6. Title: Further Readings

ⓘ This page is OPTIONAL.

The “Further Readings” section **provides additional resources** related to the topics we explore each week.

Please note that these resources are optional and are **intended to enhance your learning and understanding**. There is no obligation for you to review them all.

You may choose to revisit these resources at your convenience. They are designed to serve as valuable references for your continued learning journey.

Table of Contents

Base Language Models vs. Instruction Tuned Language Models

Large language models (LLMs) are a category of foundation models trained on immense amounts of data making them capable of understanding and generating natural language and other types of content to perform a wide range of tasks.

Base Language Models

A Base Language Model (Base LLM or also known as pre-trained LLM) represents the fundamental model obtained after initial AI training on a broad dataset. For instance, models like GPT-3 or BERT fall under this category. They are trained on large volumes of internet text to understand and predict language patterns. However, they don't inherently follow instructions provided in the prompt.

Base LLM predicts the next word, based on the text training data:

Python

```
prompt = "Once  upon a time, there was a unicorn..."
#[Out] that lived in a magical forest with all unicorn friends
```

Now, a downside of this is that if you were to prompt it with "What is the capital of France?", it is quite possible that there might be a list of quiz questions about France somewhere on the internet that the model has already been trained on. So it may complete this with "What is France's largest city, what is France's population?", and so on. But what you really want is you want is for it to tell you what is the capital of France, probably, rather than list all these questions.

Python

```
promtp = "Where is the capital of France?"
#[Out] What is France's largest city?
#      What is France's population?
#      What is the currency of France?
```

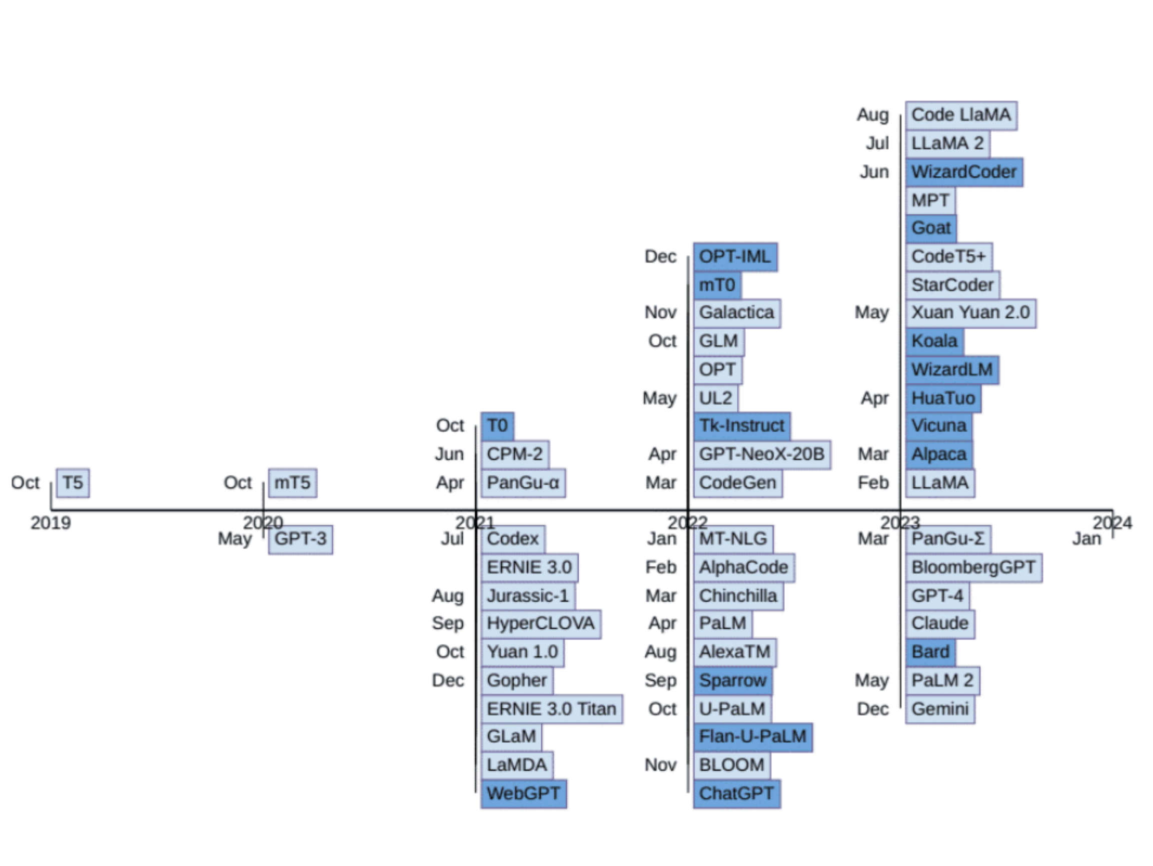
Instruction-tune Language Model

✦ An Instruction-tuned Language Model, on the other hand, undergoes an **additional round of training on a narrower dataset**, specifically designed to fine-tune its performance. This secondary training aims to enable the model to better understand and respond to specific instructions provided in the prompt.

✦ During this secondary training, the model is being fine-tuned on examples of where the output follows an input instruction.

- The process may also involve obtaining human-ratings of the quality of different outputs from the LLM, based on criteria such as whether it is helpful, honest, and harmless.
- The human-rated data is used to further tune the LLM to increase the probability for the model to generate more highly rated outputs.
- This process is called **Reinforcement Learning from Human Feedback (RLHF)**. If you're interested, you can find out more about RLHF from here .

✦ With an Instruction-tuned LLM, the model will be much more likely to follow instructions closely. This enhanced responsiveness to instructions can lead to more accurate, efficient, and satisfactory outcomes for users.



Light blue rectangles represent 'pre-trained' models, while dark rectangles correspond to 'instruction-tuned' models. Models on the upper half signify open-source availability, whereas those on the bottom half are closed-source.

source: Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435).

Works & Research on Prompt Engineering

Survey of Prompting Use-cases

✦ Large Language Models (LLMs) can perform a wide range of tasks using zero-shot prompting techniques. Research argues these abilities emerge in LLMs at a certain scale in terms of parameter size.

✦ LLMs have been applied in academic contexts for tasks such as knowledge probing, information extraction, question answering, text classification, natural language inferenceLinks to an external site. , dataset generation , and more.

Infrastructure for Prompt Engineering

✦ Prompt engineering, a relatively new concept, requires new interfaces for application development.

- We discussed about "scientists and enthusiasts can only experiment with different prompts, trying to make models perform better." in 2. Prompt Engineering.
- Here we include several example projects have been released to facilitate easier prompt design.
    - Bach and Sanh et al. built *PromptSource*, an integrated development environment to systematize and crowdsource best practices for prompt engineering.
    - Strobelt et al. developed *PromptIDE*, a visual platform to experiment with prompt variations, track prompt performance, and iteratively optimize prompts.
    - Wu et al. proposed *PromptChainer*, a tool to design multi-step LLM applications, tying together not just prompting steps but also external API calls and user inputs.

Prompt Engineering Security

✦ An interesting and concerning phenomenon in building LLM applications is the appearance of prompt-based security exploits .

✦ By leveraging carefully-crafted inputs, LLMs can reveal the “secret” prompts they use in the backend and leak credentials or other private information, similar to SQL injection attacks.

✦ Currently, there are no robust mechanisms to address this issue. We will look into this topic in a subsequent session of this training.

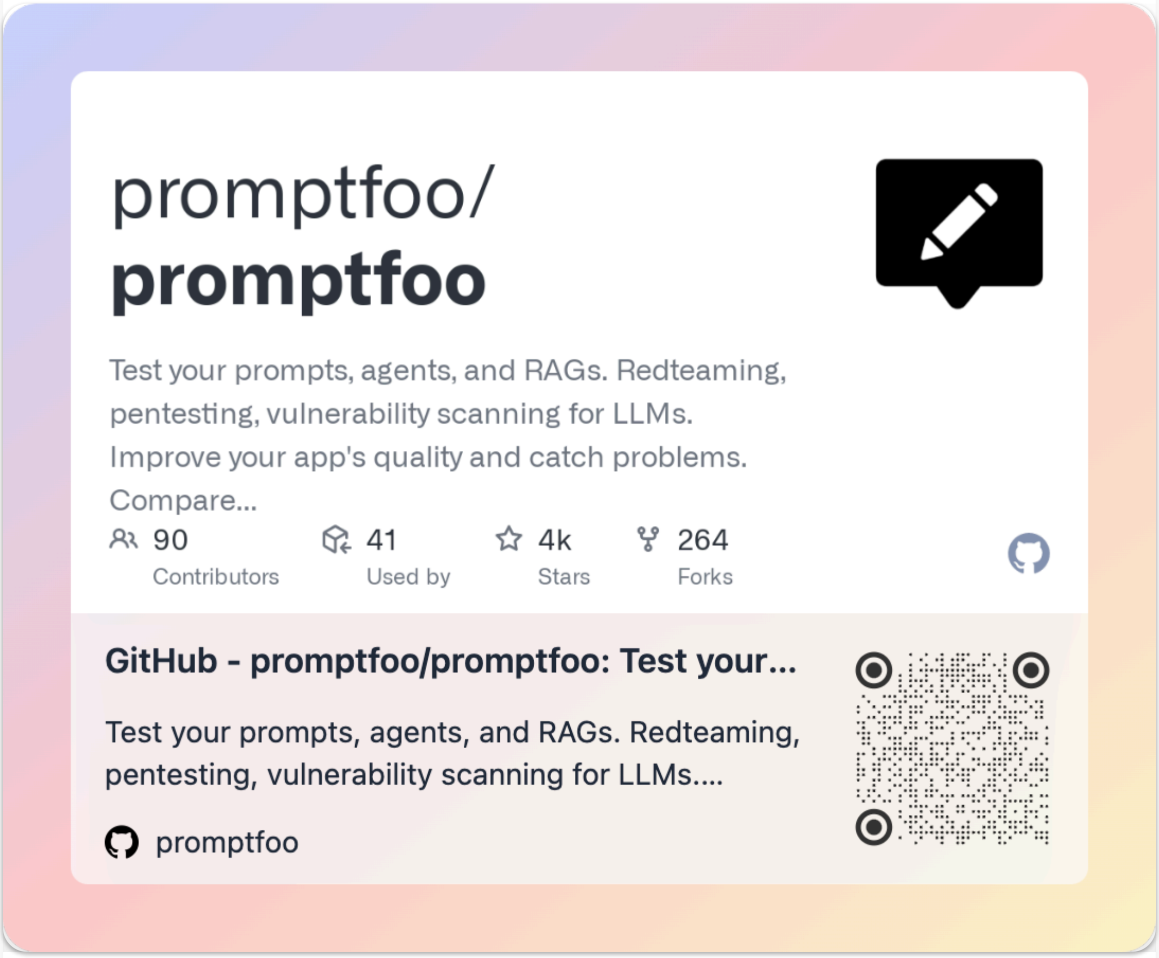
✦ Workarounds using different formatting of the inputs have been proposed, but more work needs to be done to prevent these vulnerabilities, especially as LLMs increasingly power more functionality for future use-cases.

Other Useful Resources

Repositories/Tools Useful for Prompt Engineering

Priomptfoo

`promptfoo` is a tool for testing, evaluating, and red-teaming LLM apps. With promptfoo, you can Build reliable prompts with benchmarks specific to your use-case



promptfoo/
promptfoo

Test your prompts, agents, and RAGs. Redteaming, pentesting, vulnerability scanning for LLMs. Improve your app's quality and catch problems. Compare...

90 Contributors 41 Used by 4k Stars 264 Forks

GitHub - promptfoo/promptfoo: Test your...

Test your prompts, agents, and RAGs. Redteaming, pentesting, vulnerability scanning for LLMs....

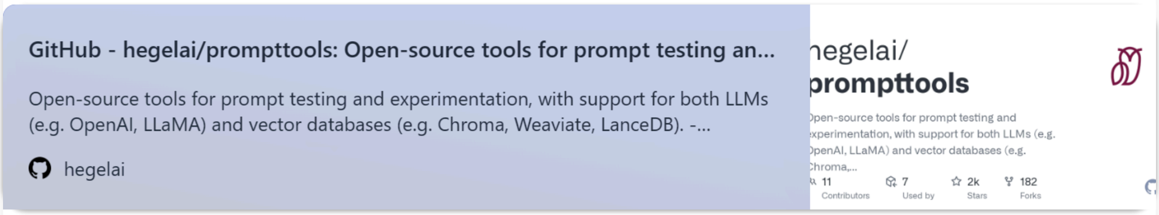
promptfoo

OPEN

PromptTools

This repo offers a set of open-source, self-hostable tools for experimenting with, testing, and evaluating LLMs, vector databases, and prompts. The core idea is to enable developers to evaluate using familiar interfaces like *code*, *notebooks*, and a local *playground*.

In just a few lines of code, you can test your prompts and parameters across different models (whether you are using OpenAI, Anthropic, or LLaMA models).



GitHub - hegelai/prompttools: Open-source tools for prompt testing an...

Open-source tools for prompt testing and experimentation, with support for both LLMs (e.g. OpenAI, LLaMA) and vector databases (e.g. Chroma, Weaviate, LanceDB). -...

hegelai

hegelai/
prompttools

Open-source tools for prompt testing and experimentation, with support for both LLMs (e.g. OpenAI, LLaMA) and vector databases (e.g. Chroma,...

11 Contributors · 7 Used by · 2k Stars · 182 Forks

OPEN

Awesome ChatGPT Prompts

This is a collection of prompt examples to be used with the ChatGPT model.

GitHub - f/awesome-chatgpt-prompts: This repo includes ChatGPT...

This repo includes ChatGPT prompt curation to use ChatGPT better. - f/awesome-chatgpt-prompts

f

f/**awesome-chatgpt-prompts**

This repo includes ChatGPT prompt curation to use ChatGPT better.

78 Contributors · 19 Used by · 102k Stars · 14k Forks

OPEN

Relevant Courses for Deepening the Understanding

✦ Find out the evolution of LLMs and how it works

Learning about Large Language Models -...

Join Jonathan Fernandes for an in-depth discussion in this video, Learning about Large Language Models, part of...

in Jonathan Fernandes

OPEN