3. Engenharia de Prompt

Davi Bezerra Barros

Definição

Engenharia de prompt é o estudo das melhores formas de criar prompts para solicitar que uma LLM resolva uma ou mais tarefas, e segue o princípio de que as máquinas estão aprendendo cada vez melhor a nossa língua, ao invés de nós a dela. As instruções dadas às llms devem ser claras diretas, assim como devem ser as instruções para um ser humano. A **regra de ouro** é: Como deve ser dito para alguém realizar uma tarefa? o prompt deve estruturado de forma semelhante.

P.R.O.M.P.T

Estrutura de prompt apresentada no vídeo:

1. **Persona:** Personalidade que o modelo deve adotar, para guiar sua linha de raciocínio a resolver a tarefa naquele contexto.
2. **Roteiro**: Roteiro da tarefa a ser realizada, diz ao modelo o que ele deve fazer.
3. **Objetivo**: Objetivo da tarefa a ser realizada, o que se pretende alcançar com ela.
4. **Modelo:** Descrição do formato esperado para o resultado
5. **Panorama:** Descreve o contexto da tarefa, com exemplos e detalhes.
6. **Transforma**r: Feedbacks ao modelo, o direcionando para chegar no resultado ideal.

Processo

As etapas utilizadas para refinar e construir um prompt mais robustos são:

1. Definir a tarefa, e os critérios de sucesso;
2. Desenvolver casos de teste;
3. Escrever o prompt inicial;
4. Testar o prompt contra exemplos;
5. Refinar o promt;
6. colocar em produção;

Técnicas

Markdown: Escrever o texto em formato markdown para o modelo entender melhor a estrutura do prompt. O modelo também responderá em formato markdown.

XML: Utilizar um header XML para separar os elementos do prompt.

Prompt de Sistema: utilizar a seção de prompts do sistema para fornecer contextos e mais informações ao modelo antes de pedi-lo para realizar uma tarefa. Serve para o modelo compreender melhor a entrada do usuário.

Zero-Shot: Técnica em que se faz um prompt sem nenhum exemplo prévio para demonstrar o formato ou tipo de saída esperada.

Prompt de Estimulo Direcional: É um a técnica de engenharia de prompt que além das instruções, o usuário dá dicas de como a resposta deve ser estruturada.

Few-Shot: Técnica de engenharia de prompt onde se fornece alguns exemplos de como o modelo deve responder.

Chain of Thought: Técnica de engenharia de prompt onde o modelo é instruído a "pensar em voz alta", mostrando o raciocínio que utilizou para chegar à resposta.

Contrasting Chain of Thought: Técnica avançada de engenharia de prompt onde o usuário gera múltiplas cadeias de raciocínio divergentes, para estimular o modelo a refletir e compará-las.

Self Consistency: O modelo é solicitado a gerar várias cadeias de raciocínio independentes, e escolher a resposta comum entre elas, por meio de uma "votação".

Tree of Thought: Abordagem de prompt que simula a tomada de decisão humana, estruturada em forma de árvore. Ao invés de seguir em uma única linha de raciocínio, o modelo explora várias opções, e avalia qual é a melhor. É útil para explorar problemas com múltiplas soluções possíveis.

Skeleton of Thought: O skeleton-of-thought separa o processo de geração em duas fases;

1. O modelo gera um equeleto das idéias principais, como um roteiro.
2. O roteiro é expandido, desenvolvendo cada ponto com detalhes.
   Este processo ajuda o modelo a manter a coesão e organização, principalmente para respostas mais longas.

Generated Knowledge Prompting for Commonsense Reasoning: O objetivo desta técnica é melhorar o raciocínio do modelo, gerando conhecimento antes de responder. o modelo primeiro gera fatos ou suposições relacionados ao problema, e em seguida, utiliza isso para resolve-lo.

Maiueutic Prompting: O objetivo desta técnica é fazer com que o modelo quebre o raciocínio em afirmações que serão verificadas uma a uma. O modelo gera as afirmações parciais relacionadas a resposta, e cada afirmação é avaliada antes de continuar. Isto reduz falhas lógicas.

Geração Aumentada de Recuperação: Nesta técnica, o usuário busca combinar o prompt com uma base de documentos externos, para que o modelo se baseie nisso ao gerar sua resposta.

React: Synergizing Reasoning and Acting in Language Models: Combina as capacidades de reasoning, onde o modelo "pensa" em voz alta, e as ações que ele está performando, interagindo com o ambiente para resolver problemas. A interação do pensamento com a ação simula o comportamento de um agente cognitivo.

Como evitar alucinações**

Para evitar que o modelo alucine, é crucial informar ao modelo que ele deve avisar quando não souber a resposta para alguma pergunta ou não conseguir performar uma tarefa, para que ele não preencha as lacunas com informação fictícia. Outra estratégia eficiente é pedir para o modelo encontrar citações nas informações que foram dadas a ele, e construir a resposta em cima disso. Com estas duas técnicas, a resposta passa a ser mais confiável, reduzindo consideravelmente o risco de gerar conteúdos fictícios.

Conclusão

A engenharia de prompt é uma prática fundamental para extrair o máximo de desempenho das LLMs, fazendo com que suas respostas tenham maior precisão. As técnicas apresentadas na aula como Chain of Thought, ReAct ou Geração Aumentada por Recuperação melhoram os resultados gerados pelo modelo ao moldar o seu comportamento para se alinhar com os objetivos do usuário. Em suma, a Engenharia de Prompt serve como uma ponte entre a linguagem humana e a inteligência artificial.