There are 2 types of LLMs

1. Base LLM - predicts next word, based on training text data
    - It predicts the next words/questions, instead of answering questions
    - E.g.: if we ask it, “what is the capital of Germany?”, it will not answer Berlin, instead it will autocomplete and generate a new set of questions relating to Germany.

2. Instruction tune LLM - fine-tune on instructions and good attempts at following instructions
    - E.g.: What is the capital of Georgia (country)?
    - Answer: the capital of Georgia is Tbilisi.
    - It further improves by using RLHF (reinforcement learning with human feedback)

Principles of Prompting

- Write clear and specific instructions
    1. Tactic 1: use delimiters
    2. Tactic 2: ask for structured output - HTML, JSON etc
    3. Check whether conditions are satisfied and check assumptions required to do the task
    4. Tactic 4: use Few-shot prompts: give successful examples of completing tasks then ask the model to perform the task.

- GIve the model time to think
    1. Tactic 1: specify the steps to complete a task:
       Step 1:.....
       Step 2:.....
       Step N:.....

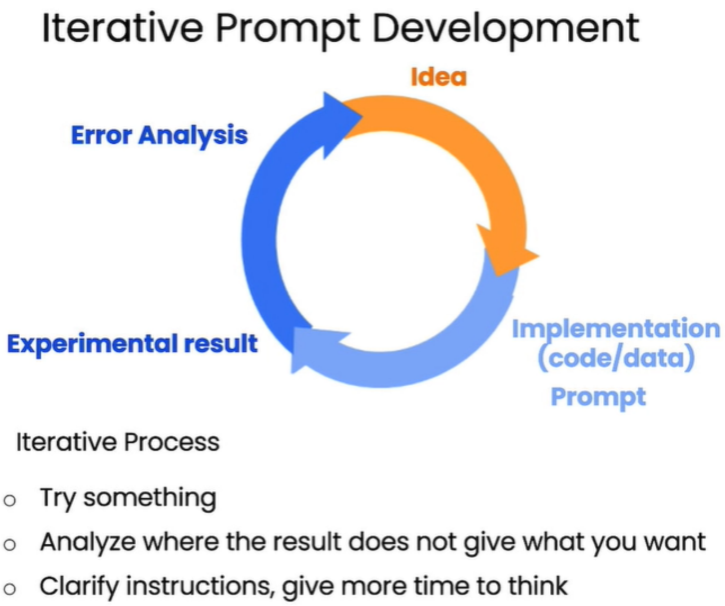
    2. Tactic 2: Instruct the model to work out its own solution before rushing to a conclusion

Model weakness

- Sometimes it will fabricate data that sounds descriptive, i.e. it will give hallucinations

Iterative Engineering

- An LLM is very literal and it will produce information only as per what the user has fed them. Sometimes, it may generate texts which are too long to read and might not be able to get the message across.
- It becomes important to be super specific. Iterative prompt engineering is the engineering of prompt in an iterative manner so that the output is exactly or almost exactly what we had expected the LLM to produce.



Iterative Prompt Development

Iterative Process

- Try something
- Analyze where the result does not give what you want
- Clarify instructions, give more time to think

Credit: DeepLearning AI

Summarizing

- This could be useful to summarize texts such as reviews and could be sent to the respective departments concerned with the specifics of the review/complaint/feedback.

Inferring

- Useful for sentiment analysis and questions surrounding reviews or to get top highly discussed topics in an article. etc. sort of like reference-to-context based questions.

Transforming

- Useful for translation and grammar/spell checks
- Also could be used to change from one data format to another (json to html e.g.)
- Another use could be to convert the text from a paragraph to something like a text in the APA format of paper writing (there are many other formats available)

Expanding

- Based on small text, generate a whole paragraph or article or email.