Overview of Large Language Models and Prompt Engineering

Yogesh Haribhau Kulkarni

YHK

Outline

YHK

About Me

YHK

Yogesh Haribhau Kulkarni

Bio:

- 20+ years in CAD/Engineering software development
- Got Bachelors, Masters and Doctoral degrees in Mechanical Engineering (specialization: Geometric Modeling Algorithms).
- Currently doing Coaching in fields such as Data Science, Artificial Intelligence Machine-Deep Learning (ML/DL) and Natural Language Processing (NLP).
- Feel free to follow me at:
  - Github (github.com/yogeshhk)
  - LinkedIn (www.linkedin.com/in/yogeshkulkarni/)
  - Medium (yogeshharibhaukulkarni.medium.com)
  - Send email to yogeshkulkarni at yahoo dot com

Office Hours: Saturdays, 2 to 5pm (IST); Free-Open to all; email for appointment.

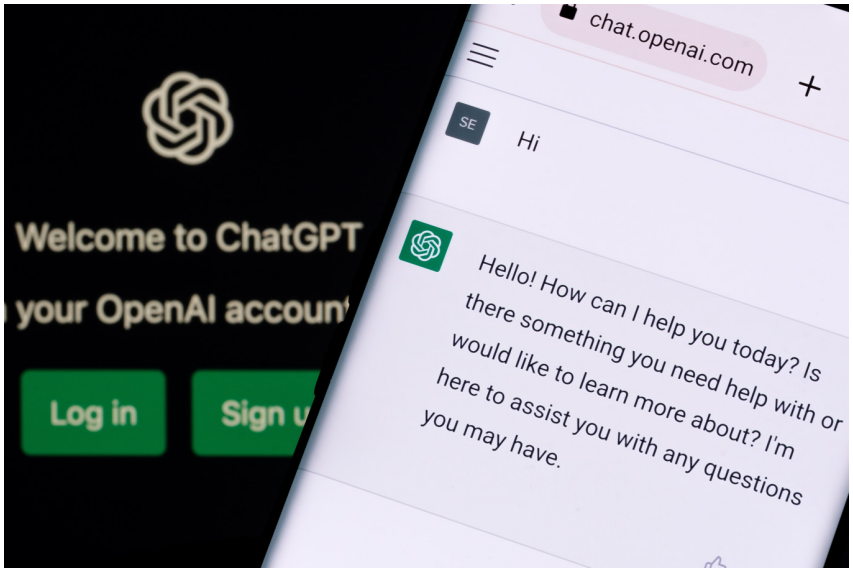
YHK

Overview of Large Language Models

YHK

What is ChatGPT?

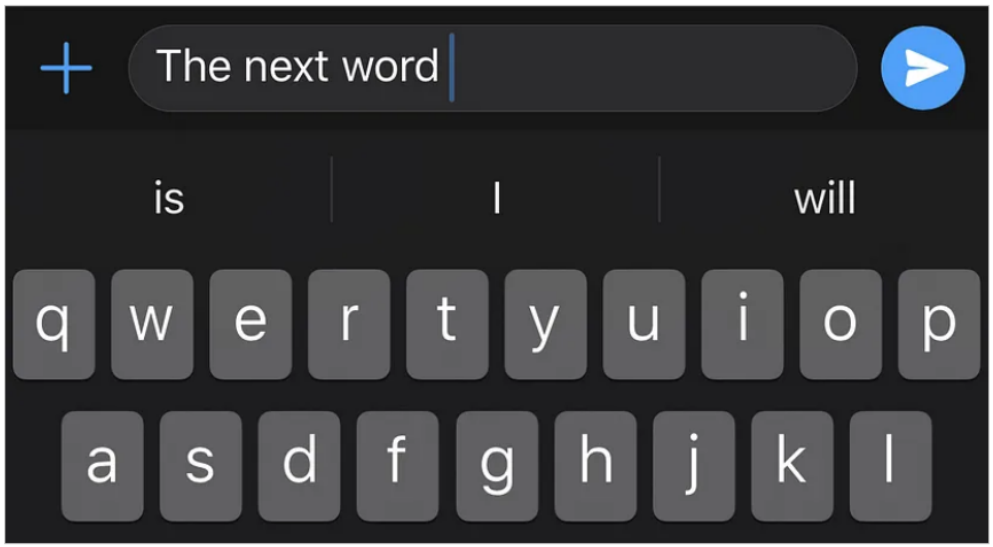
- A Chatbot
- A language model, that takes 'prompt' from user and generates a response.
- Built by OpenAI and released in Nov 2022
- Got 1m users in 5 days (Insta 2.5 months, Spotify 5m, Facebook 10m, Netflix 3.5yrs)
- Link: https://chat.openai.com/chat
- Details: https://openai.com/blog/chatgpt/



(Ref: OpenAI, creator of ChatGPT, casts spell on Microsoft - The Jakarta Post

What is a Language Models?

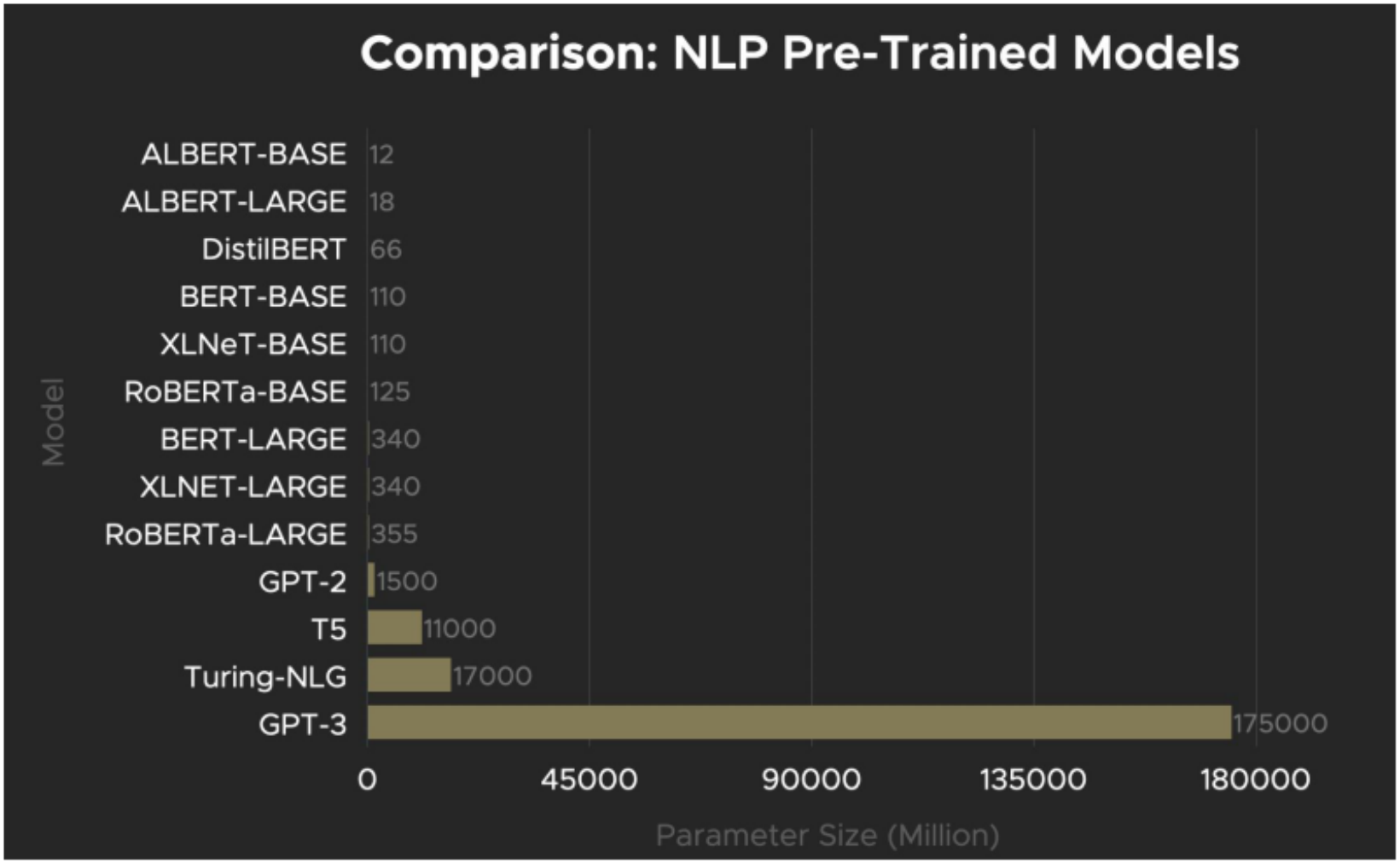
- While typing SMS, have you seen it suggests next word?
- While typing email, have you seen next few words are suggested?
- How does it suggest? (suggestions are not random, right?)
- In the past, for "Lets go for a . . . ', if you have typed 'coffee' 15 times, 'movie' say 4 times, then it learns that. Machine/Statistical Learning.
- Next time, when you type "Lets go for a ", what will be suggested? why?
- This is called Language Model. Predicting the next word. When done continuously, one after other, it spits sentence, called Generative Model.



Next word prediction using language modeling in keyboards(Mandar Deshpande)

YHK

Large Language Models - Comparison

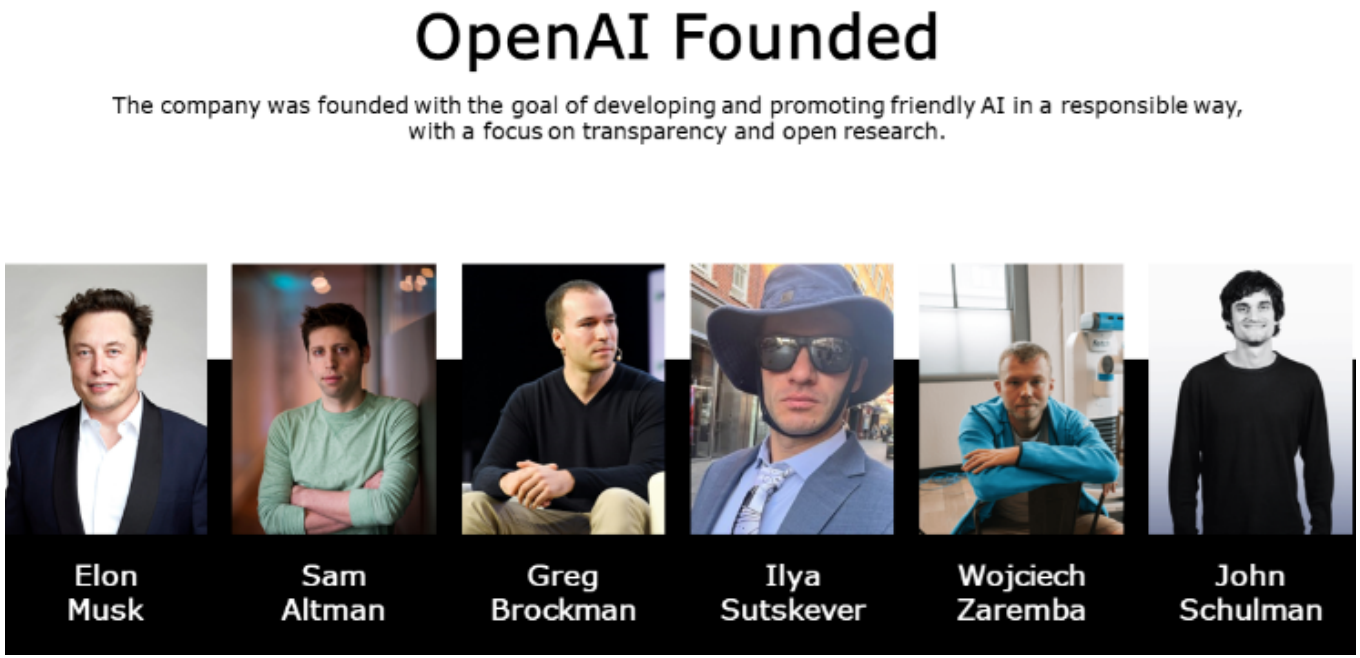


(Ref: Deus.ai https://www.deus.ai/post/gpt-3-what-is-all-the-excitement-about)

YHK

Open AI: Who are these guys?

- San Francisco-based artificial intelligence company
- Famous for its well-known DALL-E, generates images from prompts.
- Initially supported by Elon Musk
- The CEO is Sam Altman, who previously was president of Y Combinator.
- Microsoft is a partner and investor.
- Mission: To ensure that artificial general intelligence benefits all of humanity. Prevent misuse of AI



(Ref: https://www.slideegg.com/open-ai-chat-gpt)

GPTs Training

GPT: Generative Pre-trained Transformers

- GPT-1 is pre-trained on the BooksCorpus dataset, containing 7000 books amounting to 5GB of data
- GPT-2 is pre-trained using the WebText dataset which is a more diverse set of internet data containing 8M documents for about 40 GB of data
- GPT-3 uses an expanded version of the WebText dataset, two internet-based books corpora that are not disclosed and the English-language Wikipedia which constituted 600 GB of data

YHK

GPTs Training compared to human reading

- GPT-3 was trained on 499B tokens; GPT-4, on 1.4T tokens.
- In comparison, if you spent 12 hours a day reading for an entire lifetime (80 years) at average speed (250 words / minute), we would absorb 5.26B words (tokens).
- That's a ratio of 100:1 between the training data used for GPT-3 and the amount of data that can ever be read by a human, and 260:1 for GPT-4.

(Ref: LinkedIn post by Dr Jennifer Prendki)

YHK

GPT3 vs ChatGPT

	GPT-3	ChatGPT
Training Parameters	175 billion	175 BillionRLHF TechniqueHuman conversational text
Designed For	NLP Tasks	Chatbot apps
Content Retention Ability	No	Yes
Use Cases	Creative Writing, Business Writing, Ads, Translation, Summarization, Data Analysis	Customer Service, Data Analysis, Blogging, Copywriting, Coding, Debugging, etc.

(Ref:ChatGPT Explained: Complete A-Z Guide - Kripesh Adwani)

YHK

Background

Progression

Models for prediction:

- On data, derive features, put statistical techniques like regression. One model per task. That's Machine Learning.
- Feed raw data, employ neural networks. One model per task. That's Deep Learning.
- Use Text data, get embeddings, use ML/DL, say for classification. One model per task. That's Natural Language Processing.
- Train neural network on large corpus, store weights and architecture, then add final layers for say classification on custom data+labels. That's Pretrained model. One model, many tasks.
- Train Large Language Model, just supply instructions on what to do, works. One model many tasks. Zero-shot, few-shots.

(More info at SaaS LLM https://medium.com/google-developer-experts/saasgpt-84ba80265d0f)

YHK

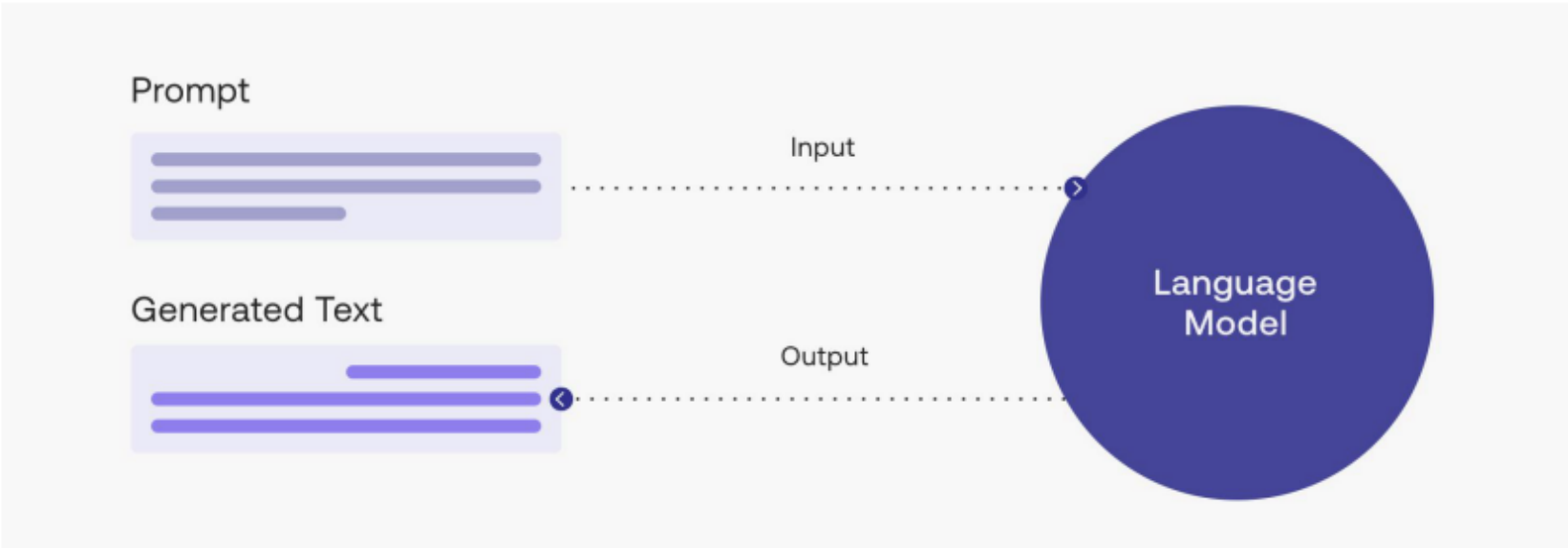
New Programming Language?

(Ref: Prompt Engineering Sudalai Rajkumar)

YHK

What is Prompt Engineering?

Prompt engineering is a NLP concept that involves discovering inputs that yield desirable or useful results

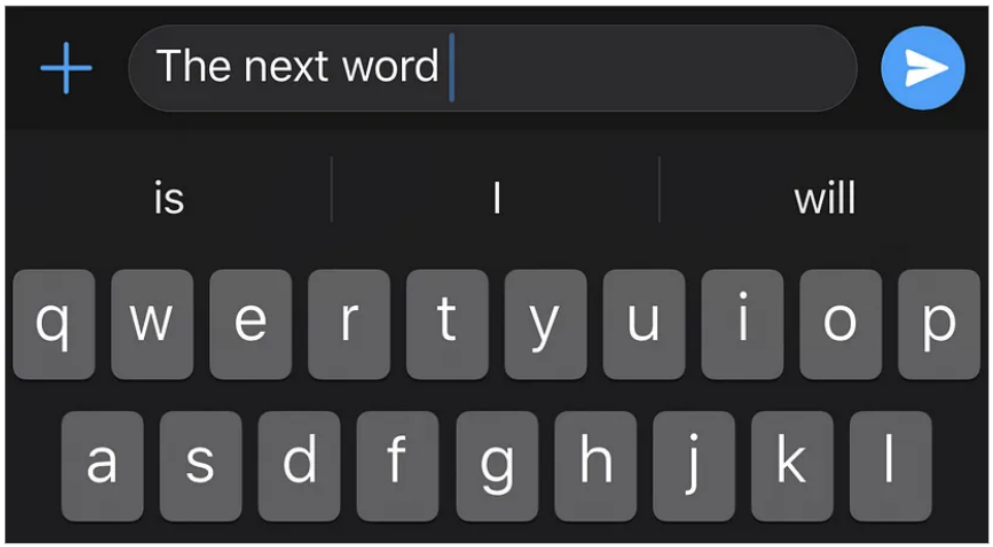


(Ref: Cohere https://docs.cohere.ai/docs/prompt-engineering)

YHK

What is a Language Models?

- While typing SMS, have you seen it suggests next word?
- While typing email, have you seen next few words are suggested?
- How does it suggest? (suggestions are not random, right?)
- In the past, for "Lets go for a . . . ', if you have typed 'coffee' 15 times, 'movie' say 4 times, then it learns that. Machine/Statistical Learning.
- Next time, when you type "Lets go for a ", what will be suggested? why?
- This is called Language Model. Predicting the next word. When done continuously, one after other, it spits sentence, called Generative Model.



Next word prediction using language modeling in keyboards(Mandar Deshpande)

YHK

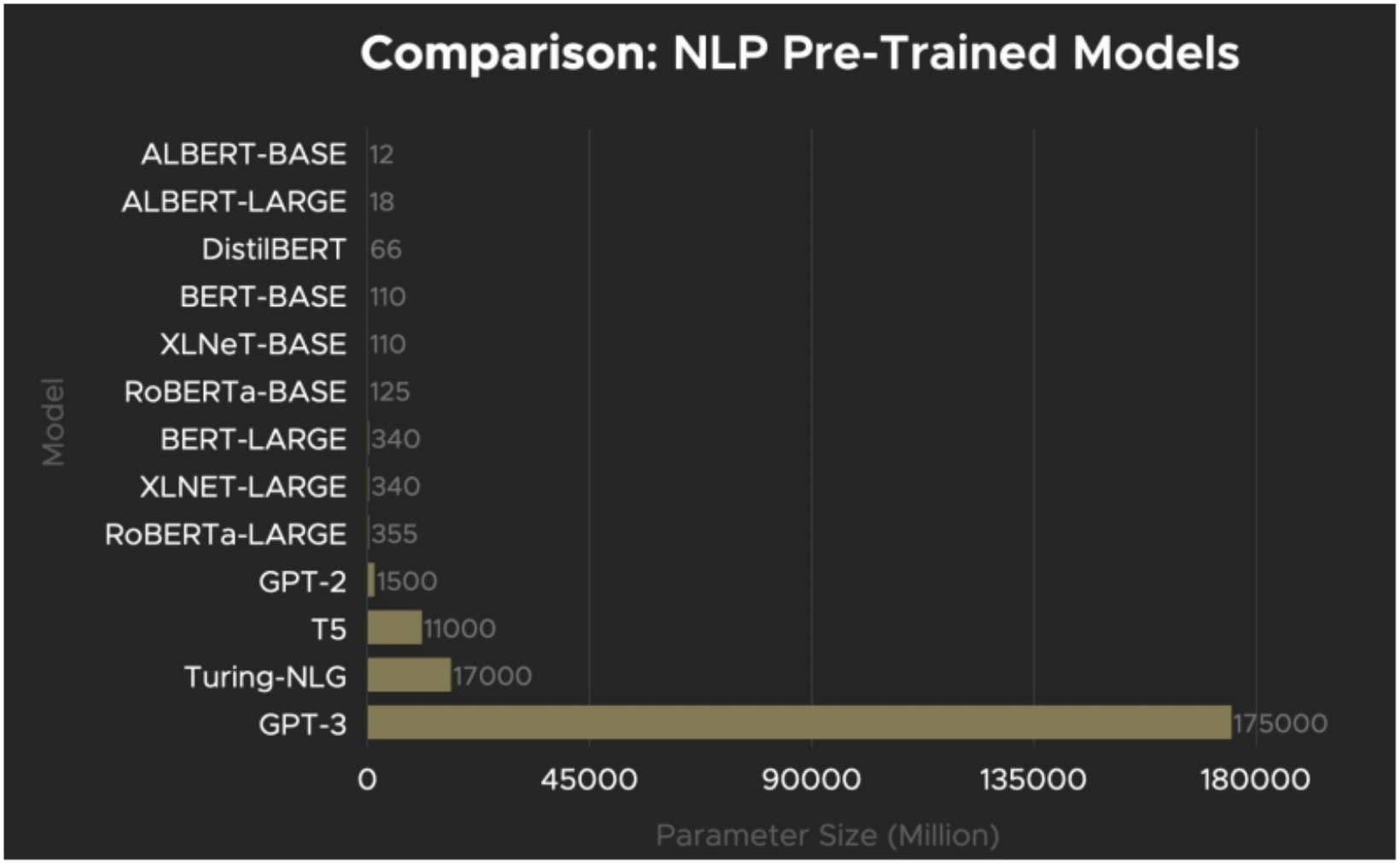
Evolution of Language Models

Language Models can be statistical (frequency based) or Machine/Deep Learning (supervised) based. Simple to complex.

(Ref: Analytics Vidhya https://editor.analyticsvidhya.com/uploads/59483evolution_of_NLP.png)

YHK

Large Language Models - Comparison



(Ref: Deus.ai https://www.deus.ai/post/gpt-3-what-is-all-the-excitement-about)

What is Prompt Engineering?

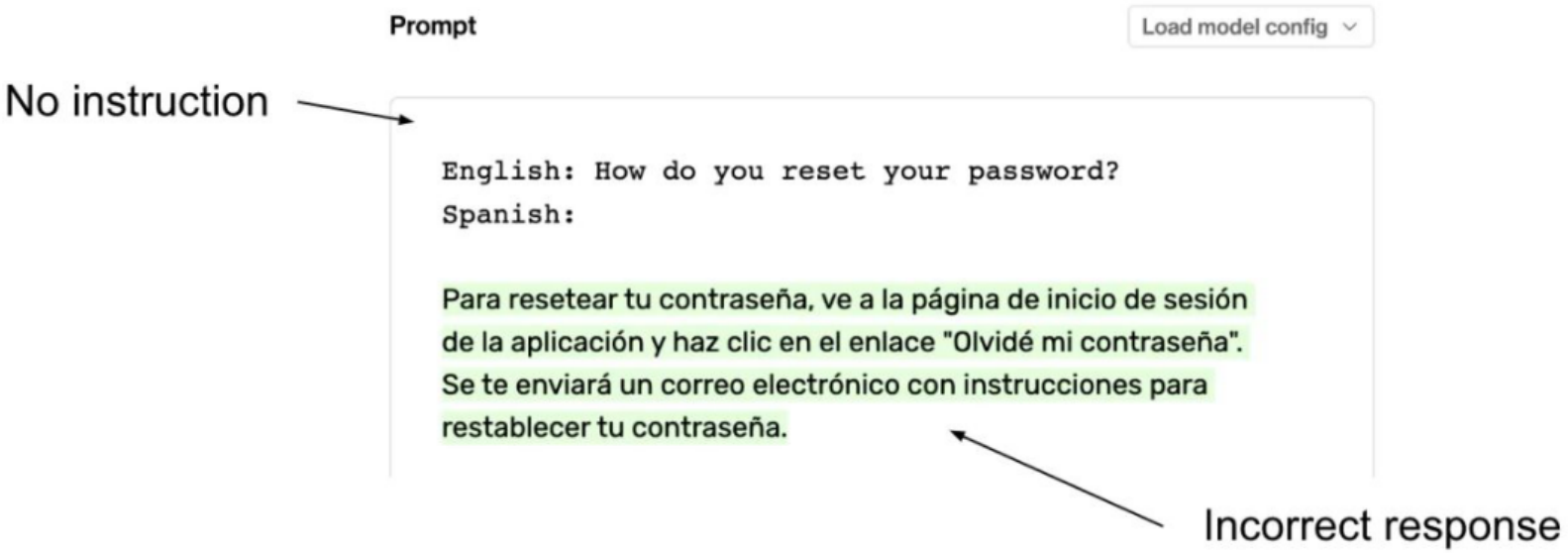
How to talk to AI to get it to do what you want

(Ref: Human Loop https://humanloop.com/blog/prompt-engineering-101)

YHK

What is Prompt Engineering?

But need to tell, for sure, else, nothing



(Ref: Human Loop https://humanloop.com/blog/prompt-engineering-101)

YHK

What is Prompt Engineering?

- For prompt `What is 1,000,000 * 9,000?` GPT-3 (text-davinci-002) (an AI) sometimes answers 9,000,000 (incorrect). This is where prompt engineering comes in.
- If, instead of asking What is `1,000,000 * 9,000?`, we ask `What is 1,000,000 * 9,000? Make sure to put the right amount of zeros, even if there are many:`, GPT-3 will answer 9,000,000,000 (correct).
- Why is this the case? Why is the additional specification of the number of zeros necessary for the AI to get the right answer? How can we create prompts that yield optimal results on our task?
- That's Prompt Engineering.

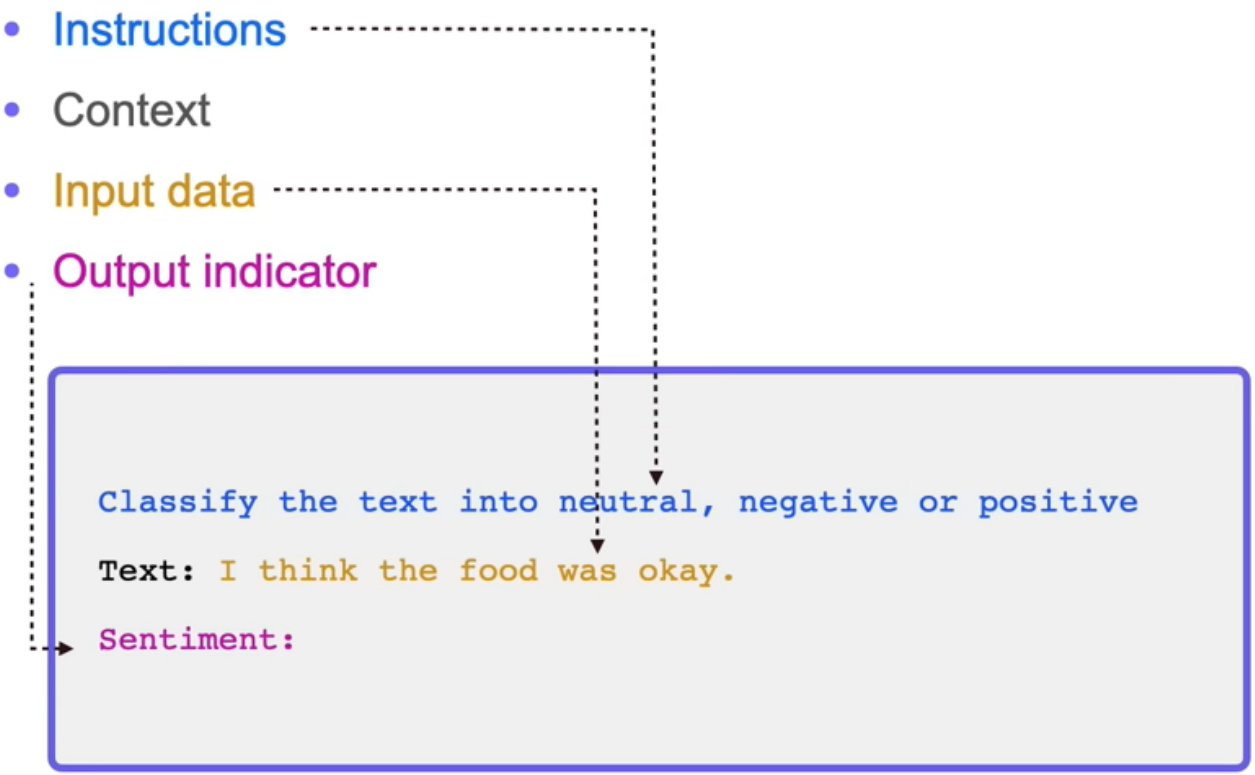
(Ref: https://learnprompting.org/docs/basics/prompting)

YHK

Concepts

Elements of Prompt

A prompt is composed of:



(Ref: Prompt Engineering Overview - Elvis Saravia)

YHK

Settings of Prompt

- 'temperature': before applying the softmax function, temperature is used scale the logits. With it, creativity or variability is allowed. If you re-run the prompt, with 0, no change, but with 1, lots of variation. Default is 0.7. With a temperature between 0 and 1, we can control the randomness and creativity of the model's predictions. Temperature defines how likely it is to choose less probable words. T=0 gives the same response every time because there's a 0
- 'top_p' or 'nucleus sampling': specifies a sampling threshold during inference time, words passing the threshold are sampled for the output. Top-p goes for a minimal set of words, the probability of which does not exceed exceeds p. In practice, this means the following: if you choose reasonably high p, like 0.9, you would likely get a set of the most likely words for the model to choose from
- Like the temperature, the top p parameter controls the randomness and originality of the model.
- OpenAI documentation recommends using either one parameter or the other and setting the unused parameter to the neutral case, i.e. 1.0.

YHK

Example: Prompting by Instruction

S write a reminder mail to customer on the upcoming home loan EMI

Subject: Reminder Regarding Your Upcoming Home Loan EMI

Dear [Customer Name],

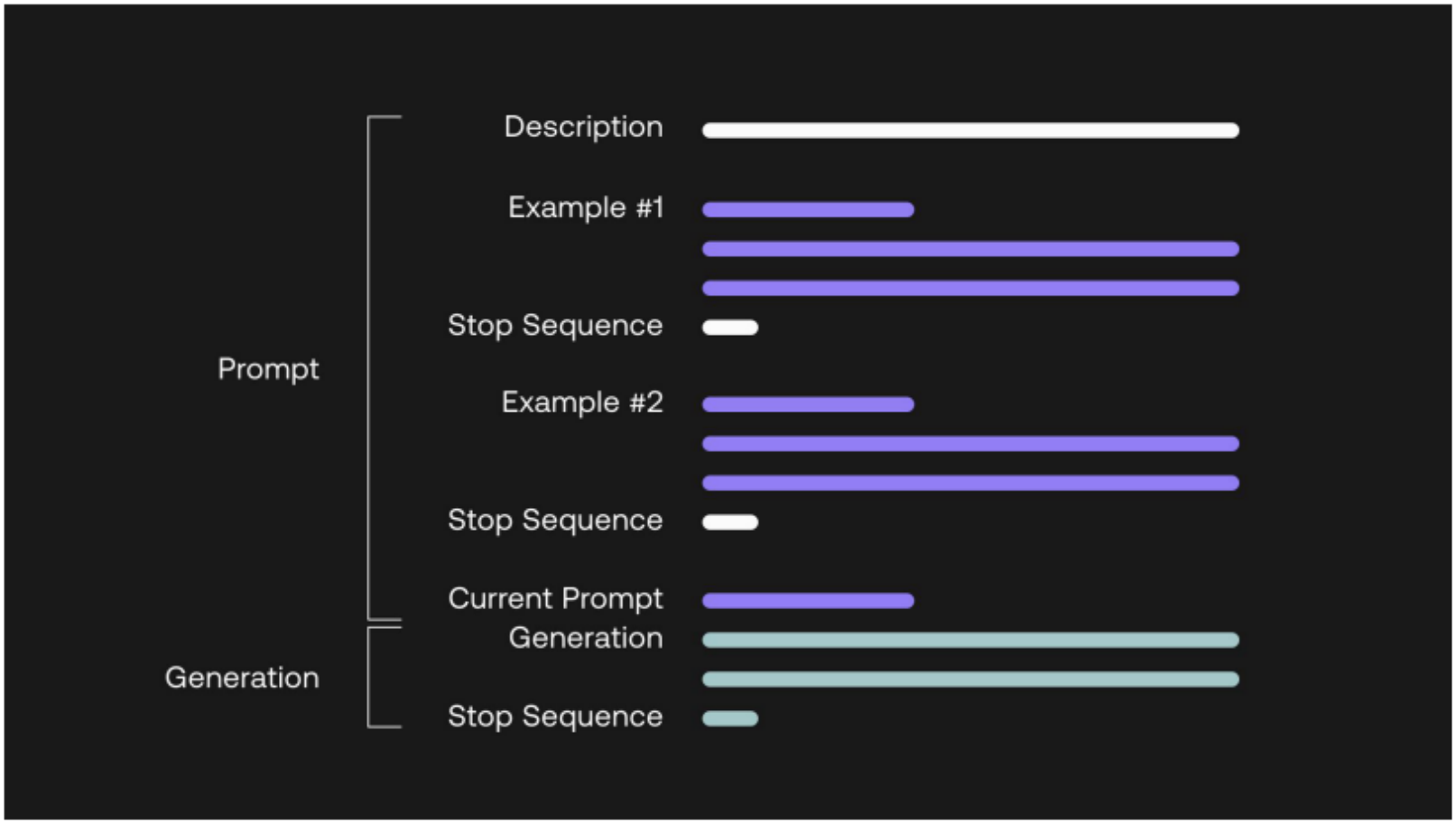
I hope this email finds you in good health and spirits. This email is a friendly reminder regarding your upcoming home loan EMI.

As per our records, your next EMI is due on [due date]. The amount due is [EMI amount]. We would like to remind you to make the payment before the due date to avoid any late payment charges.

(Ref: Cohere https://txt.cohere.ai/generative-ai-part-1/)

YHK

Prompting by Examples



(Ref: Cohere https://txt.cohere.ai/generative-ai-part-1/)

YHK

Example: Prompting by Examples

Text: I am happy with the service
Sentiment: Positive

Text: The service was not up to the mark.
Sentiment:

Negative

(Ref: Cohere https://txt.cohere.ai/generative-ai-part-1/)

YHK

Length Control

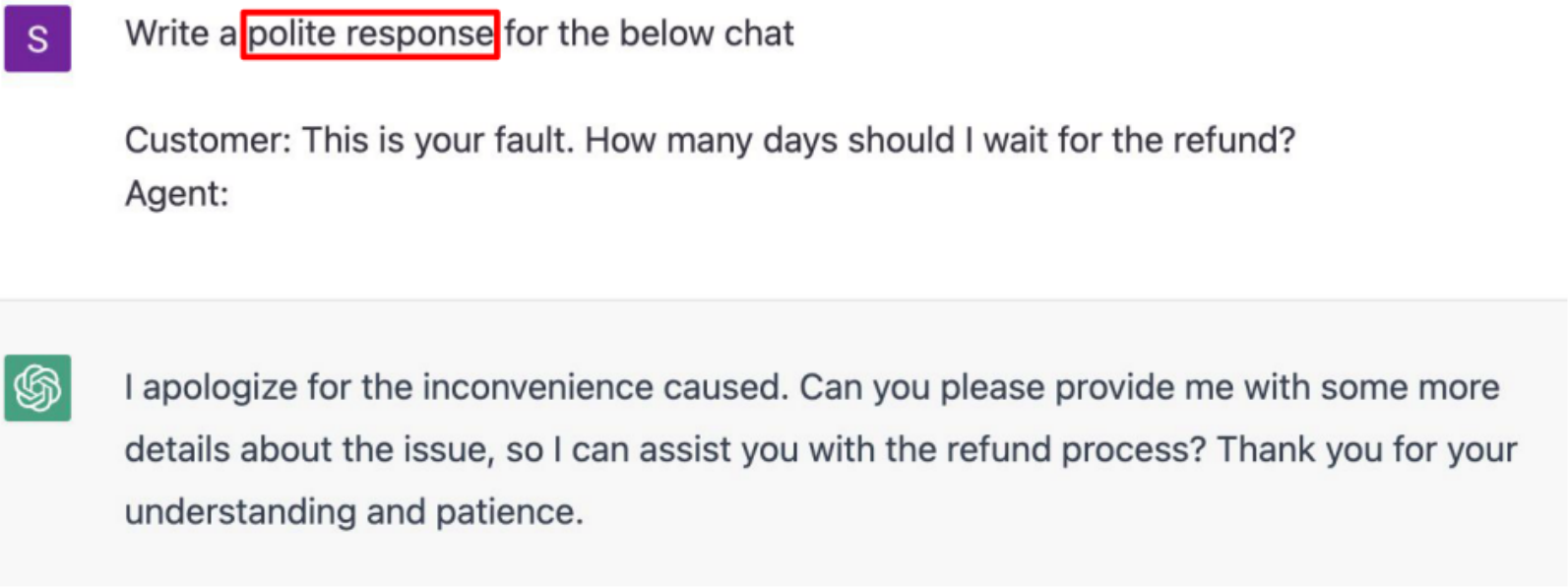
Specify a desired word count or character count as part of the prompt

(Ref: Prompt Engineering Sudalai Rajkumar)

YHK

Tone Control

Specify specific words or phrases that indicate the desired tone



(Ref: Prompt Engineering Sudalai Rajkumar)

YHK

Style Control

Specify the desired writing style.

S

Write a short mail to the banking customer with their last weeks transactions as a table and the number of transactions and total as bullet points.

Subject: Your Recent Transactions

Dear [Customer Name],

I hope this email finds you well. This email is to provide you with the details of your recent transactions.

Please find below a table summarizing your transactions from [date range]:

Date	Description	Amount
[date 1]	[description 1]	[amount 1]
[date 2]	[description 2]	[amount 2]
[date 3]	[description 3]	[amount 3]

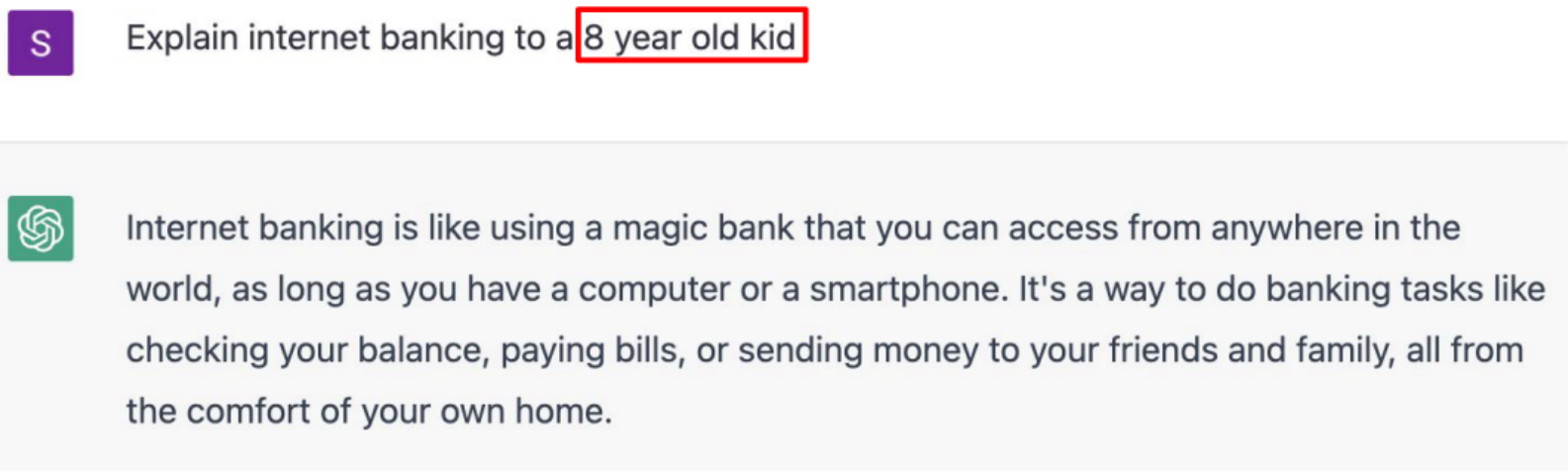
In summary:

- Total number of transactions: [number of transactions]
- Total amount: [total amount]

YHK

Audience Control

Specify the desired audience.

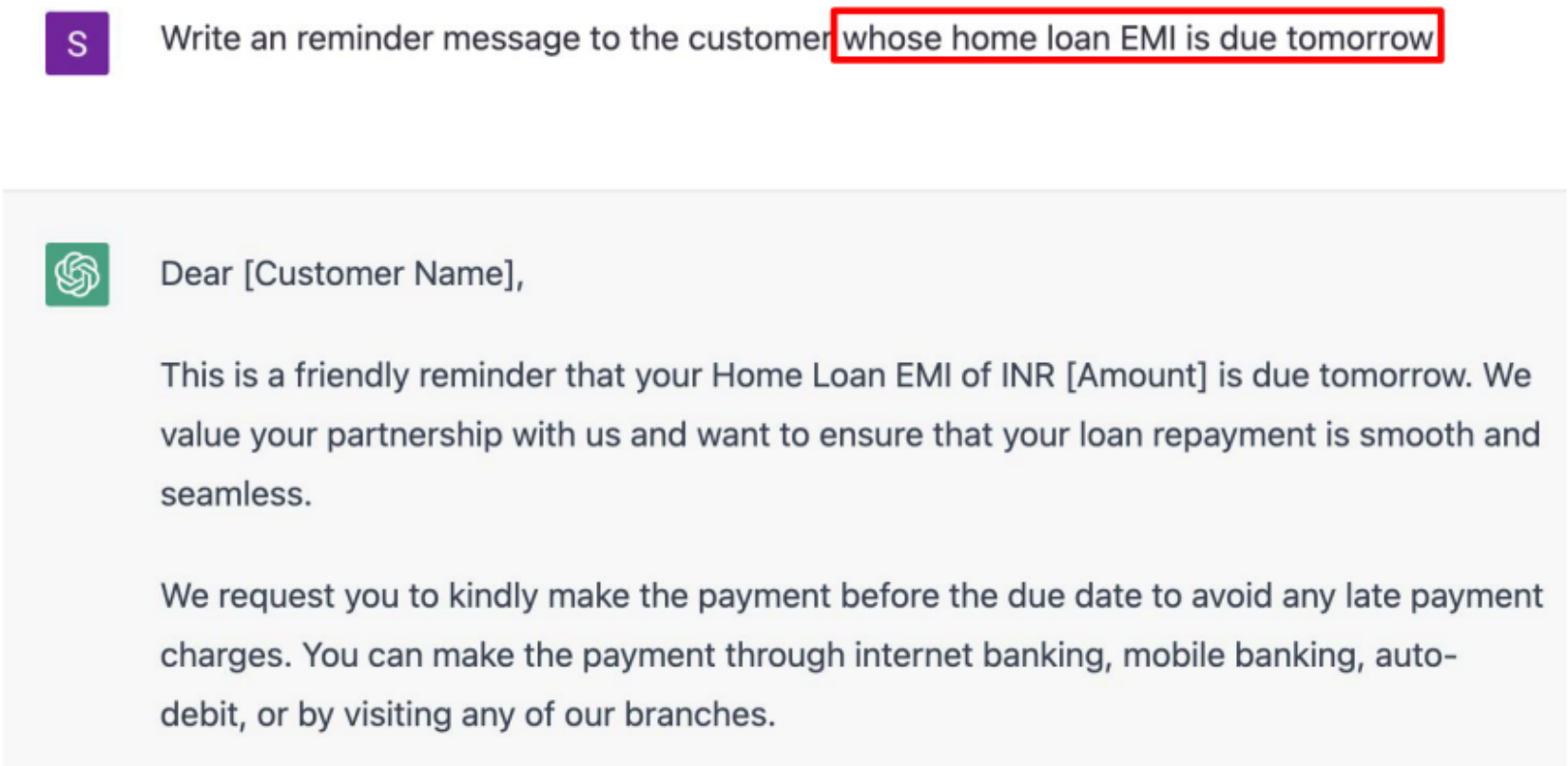


(Ref: Prompt Engineering Sudalai Rajkumar)

YHK

Context Control

Specify the information about the context.



(Ref: Prompt Engineering Sudalai Rajkumar)

YHK

Calling of Prompt

```python
import openai
import os

from dotenv import load_dotenv, find_dotenv
_ = load_dotenv(find_dotenv())

openai.api_key  = os.getenv('OPENAI_API_KEY') # for langchain it does it automatically

def get_completion(prompt, model="gpt-3.5-turbo"):
    messages = [{"role": "user", "content": prompt}]
    response = openai.ChatCompletion.create(
        model=model,
        messages=messages,
        temperature=0, # this is the degree of randomness of the model's output
    )
    return response.choices[0].message["content"]
```

Principles

- Give clear and specific instructions
- Give the model time to “think”

(Ref: ChatGPT Prompt Engineering for Developers - Deep Learning AI)

YHK

Applications of Prompts

Text Generation

S LLMs are

Language models (LMs), also known as Large language models (LLMs), are artificial intelligence models that are trained to generate text that is similar to human language. They are trained on vast amounts of text data and can generate coherent text, perform question answering, translation, and many other NLP tasks. The most famous LLMs currently are GPT-3 (Generative Pre-trained Transformer-3) developed by OpenAI.

(Ref: Prompt Engineering Sudalai Rajkumar)

YHK

Text Classification

S

Classify the following text into one of the classes - credit card, internet banking, others

Text: I was not able to log into the banks website

Answer:

Internet Banking

(Ref: Prompt Engineering Sudalai Rajkumar)

YHK

Text Translation

(Ref: Prompt Engineering Sudalai Rajkumar)

YHK

Text Comprehension

S

Context: The Reserve Bank of India (RBI) on Wednesday hiked the repo rate by 25 basis points (bps) to 6.50 percent by a majority of 4-2. The three-day meeting of the Monetary Policy Committee (MPC) started on Monday.

Question: What is the new interest rate?

Answer:

The new interest rate is 6.50 percent.

(Ref: Prompt Engineering Sudalai Rajkumar)

YHK

Text Summarization

S

Summarize the below paragraph in less than 15 words

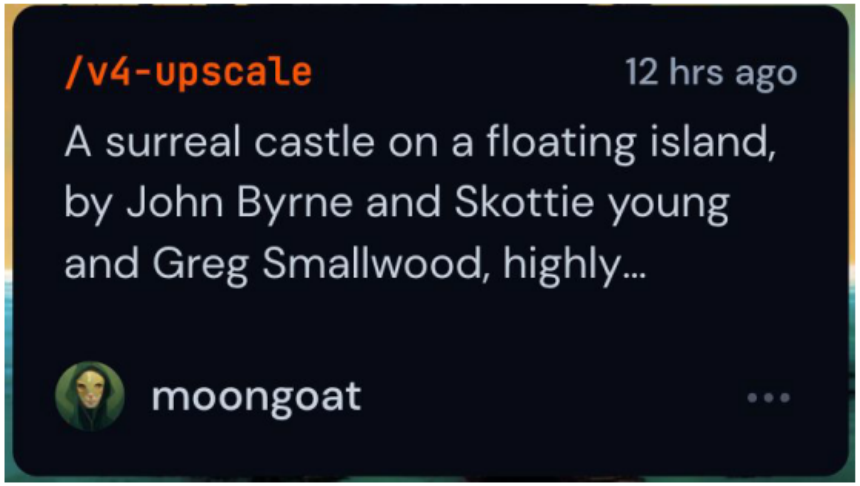
The RBI announced a hike of 25 basis points in the repo rate on February 8, as envisaged by many experts earlier. The move is aimed at containing inflationary pressures in the economy, with a high-interest-rate regime. The MPC has taken note of the moderation in headline consumer inflation numbers in recent times. However, it highlighted various risks that may keep inflation elevated, including global commodity prices.

RBI hikes repo rate 25 bps to control inflation, despite moderating consumer inflation.

(Ref: Prompt Engineering Sudalai Rajkumar)

YHK

Image Generation



(Ref: Prompt Engineering Sudalai Rajkumar)

Models / Tools: Dall-E , Midjourney, Stable Diffusion

YHK

For generating code using Codex

Provide Codex with a prompt consisting of the following:

- High level task description: Tell the model to use a helpful tone when outputting natural language
- High level context: Describe background information like API hints and database schema to help the model understand the task
- Examples: Show the model examples of what you want
- User input: Remind the model what the user has said before

(Ref: https://microsoft.github.io/prompt-engineering/)

YHK

Conclusions

ChatGPT Ultimate Prompting Guide

- Tone: Specify the desired tone (e.g., formal, casual, informative, persuasive).
- Format: Define the format or structure (e.g., essay, bullet points, outline, dialogue).
- Act as: Indicate a role or perspective to adopt (e.g., expert, critic, enthusiast).
- Objective: State the goal or purpose of the response (e.g., inform, persuade, entertain).
- Context: Provide background information, data, or context for accurate content generation.
- Scope: Define the scope or range of the topic.
- Keywords: List important keywords or phrases to be included.
- Limitations: Specify constraints, such as word or character count.
- Examples: Provide examples of desired style, structure, or content.
- Deadline: Mention deadlines or time frames for time-sensitive responses.

(Ref: LinkedIn post by Generative AI, Twitter by Aadit Sheth, Source : Reddit)

YHK

ChatGPT Ultimate Prompting Guide

- Audience: Specify the target audience for tailored content.
- Language: Indicate the language for the response, if different from the prompt.
- Citations: Request inclusion of citations or sources to support information.
- Points of view: Ask the AI to consider multiple perspectives or opinions.
- Counter arguments: Request addressing potential counterarguments.
- Terminology: Specify industry-specific or technical terms to use or avoid.
- Analogies: Ask the AI to use analogies or examples to clarify concepts.
- Quotes: Request inclusion of relevant quotes or statements from experts.
- Statistics: Encourage the use of statistics or data to support claims.
- Visual elements: Inquire about including charts, graphs, or images.
- Call to action: Request a clear call to action or next steps.
- Sensitivity: Mention sensitive topics or issues to be handled with care or avoided.

(Ref: LinkedIn post by Generative AI, Twitter by Aadit Sheth, Source : Reddit)

YHK

Interaction Guidelines: Avoid Misuses

- Factual Accuracy: Interactions must be free from factual inaccuracies that can be challenged by social media or journalists.
- Negative Debates: Avoid discussing topics that fuel negative or concerning online debates, such as AI sentience, AI in education, AI-driven job displacements, and politically divisive issues.
- Minors' Involvement: Do not include use cases specifically targeting or involving individuals under 18 years old.
- Sensitivity and Misinformation: Prevent the inclusion of sensitive, misleading, or hazardous responses.
- Search and Google Assistant: Interactions that require basic, straightforward answers are better suited for Search or Google Assistant.
- Financial/Legal/Medical Advice: Refrain from providing advice related to financial matters, legal issues, or medical concerns.
- Brand Names and Trademarks: Avoid mentioning specific brand names, trademarks, or public figures (except historical figures).
- No Reviews or Tweets: Do not request reviews of restaurants, businesses, or tweets to minimize the risk of associating with bots.
- Avoid Personification: Refrain from personifying the product or brand and from encouraging users to address Bard by name.

YHK

Limitations

Boie is a real company, the product name is not real. So, see what you get . . .

```
prompt = f"""
Tell me about AeroGlide UltraSlim Smart Toothbrush by Boie
"""
response = get_completion(prompt)
print(response)
```

New Roles?

Coming up with good prompt is a combination of art and science

(Ref: Prompt Engineering Sudalai Rajkumar)

YHK

Read on to learn how to engineer good prompts!

- Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., & Singh, S. (2020). AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.18653/v1/2020.emnlp-main.346
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2022). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Computing Surveys. https://doi.org/10.1145/3560815
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners.
- Zhao, T. Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate Before Use: Improving Few-Shot Performance of Language Models.

YHK

My Sketchnote

(Ref: https://medium.com/technology-hits/prompting-is-all-you-need-5dddb82bd022)

Take Aways

Prompt Engineering is an Iterative Process:

- Try something
- Analyze where the results do not match the expectations
- Clarify instructions, gives examples, specify output format, specify constraints, etc
- Test on a batch of known results.

YHK

Resources

- Prompt Engineering Guide
  https://github.com/dair-ai/Prompt-Engineering-Guide
- Awesome ChatGPT Prompts
  https://github.com/f/awesome-chatgpt-prompts/
- ChatGPT Prompt Engineering for Developers - Deep Learning AI

YHK

Thanks . . .

- Search **"Yogesh Haribhau Kulkarni"** on Google and follow me on LinkedIn and Medium
- Office Hours: Saturdays, 2 to 5pm (IST); Free-Open to all; email for appointment.
- Email: yogeshkulkarni at yahoo dot com

(Generated by Hugging Face QR-code-AI-art-generator, with prompt as "Follow me")

YHK