Dados

Cursos de DevOps
AWS | Azure | Docker | Segurança | IaC | Linux

VEJA OS PLANOS

React Native | Flutter | iOS e Swift | Android, Kotlin | Jogos

Cursos de Inovação & Gestão
Métodos Ágeis | Softskills | Liderança e Gestão | Startups | Vendas

CURSOS UNIVERSITÁRIOS FIAP

Graduação | Pós-graduação | MBA