Making Sense of LLMs 

Anish Nayanapalli
anayanapalli@gmail.com

Agenda

- What are LLMs?
- How to use them?
- Why be careful with them?
- Q&A

Let’s get started

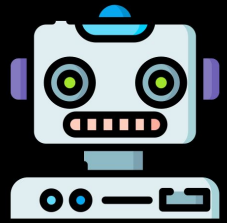
What are Large Language Models?

> *“Large”* - Language Models

What are Language Models?

LMs assign probabilities to word sequences and find the *most likely next word.*

Context: The sky is [BLANK]
Answer: More likely blue than any other word



Language Model

Generates text

The sky is blue

You might be thinking…

Predictive text; tap a suggestion to apply.

The difference is *scale.*

What do I mean?

An LLM is a *“large”* LM and is trained on *enormous* data.

An average person reads ~700 books in a lifetime. Chat GPT was trained on over **10 million** books in a few months 🤯

In short…

An LLM is a next ~~*word*~~ token predictor. A *very very* good one.

**token = part of a word*

Some Popular LLMs…

ChatGPT

Gemini

BARD AI

How do we use these things 🤔

Unfortunately…

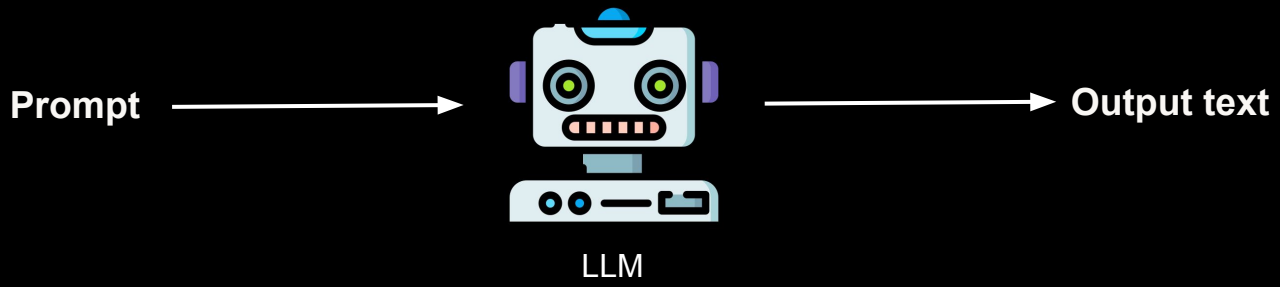
- They don’t come with a manual
- Getting them to do *“exactly”* what you want them to do is not easy
- More of an *art* than a science

For best results

- *Understanding* of how these models work
- *Domain knowledge* of the task you want to get done
- *Experience* gained by playing around with these models

Prompts

Inputs or *queries* to LLMs to do “stuff.”



“Stuff”

- Summarization
- Sentiment analysis
- Translation
- Text classification
- Text generation
- …

Prompts can be

- Natural language sentences
- Questions
- Code snippets
- Commands
- Emojis
- …

...basically any text!

A *good* prompt comprises

- Instructions
- Context
- Input/question
- Output type/format

Example

Instructions: *Write a creative and engaging short story.*

Context: *You are a detective investigating a mysterious crime in a futuristic city.*

Input/Question: *Describe the initial scene of the crime and introduce the main characters. Include unexpected twists to capture the reader's attention.*

Output Type/Format: *A narrative paragraph with vivid details and plot development.*

Instead of…

Tell me a great crime story.

A few more tips…


Hey Google, play us some music please?
Why are you being so polite?
Just in case.
KEEP THAT ONE ALIVE.
HE ALWAYS SAID "PLEASE".
seebangnow

~~*Less is more*~~, more is more

A well-thought out, articulate, clear, specific prompt - similar response.

A vague, sloppy, lazy prompt - similar response. *GIGO.*

Words you use *matter!*

Using the *“right”* word(s) for a specific task, more likely to give you a better response.

“Incantation” to do your task.

Don’t argue!

If you are not getting useful responses, try to *fix the prompt early* than continuing the conversation and arguing with the LLM.

Bad responses are more likely to be followed by similar responses.

Give *time* to think

Break down the task into multiple *steps* that the model can work on *incrementally*.

Step 1: *…*

Step 2: *…*

…

Step N: *…*

Use words like *step-by-step* that force the model to think incrementally.

Why should we be careful?

They *hallucinate!*

What do you mean?

> The generated content is *nonsensical* or *untrue* to the original data.

They are very good at generating a.k.a *making things up.*

Intrinsic vs. *extrinsic* hallucination

Intrinsic

Output contradicts the source.

Source:

The first Ebola vaccine was approved by the FDA in 2019, five years after the initial outbreak in 2014.

Output:

The first Ebola vaccine was approved in 2021.

Extrinsic

Cannot verify output from the source, but might not be wrong.

Source:

Alice won first prize in running last week.

Output:

Alice won first prize in running last week and *she was ecstatic*.

But *why?*

- Because of data that it is trained on, *non-factual* information, *duplicate* data…
- Just the *nature* of generative tasks, a *next token predictor* remember?

How to *mitigate?*

One approach is to *feed appropriate context* to the LLM before it answers.

Stay tuned for the *RAG (Retrieval Augmented Generation)* talk *next week.*

Let’s learn together

Q&A