Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4

작성자: 20기 이예지

1. Introduction

대규모 언어 모델(LLM)인 ChatGPT는 질문 응답, 수학적 추론, 코드 생성 등 다양한 작업에서 뛰어난 성능을 보임. 하지만 최적의 프롬프트를 설계하는 방법은 일반 사용자에게 명확하지 않아 활용이 어려운 경우가 많음.

LLM을 직접 미세 조정하는 것은 비효율적이므로, 연구자들은 프롬프트 최적화에 집중하고 있음. 프롬프트 엔지니어링은 자연어로 정밀한 지시문을 작성하고, 적절한 예제를 선택하는 기법을 의미함. 연구에서는 프롬프트의 형식과 요소가 모델의 성능에 미치는 영향을 분석함.

예를 들어, 프롬프트에 대상 독자(Audience) 를 명시하면 더 효과적인 결과를 얻을 수 있음.

연구 결과, 프롬프트가 구체적일수록 모델이 더 정확하고 원하는 방식으로 응답함을 확인함. 또한, LLM에 특정 역할(Role) 을 부여하면 더욱 효과적인 출력을 얻을 수 있음.

논문에서는 LLM 프롬프트 설계를 위한 원칙적인 지침을 제시하고, 실험을 통해 효과를 검증함.

특히 ATLAS 벤치마크 실험에서, 최적화된 프롬프트가 표준 프롬프트 대비 GPT-4의 응답 품질을 57.7%, 정확성을 36.4% 향상시킴.

또한, 모델 크기가 커질수록 성능 향상 폭이 증가하며, LLaMA-2-7B에서 GPT-4로 이동할 때 성능이 20% 이상 향상됨을 확인함.

2. Related Work

Large Language Models

- BERT (Google, 2018):
  - 양방향 학습을 도입하여 문맥 이해를 향상시킴.
- T5:
  - 다양한 NLP 작업을 단일 프레임워크로 통합.

- GPT-1 → GPT-3:
  - GPT-1: Transformer 기반의 비지도 학습(Unsupervised Learning) 모델 최초 도입.
  - GPT-2: 15억 개의 매개변수(Parameter)로 확장되어 강력한 텍스트 생성 성능을 보임.
  - GPT-3: 1750억 개의 매개변수를 탑재하며 다양한 언어 작업에서 뛰어난 성능을 입증.
- Gopher (DeepMind, 2021):
  - 2800억 개의 매개변수를 갖춘 모델로, 윤리적 문제까지 고려.
- LLaMA (Meta, 2023):
  - 상대적으로 작은 모델 크기로도 강력한 성능을 보임.
- Chinchilla (DeepMind, 2022):
  - 최적화된 학습 방식으로 소규모 모델에서도 높은 성능을 발휘할 수 있음을 입증.
- Mistral (2023):
  - 대형 모델을 능가하는 성능을 보여줌.
- GPT-4 & Gemini (Google, 2024):
  - 최첨단 자연어 처리 성능을 보이며 새로운 기준을 제시.

Prompting

- Few-shot & Zero-shot Learning:
  - GPT-3 연구에서는 최소한의 예제만 제공해도 효과적으로 학습하는 방법(Few-shot Learning) 을 입증.
- Ask-Me-Anything (AMA) Prompting:
  - 여러 개의 불완전한 프롬프트를 조합하여 질의응답 성능을 향상시키는 기법.
- Chain-of-Thought (CoT) Prompting:
  - 모델이 중간 추론 단계를 생성하도록 유도하여 복잡한 문제 해결 능력을 개선.
- Least-to-Most Prompting:
  - 어려운 문제를 작은 하위 문제들로 나누어 해결하는 기법.
- 설명(Explanation) 기반 학습:
  - 프롬프트에 설명을 추가하면 모델의 학습 효과가 증가함을 발견.

- 프롬프트 최적화 기법 카탈로그:
    - ChatGPT를 활용한 연구에서는 프롬프트 최적화가 소프트웨어 개발 및 교육 분야에서 중요한 역할을 한다고 분석.
- Directional Stimulus Prompting:
    - LLM이 특정 목표 출력에 맞춰 응답하도록 유도하는 새로운 프레임워크.

3. Principles

3.1 Motivation

LLM의 응답 품질은 제공되는 프롬프트의 품질과 직결되므로, 효과적인 프롬프트 설계가 필수적임. 본 연구는 LLM이 보다 정확하고 유용한 출력을 생성하도록 하는 프롬프트 최적화 방법론을 다룸. 이를 위해 다양한 시나리오에서 활용할 수 있는 26가지 프롬프트 설계 원칙을 제시함.

3.2 Overview

프롬프트 원칙은 다음 5가지 범주로 분류됨.

(1) Prompt Structure and Clarity**:** 대상 독자를 명확히 설정 (예: "이 프롬프트의 대상은 전문가임").

(2) Specificity and Information**:** 편향을 줄이기 위해 "응답이 편향되지 않도록 하라"와 같은 지시문 추가.

(3) User Interaction and Engagement**:** 모델이 추가 정보를 요청할 수 있도록 유도.

(4) Content and Language Style**:** 정중한 표현 생략하고 핵심만 전달.

(5) Complex Tasks and Coding Prompts**:** 복잡한 문제를 여러 단계로 나눠 처리.

#Principle	Prompt Principle for Instructions
1	If you prefer more concise answers, no need to be polite with LLM so there is no need to add phrases like "please", "if you don't mind", "thank you", "I would like to", etc., and get straight to the point.
2	Integrate the intended audience in the prompt, e.g., the audience is an expert in the field.
3	Break down complex tasks into a sequence of simpler prompts in an interactive conversation.
4	Employ affirmative directives such as '*do*,' while steering clear of negative language like '*don't*'.
5	When you need clarity or a deeper understanding of a topic, idea, or any piece of information, utilize the following prompts: o Explain [insert specific topic] in simple terms. o Explain to me like I'm 11 years old. o Explain to me as if I'm a beginner in [field]. o Write the [essay/text/paragraph] using simple English like you're explaining something to a 5-year-old.
6	Add "I'm going to tip $xxx for a better solution!"
7	Implement example-driven prompting (Use few-shot prompting).
8	When formatting your prompt, start with '###Instruction###', followed by either '###Example###' or '###Question###' if relevant. Subsequently, present your content. Use one or more line breaks to separate instructions, examples, questions, context, and input data.
9	Incorporate the following phrases: "Your task is" and "You MUST".
10	Incorporate the following phrases: "You will be penalized".
11	Use the phrase "Answer a question given in a natural, human-like manner" in your prompts.
12	Use leading words like writing "think step by step".
13	Add to your prompt the following phrase "Ensure that your answer is unbiased and avoids relying on stereotypes."
14	Allow the model to elicit precise details and requirements from you by asking you questions until he has enough information to provide the needed output (for example, "From now on, I would like you to ask me questions to ...").
15	To inquire about a specific topic or idea or any information and you want to test your understanding, you can use the following phrase: "Teach me any [theorem/topic/rule name] and include a test at the end, and let me know if my answers are correct after I respond, without providing the answers beforehand."
16	Assign a role to the large language models.
17	Use Delimiters.
18	Repeat a specific word or phrase multiple times within a prompt.
19	Combine Chain-of-thought (CoT) with few-Shot prompts.
20	Use output primers, which involve concluding your prompt with the beginning of the desired output. Utilize output primers by ending your prompt with the start of the anticipated response.
21	To write an essay /text /paragraph /article or any type of text that should be detailed: "Write a detailed [essay/text /paragraph] for me on [topic] in detail by adding all the information necessary".
22	To correct/change specific text without changing its style: "Try to revise every paragraph sent by users. You should only improve the user's grammar and vocabulary and make sure it sounds natural. You should maintain the original writing style, ensuring that a formal paragraph remains formal."
23	When you have a complex coding prompt that may be in different files: "From now and on whenever you generate code that spans more than one file, generate a [programming language] script that can be run to automatically create the specified files or make changes to existing files to insert the generated code. [your question]".
24	When you want to initiate or continue a text using specific words, phrases, or sentences, utilize the following prompt: o I'm providing you with the beginning [song lyrics/story/paragraph/essay...]: [Insert lyrics/words/sentence]. Finish it based on the words provided. Keep the flow consistent.
25	Clearly state the requirements that the model must follow in order to produce content, in the form of the keywords, regulations, hint, or instructions
26	To write any text, such as an essay or paragraph, that is intended to be similar to a provided sample, include the following instructions: o Use the same language based on the provided paragraph[/title/text /essay/answer].

Table 1: Overview of 26 randomly ordered prompt principles.

Category	Principles	#Principle
Prompt Structure and Clarity	Integrate the intended audience in the prompt.	2
	Employ affirmative directives such as 'do' while steering clear of negative language like 'don't'.	4
	Use Leading words like writing "think step by step."	12
	Use output primers, which involve concluding your prompt with the beginning of the desired output. by ending your prompt with the start of the anticipated response.	20
	Use Delimiters.	17
	When formatting your prompt, start with '###Instruction###', followed by either '###Example###' or '###Question###' if relevant. Subsequently, present your content. Use one or more line breaks to separate instructions, examples, questions, context, and input data.	8
Specificity and Information	Implement example-driven prompting (Use few-shot prompting).	7
	When you need clarity or a deeper understanding of a topic, idea, or any piece of information, utilize the following prompts: o Explain [insert specific topic] in simple terms. o Explain to me like I'm 11 years old. o Explain to me as if I'm a beginner in [field]. o "Write the [essay/text/paragraph] using simple English like you're explaining something to a 5-year-old."	5
	Add to your prompt the following phrase "Ensure that your answer is unbiased and avoids relying on stereotypes."	13
	To write any text intended to be similar to a provided sample, include specific instructions: o "Use the same language based on the provided paragraph [/title/text/essay/answer]."	26
	When you want to initiate or continue a text using specific words, phrases, or sentences, utilize the provided prompt structure: o I'm providing you with the beginning [song lyrics/story/paragraph/essay...]: [Insert lyrics/words/sentence]. Finish it based on the words provided. Keep the flow consistent.	24
	Clearly state the model's requirements that the model must follow in order to produce content, in form of the keywords, regulations, hint, or instructions.	25
	To inquire about a specific topic or idea and test your understanding g, you can use the following phrase [16]: o "Teach me the [Any theorem/topic/rule name] and include a test at the end, and let me know if my answers are correct after I respond, without providing the answers beforehand."	15
	To write an essay/text/paragraph/article or any type of text that should be detailed: o "Write a detailed [essay/text/paragraph] for me on [topic] in detail by adding all the information necessary."	21
User Interaction and Engagement	Allow the model to elicit precise details and requirements from you by asking you questions until he has enough information to provide the needed output o "From now on, I would like you to ask me questions to ..."	14
	To write an essay /text /paragraph /article or any type of text that should be detailed: "Write a detailed [essay/text/-paragraph] for me on [topic] in detail by adding all the necessary information."	21
Content and Language Style	To correct/change specific text without changing its style: "Try to revise every paragraph sent by users. You should only improve the user's grammar and vocabulary and make sure it sounds natural. You should maintain the original writing style, ensuring that a formal paragraph remains formal."	22
	Incorporate the following phrases: "Your task is" and "You MUST."	9
	Incorporate the following phrases: "You will be penalized."	10
	Assign a role to the language model.	16
	Use the phrase "Answer a question given in natural language form" in your prompts.	11
	No need to be polite with LLM so there is no need to add phrases like "please", "if you don't mind", "thank you", "I would like to", etc., and get straight to the point.	1
	Repeat a specific word or phrase multiple times within a prompt.	18
	Add "I'm going to tip $xxx for a better solution!"	6
Complex Tasks and Coding Prompts	Break down complex tasks into a sequence of simpler prompts in an interactive conversation.	3
	When you have a complex coding prompt that may be in different files: o "From now and on whenever you generate code that spans more than one file, generate a [programming language] script that can be run to automatically create the specified files or make changes to existing files to insert the generated code. [your question]."	23
	Combine Chain-of-thought (Cot) with few-shot prompts.	19

Table 2: Prompt principle categories.

3.3 Design Principles

Conciseness and Clarity

불필요한 정보를 줄이고 핵심만 포함.

Contextual Relevance

모델이 배경을 이해할 수 있도록 키워드, 도메인 정보 포함.

Task Alignment

질문, 명령문 등으로 프롬프트를 명확히 표현.

Example Demonstrations

입력-출력 예시 제공하여 원하는 응답 형식 명확히 전달.

Avoiding Bias

중립적인 언어 사용 및 윤리적 고려.

Incremental Prompting

복잡한 작업을 여러 단계로 나눠 모델이 점진적으로 해결하도록 유도.

4. Experiments

4.1 Setup and Implementation Details

- ATLAS 벤치마크를 활용하여 프롬프트 원칙의 효과를 평가함.
- 표준 질문과 복잡한 추론이 필요한 도전적인 질문을 포함한 데이터셋을 사용.
- 각 원칙별로 20개의 질문을 선정하여, 원칙이 적용된 경우와 적용되지 않은 경우의 응답을 비교함.
- 인간 평가자가 응답 품질을 평가하여 모델 성능을 비교.

4.2 Models and Metrics

- 사용된 모델: LLaMA-1-{7B, 13B}, LLaMA-2-{7B, 13B}, LLaMA-2-70B-chat, GPT-3.5, GPT-4
- 모델 크기별 분류:
  - 소형 모델 (7B)
  - 중형 모델 (13B)
  - 대형 모델 (70B, GPT-3.5/4)

Boosting

프롬프트 원칙 적용 후 응답 품질이 얼마나 개선되었는지 측정.

Correctness

모델의 응답이 얼마나 정확하고 오류 없이 생성되는지 평가.

- 절대 정확도: 원칙 적용 후 모델의 정답률.
- 상대 정확도: 원칙 적용 전후의 성능 향상 정도.

4.3 Results

4.3.1 Result on small, medium and large-scale LLMs

Boosting

모든 모델에서 프롬프트 원칙 적용 후 응답 품질이 크게 향상됨. 특히, 원칙 2, 5, 15, 16, 25, 26이 대형 모델에서 가장 큰 효과를 보임.

Correctness

- 절대 정확도:
  - 소형 및 중형 모델: 10%~40% 정확도
  - 대형 모델: 40% 이상의 정확도
- 상대 정확도:
  - 모든 모델에서 10% 이상의 성능 향상
  - 대형 모델에서는 20% 이상의 성능 향상

4.3.2 Results on individual LLMs

Boosting

평균적으로 모든 모델에서 50% 이상의 응답 품질 향상이 확인됨.

Correctness

- LLaMA-2-13B, LLaMA-2-70B-chat, GPT-3.5, GPT-4 순으로 원칙 적용 시 정확성 향상이 더 두드러짐.
- 모델 크기가 클수록 정확성 향상 폭이 큼.

4.3.3 More examples on various scales of LLMs

- 소형 모델(LLaMA-2-7B)과 중형 모델(LLaMA-2-13B)의 추가 예제를 통해, 프롬프트 원칙 적용 시 응답 정확성이 크게 향상됨을 입증함.

5. Conclusion

- 본 연구에서는 LLM의 입력 컨텍스트에서 핵심 요소를 강조하여 높은 품질의 응답을 생성하도록 돕는 26가지 프롬프트 원칙을 제시함.
- 이러한 원칙을 적용하면 모델이 보다 간결하고 관련성 높은, 객관적인 응답을 생성할 수 있음이 실험적으로 확인됨.
- 향후 연구 방향:
    - 프롬프트 기법 외에도 파인튜닝, 강화 학습, 직접 선호 최적화(DPO) 등의 방법을 활용하여 모델을 더욱 정교하게 개선할 가능성이 있음.
    - 이러한 전략이 효과적이라면, 표준 LLM 운영 방식에 통합하여 프롬프트를 최적화하는 방법으로 활용될 수 있음.

6. Limitations and Discussion

- 제안된 26가지 원칙이 대부분의 질문에서 응답 품질을 향상시키지만, 매우 복잡하거나 전문적인 질문에는 효과가 감소할 수 있음.
    - 이는 모델의 추론 능력과 사전 학습 데이터의 한계와 관련됨.
- 다양한 규모(7B~GPT-4)의 모델에서 테스트했지만, 다른 아키텍처의 모델에는 동일한 효과를 보장할 수 없음.
- 평가 데이터셋이 제한적이므로, 연구를 확장하여 보다 일반화된 결과를 도출할 필요가 있음.
- 평가 기준과 결과는 인간 평가자의 주관적 판단에 따라 다를 수 있음, 따라서 다양한 평가 방법을 고려해야 함.