Applying Large Language Models to Interactive Information Retrieval: A Practical Exploration

Part II: Prompt Strategies

ADM+S ZUVA RMIT UNIVERSITY

From https://forgottenrealms.fandom.com/wiki/Conjuration?file=Conjuration.png

It Works Until It Doesn’t

- Re-wording prompts and/or specifying information differently can be an effective way to bend an LLM to your will
- Except despite our best efforts, it refuses to follow instructions as we want
- There has been a bunch of work on other techniques beyond prompt engineering
- Some methods may work better than others depending on the task and the model

Zero-Shot/One-Shot/Few-Shot/Multi-Shot

- **Zero-shot:** The model is asked to perform a task without being given any specific examples of that task in the prompt. It relies on its pre-trained knowledge.
- **One-shot:** The model is given a single example of the desired input-output format along with the instruction or query.
- **Few-shot (or Multi-shot)**: The model is provided with multiple (but typically a small number) of examples demonstrating the task and the expected input-output patterns.

Method	Description	Example
Zero-shot	No examples provided	**Prompt:** "Classify the sentiment of this review: 'I love this product!'" **LLM Output:** "Positive"
One-shot	One example is provided to guide	**Prompt:** "Classify the sentiment of this review: 'This product is amazing!' → Positive. Now classify: 'This is terrible!'" **LLM Output:** "Negative"
Few-shot	Few examples (~4?) are provided to *see* patterns	**Prompt:** "Classify the sentiment of the following reviews: 'Great service!' → Positive 'Too expensive.' → Negative 'Very useless staff!' → Negative Now classify: 'The product stopped working after a week.' " **LLM Output:** "Negative"
Many-shot	Large number of examples are provided to *reinforce* patterns	Few-shot with many more examples

Chain-of-Thought (CoT)

- Premise: LLMs often try to “go” directly to an answer. So encourage model to “think” about the solution.
- Can be repetitively done.
- Zero-shot: Add “Let’s think step by step” or similar to prompt.
- Multi-shot: Provide examples showing related problems being solved with reasoning.

If one is good, why not more?

- Many extensions to CoT that essentially involve generating more CoT reasoning paths and selecting the answer using them.
- Self-Consistency: Generate several CoT responses using few-shot CoT and then take a majority vote.
- Tree of Thoughts: A further extension that uses tree search algorithms and additional steps to have the LLM evaluate whether a CoT should be refined or abandoned.

Prompt Chaining

- Premise: Take a large task and break it into smaller, independent tasks
- Derived from CoT but makes the steps discrete.
- LLMs respond more reliably when given simple, unambiguous instructions
- This style of prompting also forms a basis for “agentic” workflows

Structured Generation

- Because you really, really want JSON the first time
- You can prompt models to respond with JSON
  - It works, sometimes.
- It works better if you specify the schema (and it is a larger model)
- Libraries exist to support this (Pydantic) and you get output like:

```
Return your response as a JSON array with each element taking the following format:
“‘json
{
"bucket" : character // Respond with the character representing the appropriate bucket.
"explanation" : string // The reasoning behind why this option was selected.
} “‘
```

Structured Generation

- Many libraries to help “guide” the model
  - Outlines
  - Guidance
  - Can be used with local and commercial models
- Some commercial models support this explicitly
  - OpenAI has a couple variants: JSON Mode (old) and Structured Generation
  - Gemini also has its own version

Promptimization

- There is an active body of research trying to automatically optimize prompts
- Libraries like **dspy** attempt to do this relatively painlessly
  - But your mileage may vary
- An alternative is to prompt an LLM to improve your prompt (or suggest one) for your task.
  - Can be surprisingly effective.
  - Some LLMs do this well (e.g., ChatGPT) but it may result in fragile prompts that only work on some models

Caveats

- Prompts are LLM-specific
- Interacting directly with the LLM can be problematic, try to use libraries
  - You need to know the instruction format and this can be onerous
- Lots of different strategies for prompting and more on the way
  - No single strategy is universally superior
  - https://www.promptingguide.ai/ has a good summary of techniques

Coopetition 2

Summarising and interacting

https://bit.ly/CHIIR_LLM

Table 2: Search tasks in the user study.

Domain	Task Description
Environment	What are the characteristics of pollution particulate matter in China? Your answer should cover its compositions, its time-varying patterns, and its geographical characteristics.
	Why ultraviolet disinfection cannot completely supplant chlorination when disinfecting drinking water? And what are the advantages and disadvantages of them?
Medicine	What are the most commonly-used methods for cancer treatment in clinics?
	What are the potential applications of 3D printing for “Precision Medicine”?
Politics	Political scientists have noted that the trend of political polarization during the US presidential election is increasingly evident. What are the reasons behind it?
	In order to achieve their own interests, what kind of strategies do the US interest groups often take?

Query

Why ultraviolet disinfection cannot completely supplant chlorination when disinfecting drinking water? And what are the advantages and disadvantages of them?

From Liu, Mengyang, et al. "" Satisfaction with Failure" or" Unsatisfied Success" Investigating the Relationship between Search Success and User Satisfaction." Proceedings of the 2018 world wide web conference. 2018.