

INTRODUCTION TO GENERATIVE AI

Yogesh Haribhau Kulkarni



Outline

① OVERVIEW

② OVERVIEW

③ DEMO

④ CONCLUSIONS

⑤ REFERENCES



About Me



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](mailto:yogeshkulkarni@yahoo.com)



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

Introduction to Generative AI

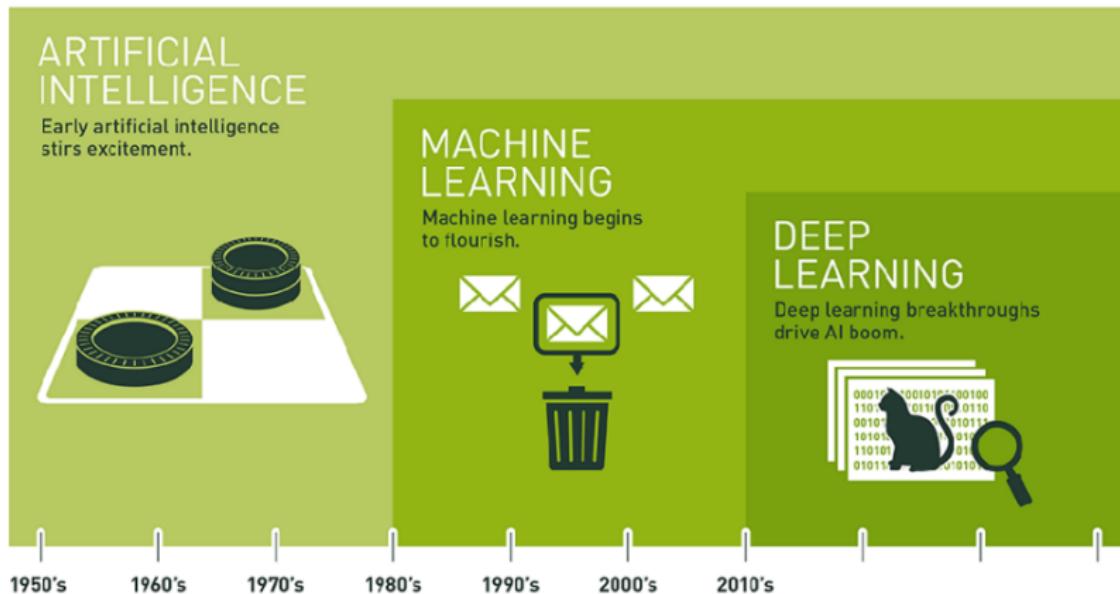


Introduction

- ▶ What is Generative AI?
- ▶ What is not Generative AI?
- ▶ How is it related to AI-ML-DL?

Relationship between AI, ML, DL

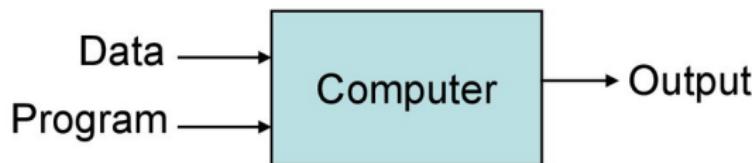
First, lets see whats AI-ML-DL and relationship among them.



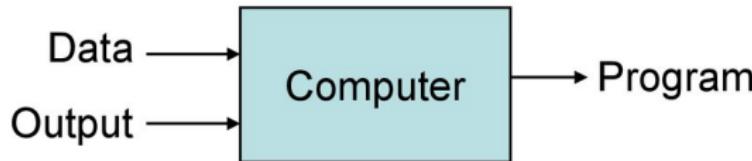
(Ref: <https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/>)

Traditional vs. Machine Learning?

Traditional Programming



Machine Learning



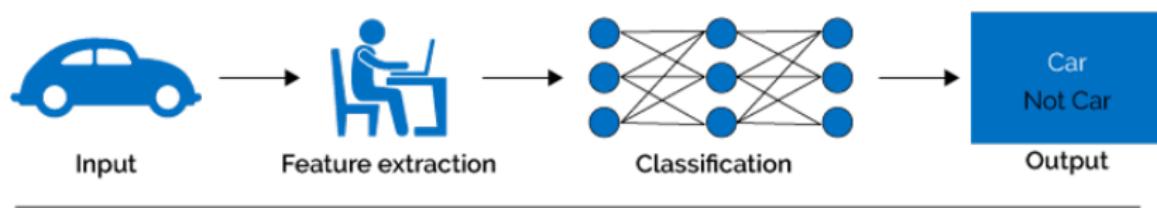
Why Machine Learning?

- ▶ Problems with High Dimensionality
- ▶ Hard/Expensive to program manually
- ▶ Job \$\$\$

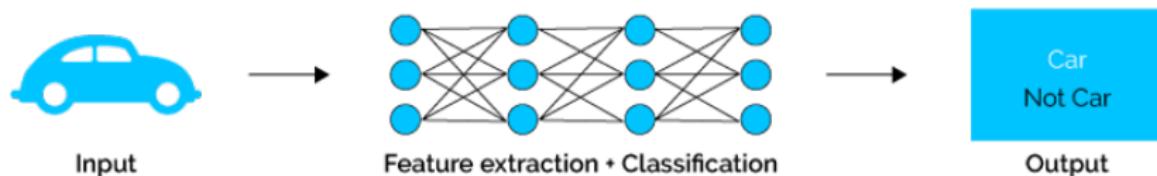
ML vs DL: What's the difference?

Deep learning algorithms attempt to learn (multiple levels of) representation by using a hierarchy of multiple layers

Machine Learning



Deep Learning



(Reference: <https://www.xenonstack.com/blog/static/public/uploads/media/machine-learning-vs-deep-learning.png>)

Use Deep Learning When ...

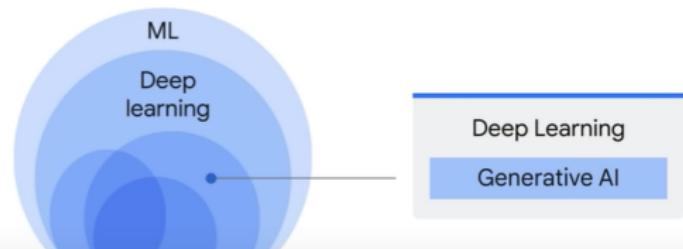
- ▶ You have lots of data (about 10k+ examples)
- ▶ The problem is “complex” - speech, vision, natural language
- ▶ The data is unstructured
- ▶ Techniques to model ‘ANY’ function given ‘ENOUGH’ data.

(Ref: Introduction to TensorFlow 2.0 - Brad Miro)



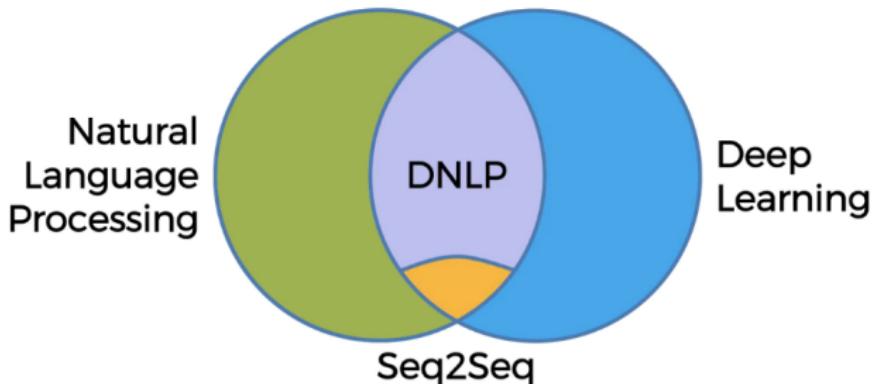
What is Gen AI wrt AI, ML, DL

Generative AI
is a **subset** of
Deep Learning



(Ref: Introduction to Generative AI - Google Cloud Tech)

What is Deep NLP



(Ref: Deep Learning and NLP A-Z - Kirill Eremenko)
(Note: Size is not indicative of importance)

Seq2Seq is heavily used technique of DNLP for sequence to sequence modeling,
eg Translation, Q & A, etc. Thats the basis of Large Language Models (LLMs)

Types of Approaches

Deep Learning Model Types



Discriminative

- Used to classify or predict
- Typically trained on a dataset of labeled data
- Learns the relationship between the features of the data points and the labels

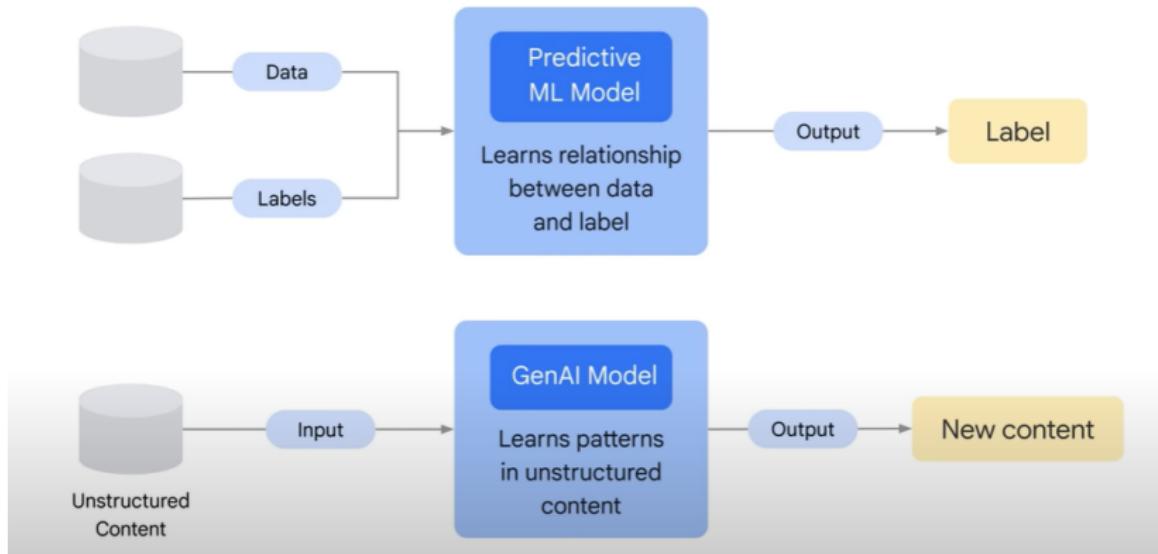


Generative

- Generates new data that is similar to data it was trained on
- Understands distribution of data and how likely a given example is
- Predict next word in a sequence

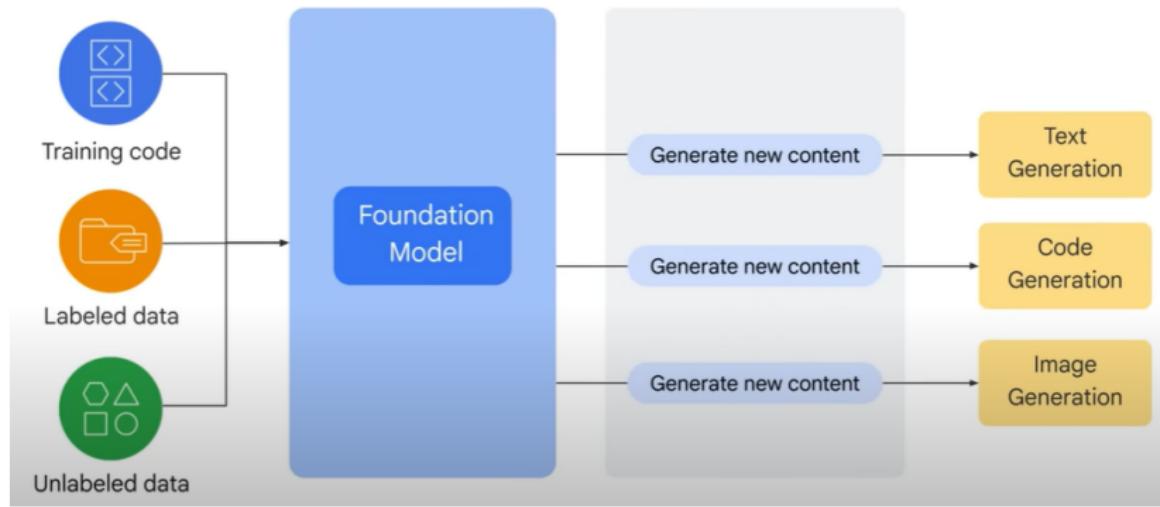
(Ref: Introduction to Generative AI - Google Cloud Tech)

Types of Approaches



(Ref: Introduction to Generative AI - Google Cloud Tech)

What is Foundation Model?



(Ref: Introduction to Generative AI - Google Cloud Tech)

Same Problem, using different Technologies



Difference across technologies, old to new

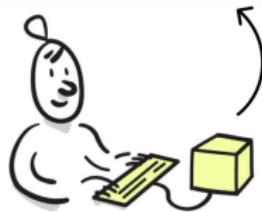
Lets see how the solutions to the problem of detecting a cat from images using traditional programming, deep learning, and generative AI, respectively.



Traditional Programming

- ▶ Traditional programming involves writing explicit rules to detect a cat in images.
- ▶ Features like color, texture, and shape can be used to define these rules.
- ▶ However, designing accurate rules for complex patterns like cat detection can be challenging.
- ▶ It requires extensive domain knowledge and might not generalize well to different images.

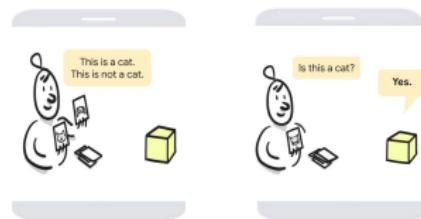
```
cat:  
  type: animal  
  legs: 4  
  ears: 2  
  fur: yes  
  likes: yarn, catnip
```



(Ref: Primer on LLM and Gen AI - Google Cloud)

Deep Learning

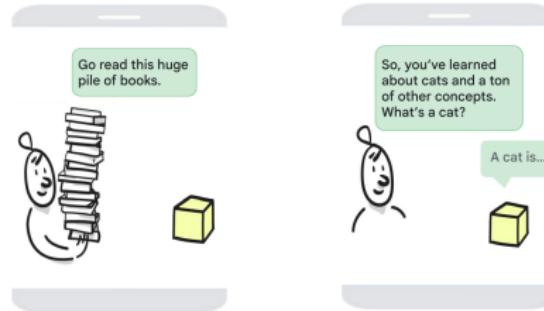
- ▶ Deep learning utilizes neural networks to automatically learn features for cat detection.
- ▶ Convolutional Neural Networks (CNNs) are particularly effective for image classification tasks.
- ▶ Large labeled datasets of cat images are used to train the network.
- ▶ The network learns to identify unique cat features and generalize them to detect cats in new images.
- ▶ Deep learning offers better accuracy and can handle complex patterns without explicit rule definition.



(Ref: Primer on LLM and Gen AI - Google Cloud)

Generative AI

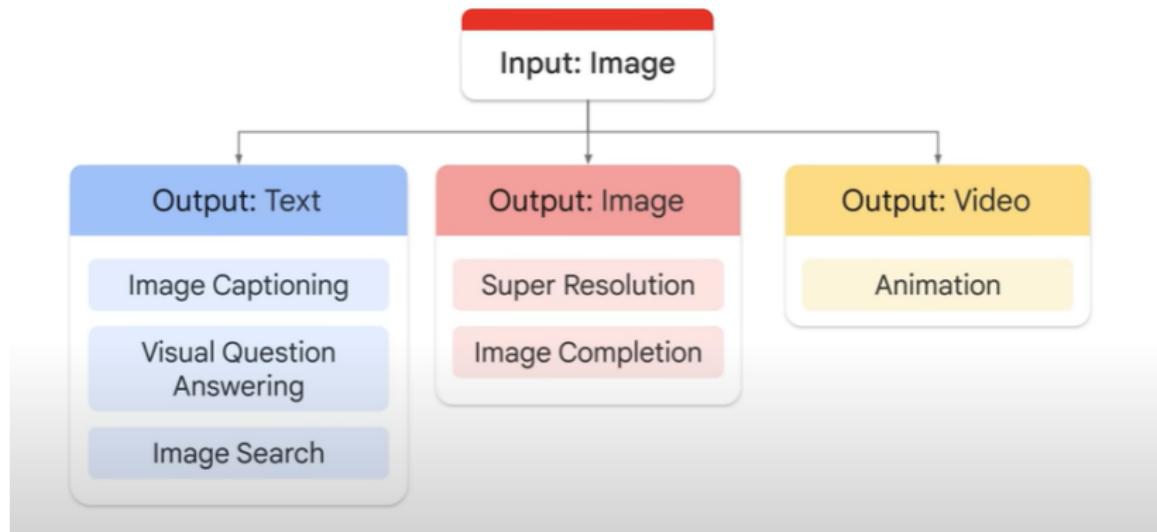
- ▶ Generative AI focuses on generating new data, including images of cats.
- ▶ Generative Adversarial Networks (GANs) are used to generate realistic cat images.
- ▶ The GAN consists of a generator and a discriminator that compete against each other.
- ▶ The generator learns to generate increasingly realistic cat images, while the discriminator learns to distinguish real from generated images.
- ▶ The generated cat images can be used to augment datasets for cat detection models.



(Ref: Primer on LLM and Gen AI - Google Cloud)

YHK

Modalities in Generative AI

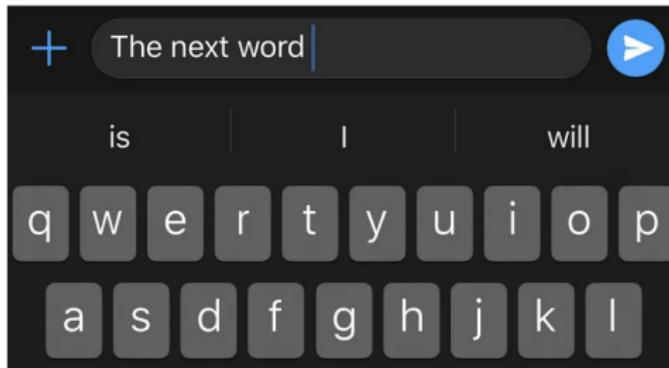


(Ref: Introduction to Generative AI - Google Cloud Tech)

Let's focus on the most popular modality ...

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)

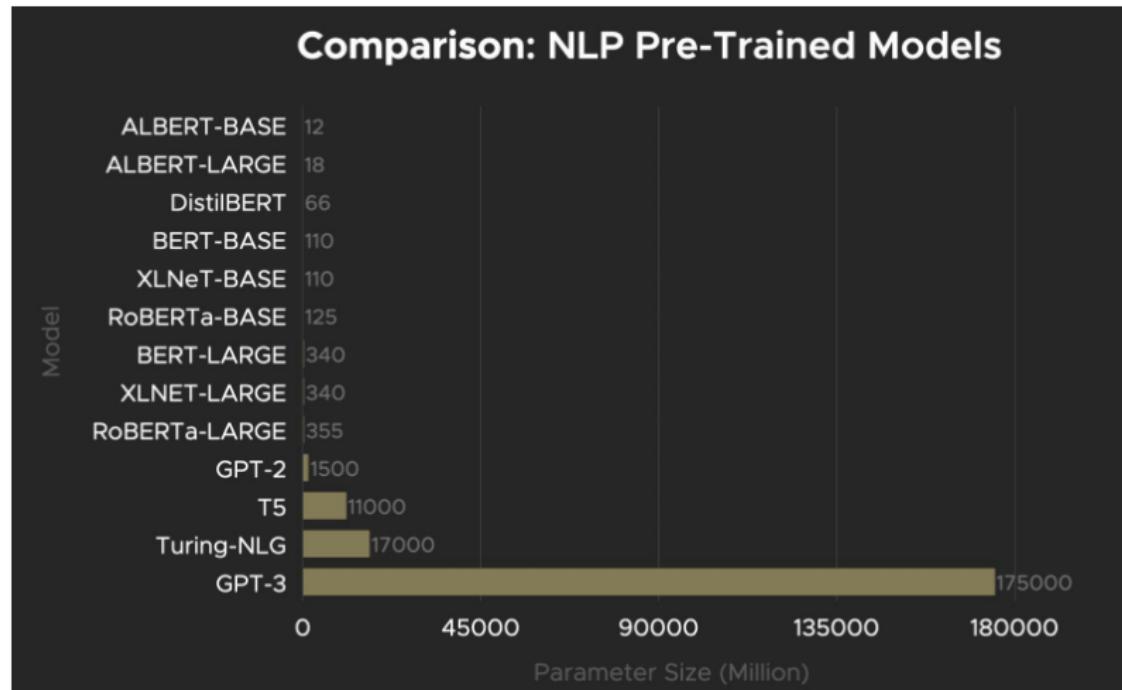
YHK

Why they are called Large? Corpus

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

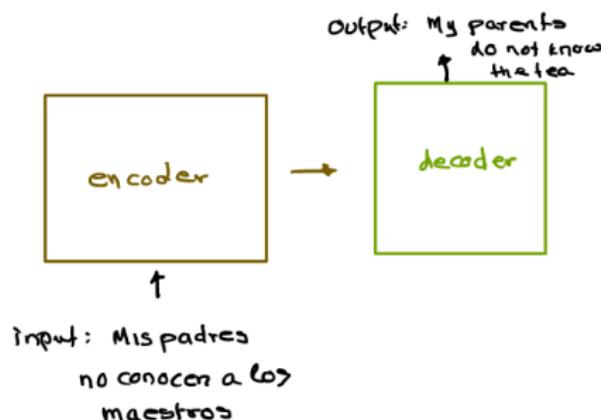
Why they are called Large? Parameters



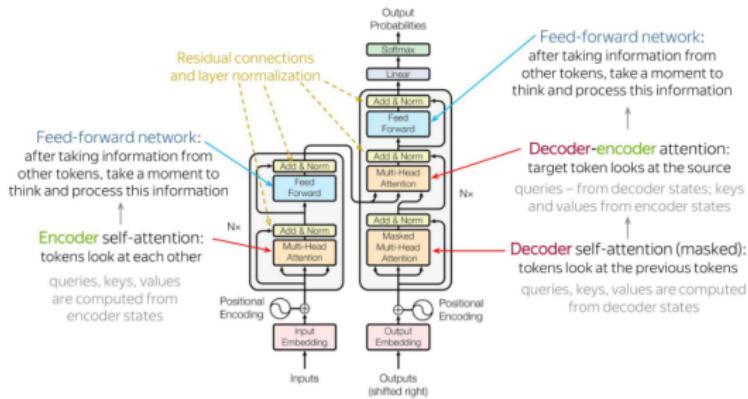
(Ref: Deus.ai <https://www.deus.ai/post/gpt-3-what-is-all-the-excitement-about>)

Underlying Architecture: Transformers

- ▶ In its heart it contains an encoding component, a decoding component, and connections between them.
- ▶ The Transformer is a model that uses attention to boost the speed with which seq2seq with attention models can be trained.
- ▶ The biggest benefit, however, comes from how The Transformer lends itself to parallelization. How?



Transformer Models

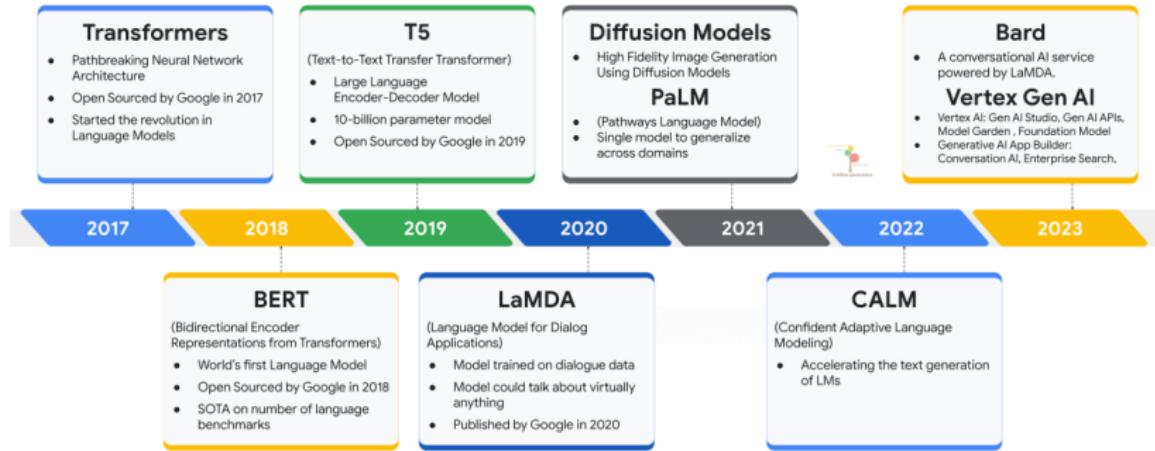


(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))

- ▶ No recurrence, so parallelization possible
- ▶ Context information captured via attention and positional encodings
- ▶ Consists of stacks of layers with various sublayers

Transformers are basis of (the most) Large Language Models

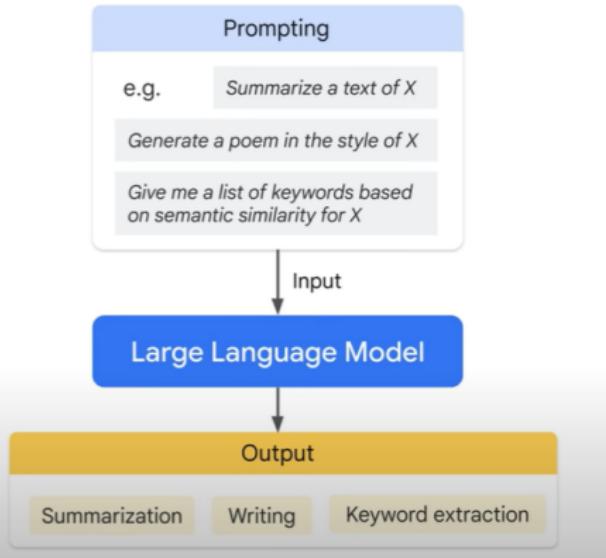
The Progress of Models ...



(Ref: Primer on LLM and Gen AI - Google Cloud)

Prompts driving Generative AI

Prompt Design:
the quality of the
input **determines** the
quality of the output.



(Ref: Introduction to Generative AI - Google Cloud Tech)

What's IN these days . . .



ChatGPT - A Tipping Point for Generative AI

- ▶ Released by OpenAI in November 2022
- ▶ Generative AI chatbot
- ▶ Rapid worldwide popularity
- ▶ 1 million users in 5 days
- ▶ Netflix took 3.5 years for same user count
- ▶ 100 million monthly active users by January 2023
- ▶ Fastest-growing application in history

What's Inside?

Technical Details

- ▶ Based on GPT3.5 Instruct architecture
- ▶ Estimated 175 billion parameters
- ▶ Fine-tuned on chat-specific task
- ▶ Curated dataset for fine-tuning



Improving ChatGPT with RLHF

- ▶ Key technique: Reinforcement Learning from Human Feedback (RLHF)
- ▶ Trains language model to align with human preferences
- ▶ Collects human feedback on model-generated text
- ▶ Updates model's parameters using feedback
- ▶ Enhances ChatGPT responses' quality
- ▶ Increases factual, informative, and creative output

Midjourney: Image Generation Model

- ▶ Developed by Midjourney Inc.
- ▶ Released in July 2022
- ▶ Architecture details undisclosed
- ▶ High-quality image generation
- ▶ Wide variety of styles and genres

Meta Releases LLaMA

Open Source LLMs Explode!

- ▶ February 2023: Meta releases LLM "LLaMA"
- ▶ LLaMA: 65-billion parameter model
- ▶ Trained on extensive text and code dataset

Significance of LLaMA Release

- ▶ One of the largest public LLMs
- ▶ Suited for complex and challenging tasks
- ▶ Open source, initially for research purposes
- ▶ Model weights leaked online, accessible to all
- ▶ Sparked development of numerous open source LLMs

Anthropic Claude

Fit Shaker: Fast and Healthy

Chat Edit Chat

Product names: HomeShaker, Fit Shaker, QuickShake, Shake Maker

Product description: A pair of shoes that can fit any foot size.

Seed words: adaptable, fit, omni-fit

Product names:

M

Here are some suggested product names based on the seed words:

- OmniFit
- AdaptiShoe
- UniSole
- One Size Fits All
- FlexiFit

Write a message...

- ▶ Uses Constitutional AI rather than RLHF
 - ▶ Constitutional AI trains to follow a set of high-level principles or rules, such as a constitution, that specify the desired behavior and outcomes of the system.
 - ▶ RLHF uses human feedback, such as ratings, preferences, or corrections, to optimize a language model or an agent's policy using reinforcement learning

(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))

Google's Latest LLM: Gemini and Gemma

Google Releases Gemini

- ▶ December 2023: Rebranded from Bard to Gemini
- ▶ Advanced multimodal LLM for language, audio, code, and video
- ▶ Trained on extensive datasets for improved contextual understanding
- ▶ Real-time information access through Google Search integration

Introducing Gemma

- ▶ Specialized version of Gemini for personal assistance, on Hugging Face
- ▶ Helps manage tasks, schedules, and reminders effectively
- ▶ Deep integration with Google Calendar and productivity tools

Overview

Overview of Prompt Engineering

▶ Introduction

- ▶ Large Language Models (LLMs) differ from traditional ML models.
- ▶ LLMs provide unique insights without requiring retraining.

▶ Transformational Impact

- ▶ LLMs have catalyzed a transformative wave in programming.
- ▶ Enables effortless computer programming through simple text prompts.

▶ Prompt Engineering Technique

- ▶ Technique for directing LLM responses without altering model weights.
- ▶ Relies on strategic in-context prompting.
- ▶ Art of effectively communicating with AI for desired outcomes.

▶ Application Spectrum

- ▶ Applied across various tasks: question-answering, arithmetic reasoning, etc.
- ▶ Serves as a versatile tool to explore LLM boundaries and potentials.



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>)

New Programming Language?



Andrej Karpathy ✅
@karpathy

...

The hottest new programming language is English

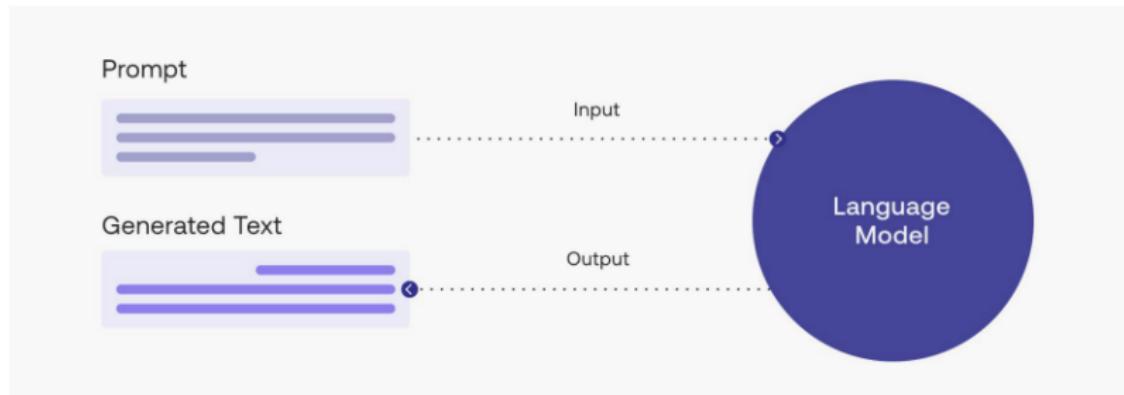
1:44 AM · Jan 25, 2023 · 1.9M Views

2,050 Retweets 284 Quote Tweets 17.9K Likes

(Ref: Prompt Engineering Sudalai Rajkumar)

What is Prompt Engineering?

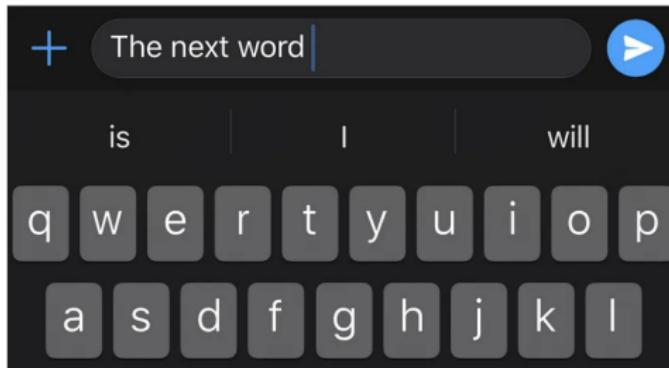
Prompt engineering is an NLP concept that involves discovering inputs that yield desirable or useful results



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

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

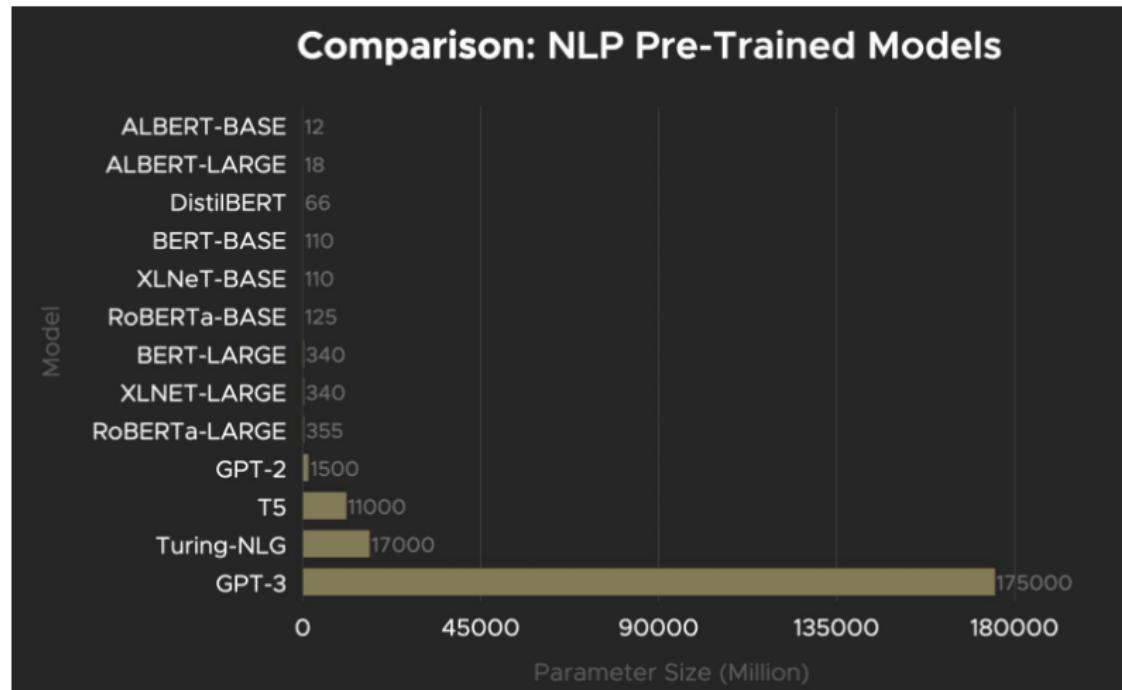
Evolution of Language Models

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>)

Large Language Models - Comparison



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

What is Prompt Engineering?

- ▶ For prompt What `is` 1,000,000 * 9,000? GPT-3 (text-davinci-002) (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>)

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>)

What is Prompt Engineering?

But need to tell, for sure, else, nothing

The screenshot shows a user interface for a generative AI system. At the top left is a button labeled "Prompt". At the top right is a button labeled "Load model config". Below these, the word "No instruction" is followed by an arrow pointing to a text input field. Inside the input field, there are two lines of text: "English: How do you reset your password?" and "Spanish:". Below this, a larger block of text is displayed in Spanish: "Para resetear tu contraseña, ve a la página de inicio de sesión de la aplicación y haz clic en el enlace \"Olvidé mi contraseña\". Se te enviará un correo electrónico con instrucciones para restablecer tu contraseña." An arrow points from the word "Incorrect response" at the bottom right towards the end of this Spanish text.

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

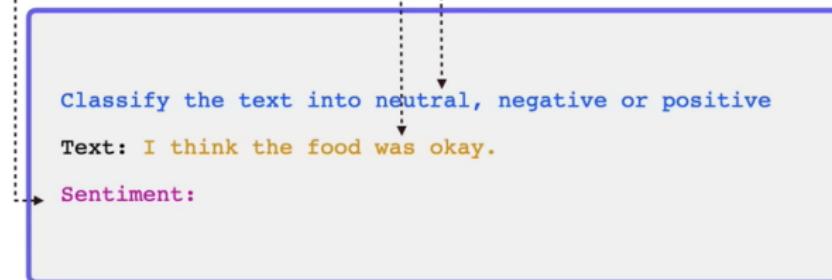
Engineering of Prompts



Elements of Prompt

A prompt is composed of:

- Instructions
- Context
- Input data
- Output indicator



(Ref: Prompt Engineering Overview - Elvis Saravia)

Settings of Prompt

- ▶ 'temperature': before applying the softmax function, temperature is used to 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 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.

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/>)

Length Control

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

S

Write a 150 word summary on UPI



UPI (Unified Payments Interface) is an innovative, instant payment system developed by the National Payments Corporation of India (NPCI). It enables instant money transfer

(Ref: Prompt Engineering Sudalai Rajkumar)

Tone Control

Specify specific words or phrases that indicate the desired tone

S

Write a **polite response** for the below chat

Customer: This is your fault. How many days should I wait for the refund?

Agent:



I apologize for the inconvenience caused. Can you please provide me with some more details about the issue, so I can assist you with the refund process? Thank you for your understanding and patience.

(Ref: Prompt Engineering Sudalai Rajkumar)

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]

(Ref: Prompt Engineering Sudalai Rajkumar)

Audience Control

Specify the desired audience.

S

Explain internet banking to a 8 year old kid



Internet banking is like using a magic bank that you can access from anywhere in the world, as long as you have a computer or a smartphone. It's a way to do banking tasks like checking your balance, paying bills, or sending money to your friends and family, all from the comfort of your own home.

(Ref: Prompt Engineering Sudalai Rajkumar)

Context Control

Specify the information about the context.

S

Write an reminder message to the customer whose home loan EMI is due tomorrow



Dear [Customer Name],

This is a friendly reminder that your Home Loan EMI of INR [Amount] is due tomorrow. We value your partnership with us and want to ensure that your loan repayment is smooth and seamless.

We request you to kindly make the payment before the due date to avoid any late payment charges. You can make the payment through internet banking, mobile banking, auto-debit, or by visiting any of our branches.

(Ref: Prompt Engineering Sudalai Rajkumar)



Zero-shot Prompting

► Definition

- Zero-shot learning: Task given to the model without specific output examples.
- No prior examples indicating the desired output.

► Example Scenario

- Input: A sentence without examples.
- Task: Model predicts sentiment of the given sentence.

► Illustrative Example (DAIR-AI)

- **Prompt:** Classify the text into neutral, negative, or positive.
- **Text:** I think the vacation is okay.
- **Output:** Neutral

Few-shot Prompting

► **Definition**

- ▶ Few-shot learning: Model provided with a small number of quality examples (input and desired output).
- ▶ Helps the model better understand human intention and criteria for accurate outputs.

► **Comparison with Zero-shot Learning**

- ▶ Few-shot learning often yields better performance compared to zero-shot learning.
- ▶ However, may consume more tokens and encounter context length limitations for long input/output text.

► **Application in Large Language Models (e.g., GPT-3)**

- ▶ LLMs excel in zero-shot capabilities but may face performance issues in complex tasks.
- ▶ Few-shot learning enhances performance by offering in-context learning through task-specific examples.



Few-shot Prompting

► Example Scenario (Brown et al.)

► Prompt:

- A "whatpu" is a small, furry animal native to Tanzania.
- Example sentence: We were traveling in Africa and we saw these very cute whatpus.
- To do a "farduddle" means to jump up and down really fast.
- Example sentence: When we won the game, we all started to farduddle in celebration.

► Output:

- The model, given one example, generates the answer for the next.

Chain-of-Thought (CoT) Prompting

► **Introduction**

- CoT Prompting introduced in Wei et al. (2022).
- Enables LLM to tackle complex tasks by breaking them down into constituent steps.
- Complex reasoning through intermediate reasoning steps.

► **Combining with Few-shot Prompting**

- Combine with few-shot prompting for better results on complex tasks.
- Requires reasoning before responding.



Chain-of-Thought (CoT) Prompting

Standard Prompting

Model 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

Model 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.

(Ref: Primers • Prompt Engineering - Aman AI)

Zero-shot CoT Prompting

- ▶ Zero-shot CoT introduced by Kojima et al. (2022).
- ▶ Involves appending "Let's think step by step" to the prompt.
- ▶ Aims to improve model performance.



Without Zero-shot CoT Prompting

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?

2

Response: 11 apples

With Zero-shot CoT Prompting

- 1 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?
- 3 Let's think step by step.
- 5 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.
- 7 Then you bought 5 more apples, so now you had 11 apples.
Finally, you ate 1 apple, so you would remain with 10 apples.

Zero-shot CoT Prompting

► Overview

- ▶ CoT prompting generates reasoning chains without additional training.
- ▶ Sequence of short sentences with an explanation and prediction pair.
- ▶ No need for extra training or fine-tuning.

► Benefits

- ▶ Step-by-step reasoning logic leading to the final answer.
- ▶ Particularly beneficial for complex reasoning tasks and larger models.



Few-shot CoT Prompting

Few-shot CoT allows the model to view a few demonstrations of high-quality reasoning chains.

- Tom **and** Elizabeth have a competition to climb a hill.
2 Elizabeth takes 30 minutes to climb the hill.
Tom takes four times as **long** as Elizabeth does to climb the hill.
4 How many hours does it take Tom to climb up the hill?

6 **Answer:**

8 It takes Tom $30 * 4 = <<30*4=120>>120$ minutes to climb the hill.
It takes Tom $120 / 60 = <<120/60=2>>2$ hours to climb the hill.
10 So the answer **is** 2.

Few-shot CoT Prompting

1 Jack **is** a soccer player. He needs to buy two pairs of socks **and** a pair of soccer shoes.

Each pair of socks cost \$9.50, **and** the shoes cost \$92.

3 Jack has \$40. How much more money does Jack need?

5 Answer:

The total cost of two pairs of socks **is** $\$9.50 \times 2 = \19 .

7 The total cost of the socks **and** the shoes **is** $\$19 + \$92 = \$111$.

Jack needs $\$111 - \$40 = \$71$ more.

9 So the answer **is** 71.

11 Question:

Marty has 100 centimeters of ribbon that he must cut into 4 equal parts.

13 Each of the cut parts must be divided into 5 equal parts.

How **long** will each final cut be?

15

Answer:

Chain-of-Thought (CoT) Prompting Example

- The odd numbers add up to an even number: 4, 8, 9, 15, 12, 2, 1.
2 A: Adding odd numbers (9, 15, 1) gives 25. The answer **is** False.
- 4 The odd numbers add up to an even number: 17, 10, 19, 4, 8, 12, 24.
A: Adding odd numbers (17, 19) gives 36. The answer **is** True.
- 6 The odd numbers add up to an even number: 16, 11, 14, 4, 8, 13, 24.
8 A: Adding odd numbers (11, 13) gives 24. The answer **is** True.
- 10 The odd numbers add up to an even number: 17, 9, 10, 12, 13, 4, 2.
A: Adding odd numbers (17, 9, 13) gives 39. The answer **is** False.
- 12 The odd numbers add up to an even number: 15, 32, 5, 13, 82, 7, 1.
14 A:
- 16 Adding **all** the odd numbers (15, 5, 13, 7, 1) gives 41.
The answer **is** False.
- 18 Wow! We can see a perfect result when we provided the reasoning step.
20 In fact, we can solve this task by providing even fewer examples,
i.e. just one example seems enough ...

Chain-of-Thought (CoT) Prompting Example (Contd.)

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

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

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

5 A:

7 Adding **all** the odd numbers (15, 5, 13, 7, 1) gives 41.

9 The answer **is** False.

11 Keep **in** mind that the authors claim that this **is** an emergent ability that arises with sufficiently large language models.



Sample (Simple) Gen AI 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)

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)

Text Translation

S

Translate the following to tamil

I made a transaction yesterday



"நான் நேற்று ஒரு பரிவர்த்தனை செய்தது"

(Ref: Prompt Engineering Sudalai Rajkumar)

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)

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)

Image Generation

/v4-upscale 12 hrs ago

A surreal castle on a floating island, by John Byrne and Skottie young and Greg Smallwood, highly...

moongoat ...



(Ref: Prompt Engineering Sudalai Rajkumar)

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

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/>)



Programmatic Calling of Prompt

```
import openai
2 import os
4
from dotenv import load_dotenv, find_dotenv
_ = load_dotenv(find_dotenv())
6
openai.api_key = os.getenv('OPENAI_API_KEY') # for langchain it does it
    automatically
8
def get_completion(prompt, model="gpt-3.5-turbo"):
10    messages = [{"role": "user", "content": prompt}]
    response = openai.ChatCompletion.create(
12        model=model,
        messages=messages,
14        temperature=0, # this is the degree of randomness of the model's output
    )
16    return response.choices[0].message["content"]
```

Making of a Sandwich



Basic

Explain how to make a peanut butter and jelly sandwich



Adding Roles

As a chef, explain to your assistant how to make a peanut butter and jelly sandwich



Adding Constraints

Make a nut-free version of the sandwich due to a customer's nut allergy

Adding Examples

Create two unique variations of the classic sandwich. Banana Nut Crunch: . . .

Triple Berry Blast: . . .



Adding Contextual Information

As the head chef at 'The Sandwich Haven,' guide your new assistant to create specials for the menu



Incorporating Feedback

Improve the sandwich based on customer feedback for less sweetness and a creative twist



Time Constraints and Prioritization

Prepare an alternative fruit version for testing within a tight deadline



Incorporating Multidisciplinary Knowledge

Use food presentation and garnishing techniques for a visually appealing sandwich



Addressing Dietary Preferences

Prepare a vegan version using plant-based alternatives for all ingredients

Reflection and Iteration

Reflect on feedback and iteratively refine the sandwich for better taste and appeal



Self-Criticism

Explain how to make a peanut butter and jelly sandwich. Please re-read your above response. Any mistakes? If so, please identify and make the necessary edits.



Chain-of-Thought

Explain how to make a peanut butter and jelly sandwich. Let's think step by step.



Self-Consistency

Here are recipes of multiple sandwiches. Sandwich 1: recipe 1. Sandwich 2: recipe 2. Explain how to make a peanut butter jelly sandwich.

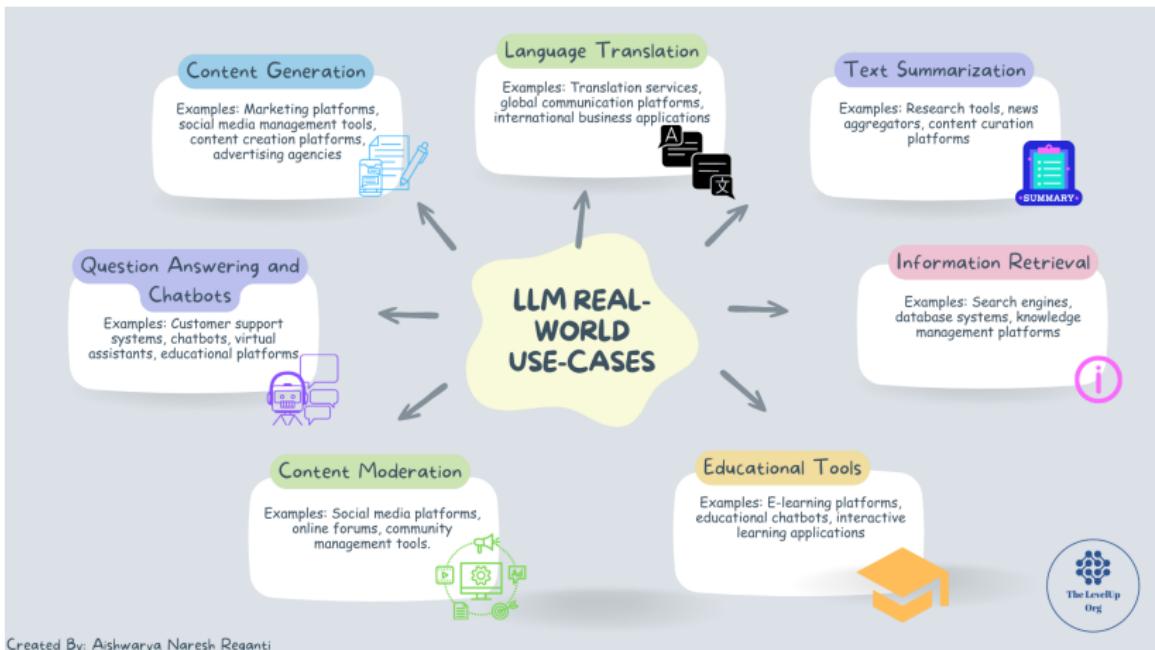


Conclusions of Generative AI

(Ref: 2023 Kaggle AI Report on Generative AI, by Trushant Kalyanpur)



LLM Real World Use Cases



(Ref: Applied LLMs Mastery 2024 - Aishwarya Reganti)

Advancements in Text-to-Image Generative AI

- ▶ Notable models: DALL-E/DALL-E 2, Midjourney, Stable Diffusion
- ▶ Creative expression, streamlined design
- ▶ Realistic, high-quality image generation
- ▶ Concerns: misuse, ethical implications
- ▶ Deepfakes, synthetic media for misinformation
- ▶ Risk of generating illegal, toxic content
- ▶ Challenges in ethical responsibility, moderation

Copilots: Revolutionizing Coding

- ▶ AI assistance for software developers
- ▶ 92% programmers use AI tools (Github survey)
- ▶ Copilot users 55% faster in project completion
- ▶ Potential \$1.5 trillion GDP boost (productivity study)
- ▶ AI tools like Copilot enhance speed, efficiency
- ▶ Fewer errors in coding processes



Industry Giants' Role in Text Generative AI

- ▶ Google, Meta, OpenAI - Pivotal contributions
- ▶ PaLM, Galactica, ChatGPT, GPT4 releases
- ▶ ChatGPT: Turning point in Text Generative AI
- ▶ LLMs for content creation, writing, storytelling
- ▶ Analyzing, organizing large textual data
- ▶ Efficient search engines, knowledge systems

Advancements in LLM Accessibility

- ▶ Meta's LLaMa: Open-source alternatives to ChatGPT
- ▶ Google's Bard: AI chatbot response to ChatGPT
- ▶ QLoRA: Fine-tuning LLMs on consumer GPUs
- ▶ Broadening access to advanced technology
- ▶ LLMs empower diverse applications
- ▶ LLMs: Bridging the gap between innovation and accessibility



References

- ▶ Introduction to Generative AI - Google Cloud Tech
- ▶ Generative AI Presentation - Laura Worden



Newsletters to subscribe

- ▶ **The Batch by DeepLearning.AI:**
 - ▶ Summarizes diverse AI news with nuanced viewpoints.
 - ▶ Andrew Ng's thought leadership adds significant value.
- ▶ **The Rundown AI by Rowan Cheung:**
 - ▶ Go-to for generative AI events and product innovations.
 - ▶ Quick rundown with bullet point details for easy comprehension.
- ▶ **AI Supremacy by Michael Spencer:**
 - ▶ Personal writing style with in-depth exploration.
 - ▶ Offers multiple perspectives on AI topics.
- ▶ **Ahead of AI by Sebastian Raschka, PhD:**
 - ▶ Technical focus covering applied deep learning and generative AI.
 - ▶ Valuable insights for those seeking in-depth technical content.
- ▶ **To Data and Beyond by Youssef Hosni:**
 - ▶ Resource hub for hands-on projects, learning roadmaps, and research papers.
 - ▶ Ideal for those looking to dive into practical aspects of AI.

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

(<https://www.linkedin.com/in/yogeshkulkarni/>, QR by Hugging Face

QR-code-AI-art-generator, with prompt as "Follow me")

