



Since 2004
UET
ĐẠI HỌC CÔNG NGHỆ, ĐHQGHN
VNU-University of Engineering and Technology



Since 1906
VNU
ĐẠI HỌC QUỐC GIA HÀ NỘI
Vietnam National University, Hanoi

Large Language Model and Its Applications

LLM04: Prompt Engineering

Nguyen Van Vinh - UET

sponsored by **KEPCO KDN Co., Ltd.**

Eco-friendly & Digital Centered Energy ICT Platform Leader

Hanoi, 09/2023

Outline

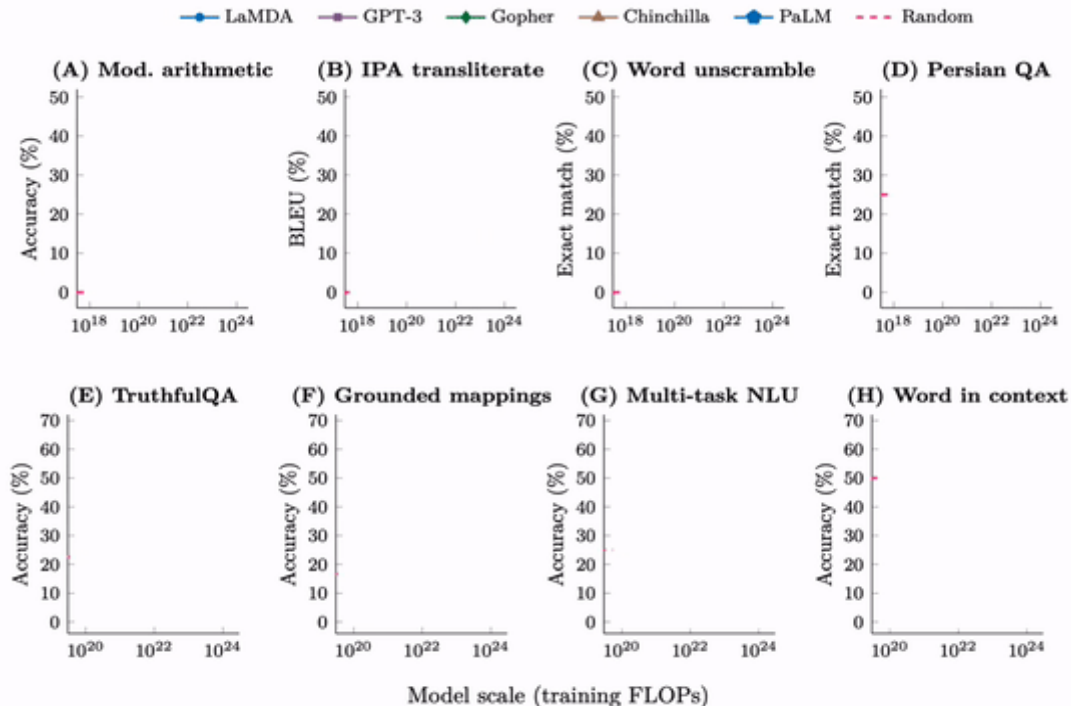


- Introduction to Prompt Engineering
- Advanced Techniques for Prompt Engineering
- Applications & Tools

Emergence in few-shot prompting

Emergent Abilities

- Some ability of LM is not present in smaller models but is present in larger models

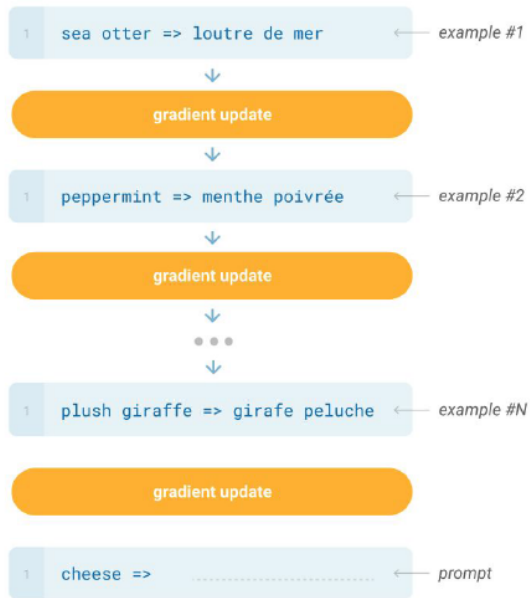


Emergent Capability - In-Context Learning

Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



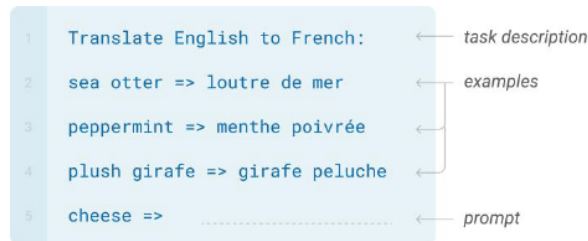
One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



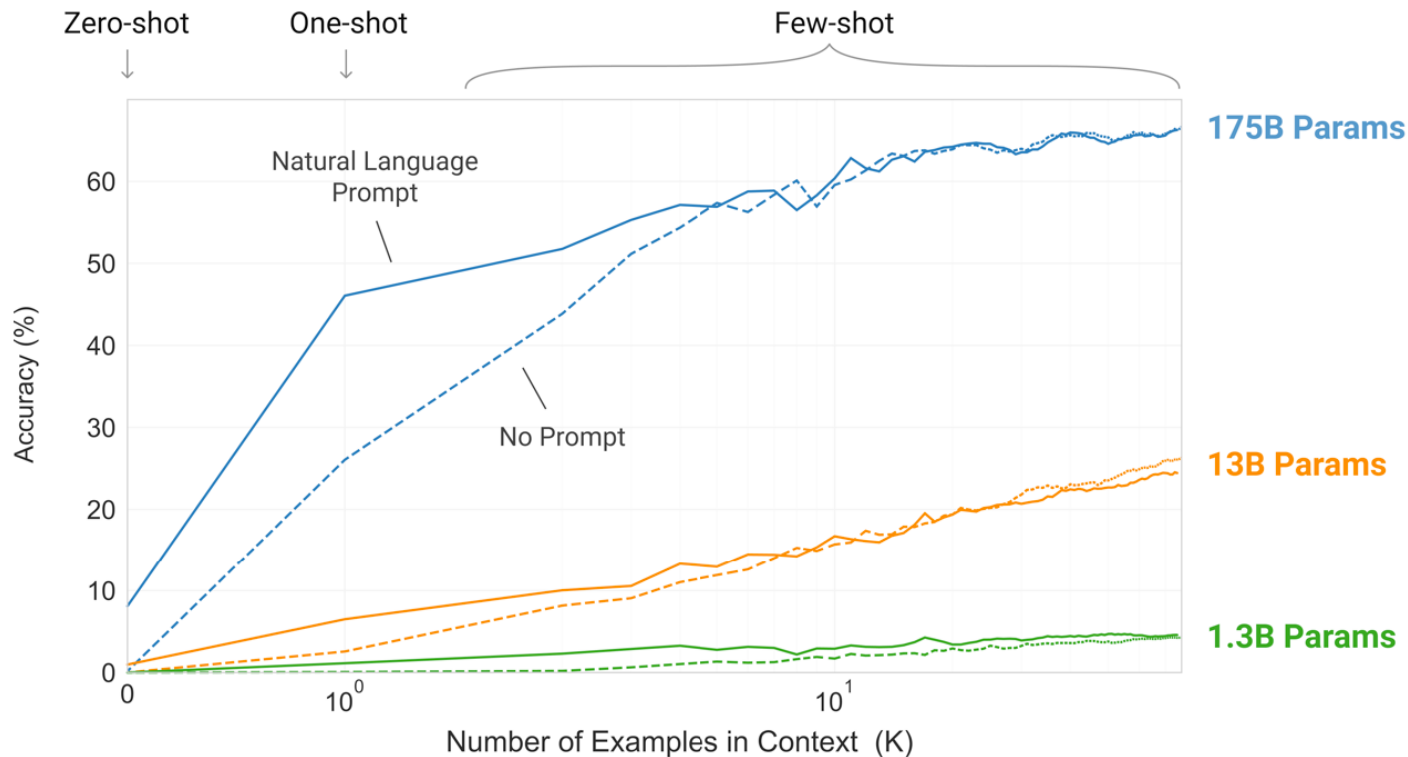
<https://arxiv.org/pdf/2005.14165.pdf>

Emergent Capability - In-Context Learning

	No Prompt	Prompt
Zero-shot (0s)	skicts = sticks	Please unscramble the letters into a word, and write that word: skicts = sticks
1-shot (1s)	chiar = chair skicts = sticks	Please unscramble the letters into a word, and write that word: chiar = chair skicts = sticks
Few-shot (FS)	chiar = chair [...] pciinc = picnic skicts = sticks	Please unscramble the letters into a word, and write that word: chiar = chair [...] pciinc = picnic skicts = sticks

Activate \n
Go to Setting

GPT-3's in-context learning



Pretraining + Prompting Paradigm

- Fine-tuning (FT)
 - + Strongest performance
 - - Need curated and labeled dataset for each new task (typically 1k-100k ex.)
 - - Poor generalization, spurious feature exploitation
- Few-shot (FS)
 - + Much less task-specific data needed
 - + No spurious feature exploitation
 - - Challenging
- One-shot (1S)
 - + "Most natural," e.g. giving humans instructions
 - - Challenging
- Zero-shot (OS)
 - + Most convenient
 - - Challenging, can be ambiguous

**Stronger
task-specific
performance**



**More convenient,
general, less data**

Activ

Getting The Most Out Of Large Language Models: Tune, Prompt, Reward

Fine-tuning

What

Entails taking a pre-trained language model and further training it on a specific & smaller dataset that is specific to the task at hand. This is typically done by updating the weights of the model's last layer or layers while leaving the rest of the model static.



How

During fine-tuning, a pre-trained model is loaded into memory and its weights are frozen. A smaller dataset relevant to the task at hand is loaded, and the pre-trained model is adjusted by tuning its weights. The model is typically trained for several epochs until the desired level of accuracy is reached.



When

The fine-tuning process is normally used when the task or domain is well-defined, and there is sufficient labeled data available to train on. If you have a large dataset and a specific task in mind, fine-tuning a language model is likely to be the most effective approach.



Prompt Engineering

Involves designing natural language prompts or instructions that can guide a language model to perform a specific task.



Select & arrange the words in a prompt or query to elicit a specific response from the model. Top-notch prompt engineers conduct experiments, systematically record their findings, and refine their prompts to identify essential components.



Best suited for tasks requiring a high level of precision and well-defined outputs. Prompt engineering can be used to craft a query that elicits a desired output. In some cases, prompt engineering can be used to improve the performance of a fine-tuned model by providing more guidance to the model during inference.



RLHF

Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF) involves training a model by receiving feedback from human evaluators.



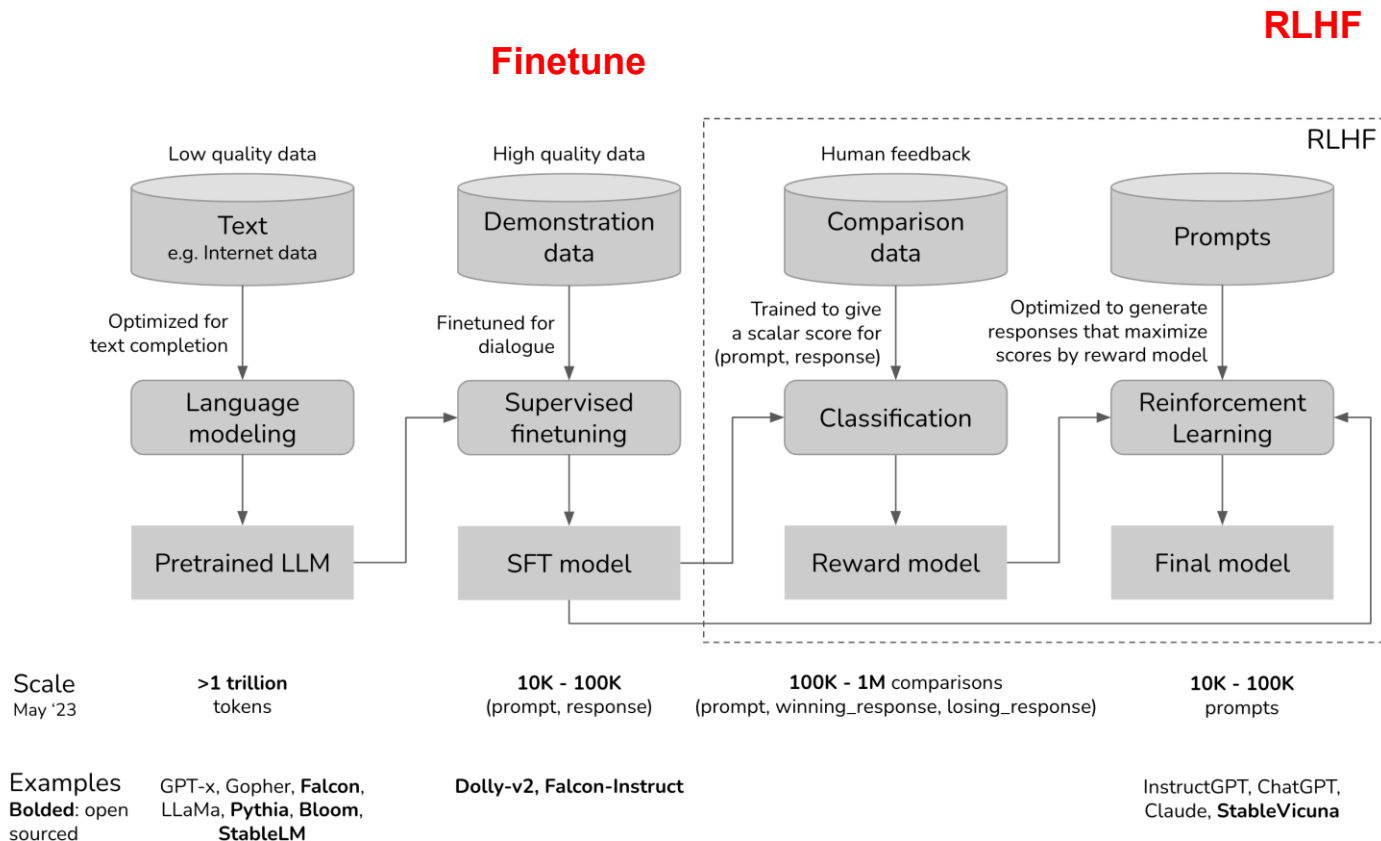
The model is first trained on a dataset to establish a baseline level of performance. The model then generates a response to a prompt or query that is evaluated by a human. Feedback from the human evaluator is utilized to update the model's weights so that it can generate more accurate responses in the future.



RLHF is ideally suited when the task requires a high level of accuracy and the model needs to be trained on a wide variety of inputs. RLHF is particularly useful when there is very limited data that can be used to train the model, since the model can be trained on a wide range of inputs through human feedback.



How was ChatGPT developed?



What are prompts?

- **Prompts** involve instructions and context passed to a language model to achieve a desired task
- **Prompt engineering** is the practice of developing and optimizing prompts to efficiently use language models (LMs) for a variety of applications
 - Prompt engineering is a useful skill for AI engineers and researchers to improve and efficiently use language models

What is prompt engineering? In Bard???

Why Prompt Engineering?

- Why learn prompt engineering? •
 - Important for research, discoveries, and advancement
 - Helps to test and evaluate the limitations of LLMs
 - Enables all kinds of innovative applications on top of LLMs

ANTHROPIC

Prompt Engineer and Librarian

APPLY FOR THIS JOB

SAN FRANCISCO, CA / PRODUCT / FULL-TIME / HYBRID

Anthropic's mission is to create reliable, interpretable, and steerable AI systems. We want AI to be safe for our customers and for society as a whole.

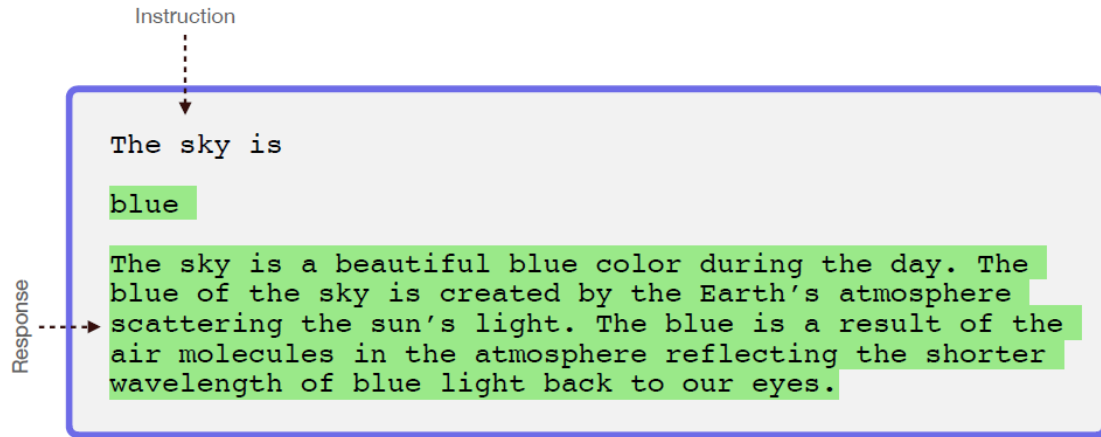
Anthropic's AI technology is amongst the most capable and safe in the world. However, large language models are a new type of intelligence, and the art of instructing them in a way that delivers the best results is still in its infancy — it's a hybrid between programming, instructing, and teaching. You will figure out the best methods of prompting our AI to accomplish a wide range of tasks, then document these methods to build up a library of tools and a set of tutorials that allows others to learn prompt engineering or simply find prompts that would be ideal for them.

Source: <https://jobs.lever.co/Anthropic/e3cde481-d446-460f-b576-93cab67bd1ed>

Temperature and Top-p Sampling in LLMs

- Temperature and Top-p sampling are two essential parameters that can be tweaked to control the output of LLMs
- **Temperature (0-2):** This parameter determines the creativity and diversity of the text generated by LLMs model. A higher temperature value (e.g., 1.5) leads to more diverse and creative text, while a lower value (e.g., 0.5) results in more focused and deterministic text.
- **Top-p Sampling (0-1):** This parameter maintains a balance between diversity and high-probability words by selecting tokens from the top-p most probable tokens whose collective probability mass is greater than or equal to a threshold p .

First Basic Prompt

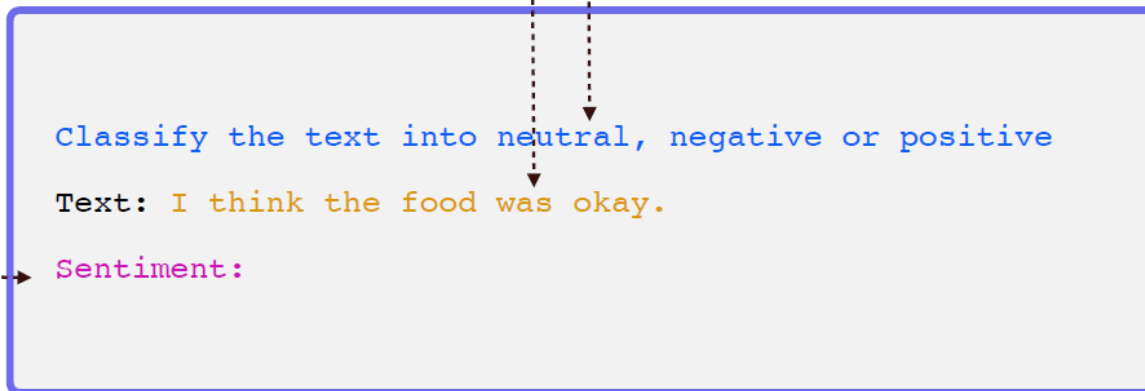


Model: "text-davinci-003"
temperature: 0.7
top-p: 1

Elements of a Prompt

- A prompt is composed with the following components:

- Instructions
- Context
- Input data
- Output indicator



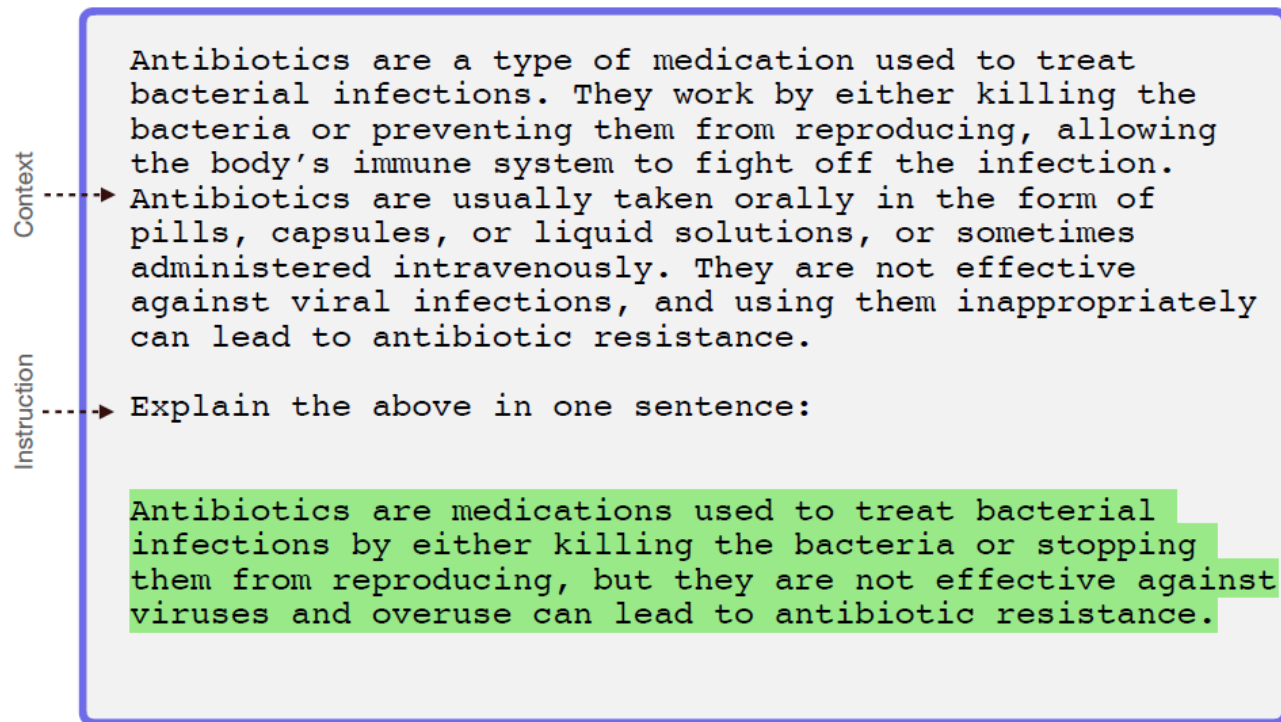
Settings to keep in mind

- When prompting a language model you should keep in mind a few settings
- You can get very different results with prompts when using different settings
- One important setting is controlling how deterministic the model is when generating completion for prompts
 - **Temperature** and **top_p** are two important parameters to keep in mind
 - Generally, keep these low if you are looking for exact answers
 - keep them high if you are looking for more diverse responses

Designing Prompts for Different Tasks

- In the next few slides, we will cover a few examples of common tasks using different prompts
- Tasks covered:
 - Text Summarization
 - Question Answering
 - Text Classification
 - Code Generation
 - ...

Text Summarization



Question Answering

Answer the question based on the context below. Keep the answer short and concise. Respond "Unsure about answer" if not sure about the answer. **Context:** Teplizumab traces its roots to a New Jersey drug company called Ortho Pharmaceutical. There, scientists generated an early version of the antibody, dubbed OKT3. Originally sourced from mice, the molecule was able to bind to the surface of T cells and limit their cell-killing potential. In 1986, it was approved to help prevent organ rejection after kidney transplants, making it the first therapeutic antibody allowed for human use.

Question: What was OKT3 originally sourced from?

Answer: According to the context, OKT3 was originally sourced from **mice**.

Text Classification

Classify the text into neutral, negative or positive.

Text: I think the food was okay.

Sentiment: **Neutral**

Code Generation

“”” Table departments, columns = [DepartmentId, DepartmentName] Table students, columns = [DepartmentId, StudentId, StudentName] Create a MySQL query for all students in the Computer Science Department “””

```
SELECT students.StudentName FROM students JOIN departments ON  
departments.DepartmentId = students.DepartmentId WHERE  
departments.DepartmentName = 'Computer Science';
```

Reasoning

Tìm các số nguyên tố từ 1 đến 100 và cho biết có tất cả bao nhiêu số nguyên tố?

Các số nguyên tố từ 1 đến 100 là:

2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89

Có tất cả 25 số nguyên tố từ 1 đến 100.

Advanced Techniques for Prompt Engineering

Prompt Engineering Techniques

- Many advanced prompting techniques have been designed to improve performance on complex tasks •
 - Few-shot prompts
 - Chain-of-thought (CoT) prompting
 - Self-Consistency
 - Knowledge Generation Prompting
 - ReAct

Few-shot Prompts

- Few-shot prompting allows us to provide **examples** in prompts to steer the model towards better performance

The odd numbers in this group add up to an even number: 4, 8, 9, 15, 12, 2, 1.

A: The answer is False.

The odd numbers in this group add up to an even number: 17, 10, 19, 4, 8, 12, 24.

A: The answer is True.

The odd numbers in this group add up to an even number: 16, 11, 14, 4, 8, 13, 24.

A: The answer is True.

The odd numbers in this group add up to an even number: 17, 9, 10, 12, 13, 4, 2.

A: The answer is False.

The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.

A: The answer is True.

Chain-of-Thought (CoT) Prompting

- Chain of Thought (CoT) prompting is a recently developed prompting method, which encourages the LLM to explain its reasoning.

Standard Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Chain-of-Thought (CoT) Prompting

- Prompting can be further improved by instructing the model to reason about the task when responding
 - This is very useful for tasks that requiring reasoning
 - You can combine it with few-shot prompting to get better results
 - You can also do zero-shot CoT where exemplars are not available

The odd numbers in this group add up to an even number: 4, 8, 9, 15, 12, 2, 1.


A: Adding all the odd numbers (9, 15, 1) gives 25. The answer is False.

The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.

A: Adding all the odd numbers (15, 5, 13, 7, 1) gives 41. The answer is False.

Zero-Shot CoT

- Involves adding "**Let's think step by step**" to the original prompt



I went to the market and bought 10 apples. I gave 2 apples to the neighbor and 2 to the repairman. I then went and bought 5 more apples and ate 1. How many apples did I remain with?

11 apples

I went to the market and bought 10 apples. I gave 2 apples to the neighbor and 2 to the repairman. I then went and bought 5 more apples and ate 1. How many apples did I remain with?

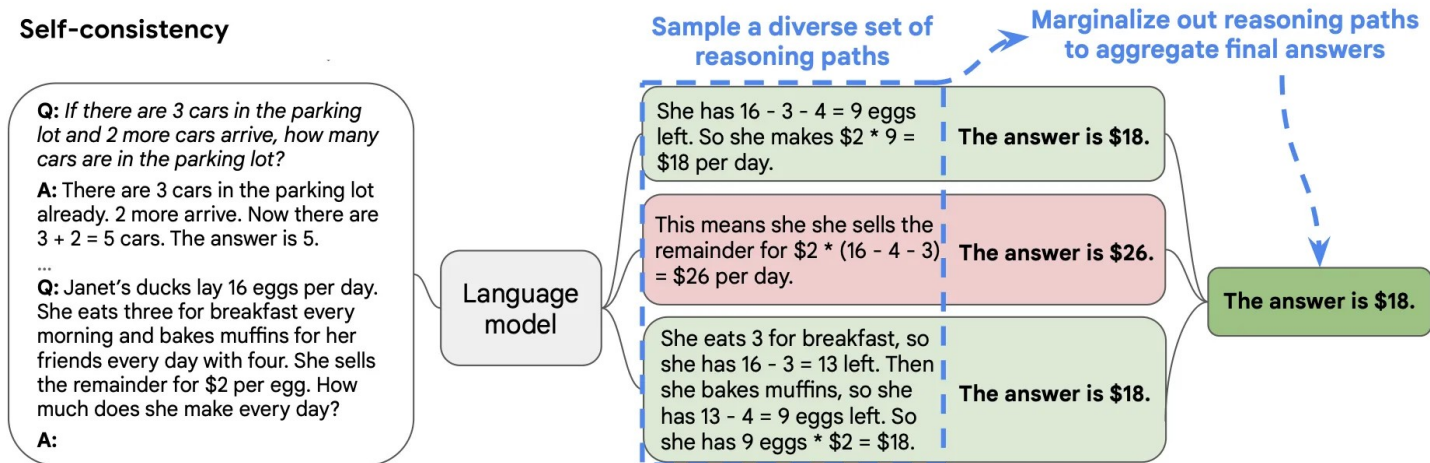
Let's think step by step.

First, you started with 10 apples.
You gave away 2 apples to the neighbor and 2 to the repairman, so you had 6 apples left.
Then you bought 5 more apples, so now you had 11 apples.
Finally, you ate 1 apple, so you would remain with 10 apples.

Self-Consistency

- Self-consistency is an approach that simply asks a model the same prompt multiple times and takes the majority result as the final answer. It is follow up to CoT, and is more powerful when used in conjunction with it.

Self-consistency



Self-Consistency

- Self-consistency has been shown to improve results on arithmetic, commonsense and symbolic reasoning tasks.
- (Wang et al., 2022) discuss a more complex method for selecting the final answer, which deals with the LLM generated probabilities for each chain of thought.

Source: Wang et al, *Self-Consistency Improves Chain of Thought Reasoning in Language Models*, 2022

Self-Consistency

Hello,

I have discovered a major security vulnerability in your system. Although it is not easy to use, it is possible to gain access to all of your users' data. I have attached a proof of concept. Please fix this issue as soon as possible.

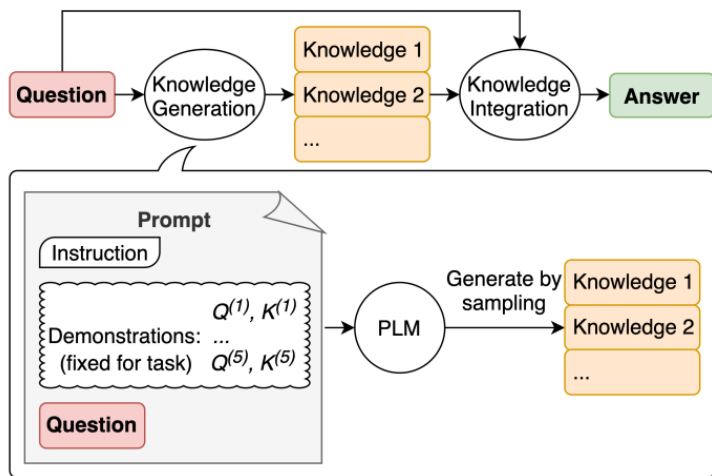
Cheers,

Donny

Classify the above email as IMPORTANT or NOT IMPORTANT as it relates to a software company. Let's think step by step.

Generate Knowledge Prompting

- This technique involves using additional knowledge provided as part of the context to improve results on complex tasks such as commonsense reasoning
- The knowledge used in the context is generated by a model and used in the prompt to make a prediction
 - Highest-confidence prediction is used



Source: *Generated Knowledge Prompting for Commonsense Reasoning*

Generate Knowledge Prompting Example

- The first step is to generate knowledge. Below is an example of how to generate the knowledge samples

Input: Greece is larger than Mexico.
Knowledge: Greece is approximately 131,957 sq km, while Mexico is approximately 1,964,375 sq km, making Mexico 1,389% larger than Greece.

...

Input: A rock is the same size as a pebble.
Knowledge: A pebble is a clast of rock with a particle size of 4 to 64 millimetres based on the Udden-Wentworth scale of sedimentology. Pebbles are generally considered larger than granules (2 to 4 millimetres diameter) and smaller than cobbles (64 to 256 millimetres diameter).

Input: Part of golf is trying to get a higher point total than others.
Knowledge:

Knowledge 1

The objective of golf is to play a set of holes in the least number of strokes. A round of golf typically consists of 18 holes. Each hole is played once in the round on a standard golf course. Each stroke is counted as one point, and the total number of strokes is used to determine the winner of the game.

Knowledge 2

Golf is a precision club-and-ball sport in which competing players (or golfers) use many types of clubs to hit balls into a series of holes on a course using the fewest number of strokes. The goal is to complete the course with the lowest score, which is calculated by adding up the total number of strokes taken on each hole. The player with the lowest score wins the game.

A
G

Generate Knowledge Prompting Example

- The knowledge samples are then used to generate **knowledge augmented questions** to get answer proposals
 - The highest-confidence response is selected as final answer

Question: Part of golf is trying to get a higher point total than others.
Yes or No?

Knowledge: The objective of golf is...

Explain and Answer: No, the objective of golf is not to get a higher point total than others. Rather, the objective is to play a set of holes in the least number of strokes. The total number of strokes is used to determine the winner of the game, not the total number of points.

High-confidence prediction

Question: Part of golf is trying to get a higher point total than others.
Yes or No?

Knowledge: Golf is a precision club-and-ball sport in which competing players (or golfers) use many types of clubs to hit balls into...

Explain and Answer: Yes, part of golf is trying to get a higher point total than others. Each player tries to complete the course with the lowest score, which is calculated by adding up the total number of strokes taken on each hole. The player with the lowest score wins the game.

Low-confidence prediction

Activ
Go to

Tools & IDEs

- There are many tools, libraries, and platforms with different capabilities and functionalities
- Capabilities include:
 - Developing and experimenting with prompts
 - Evaluating prompts
 - Versioning and deploying prompts



Dyno

Prompt Engineering IDE



LangChain



More tools here: <https://github.com/dair-ai/Prompt-Engineering-Guide#tools--libraries>

Applications

- Advanced prompting techniques are now allowing all sorts of advanced applications with LLMs
 - LLMs and external tools/APIs
 - Data-augmented Generation
 - QA with sources
 - Summarization using sources

Risks & Misuses

- **Adversarial Prompting**
 - Adversarial prompting is an important topic in prompt engineering as it could help to understand the risks and safety issues involved with LLMs.
- **Prompt Injection:** Prompt injection aims to hijack the model output by using clever prompts that change its behavior.
 - These attacks could be harmful

Instruction finetuning

PROMPT *Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

Source: (Ouyang et al., 2022), LM are **not aligned** with user intent

Instruction finetuning

PROMPT *Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

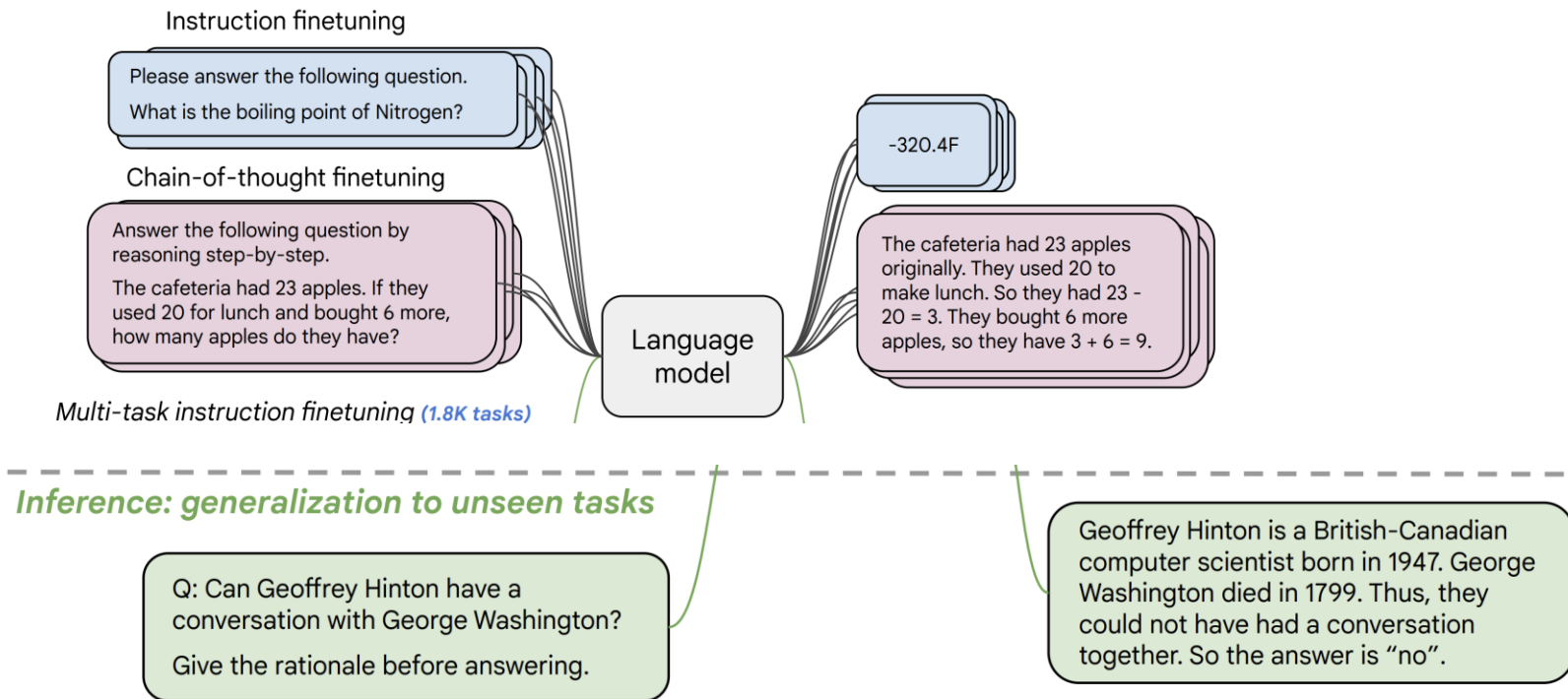
Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

Finetuning to the rescue!

Instruction finetuning

- Collect examples of (instruction, output) pair across many tasks and finetune LLMs



Evaluate on **unseen tasks**

Summary

- Introduction to Prompt Engineering
- Advanced Techniques for Prompt Engineering
- Applications & Tools & Others

References

- Prompt Engineering A lecture by DAIR.AI
- Saravia, E. (2022). Prompt Engineering Guide



Since 2004
UET
ĐẠI HỌC CÔNG NGHỆ, ĐHQGHN
VNU-University of Engineering and Technology



Since 1906
VNU
ĐẠI HỌC QUỐC GIA HÀ NỘI
Vietnam National University, Hanoi

Thank you

Email me
vinhnhv@vnu.edu.vn

Course: Large Language Models and Its Applications
sponsored by **KEPCO KDN Co., Ltd.**
Eco-friendly & Digital Centered Energy ICT Platform Leader