

INTRODUCTION TO AI AND GENERATIVE AI

Yogesh Haribhau Kulkarni



Outline

① GENAI OVERVIEW

② PROMPTENGG DEMO

③ CONCLUSIONS

④ REFERENCES

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](mailto:yogeshkulkarni@yahoo.com)



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

Introduction to Generative AI

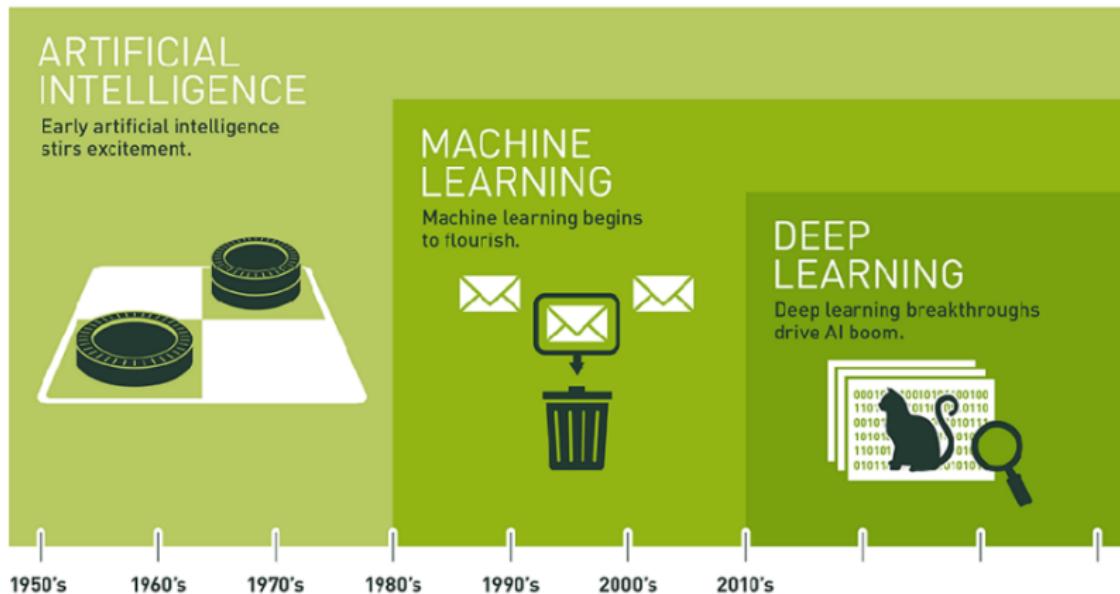
YHK

Introduction

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

Relationship between AI, ML, DL

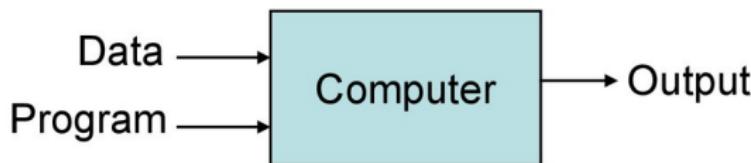
First, lets see whats AI-ML-DL and relationship among them.



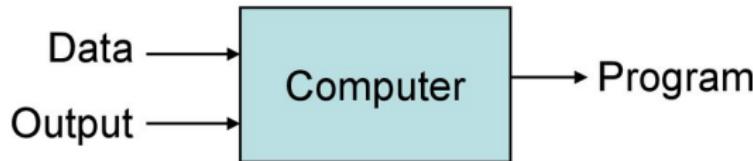
(Ref: <https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/>)

Traditional vs. Machine Learning?

Traditional Programming



Machine Learning



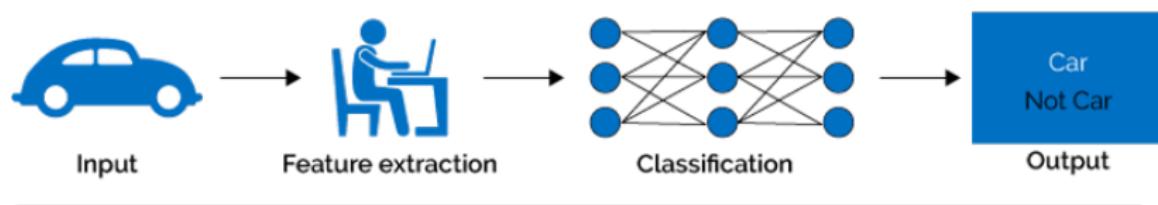
Why Machine Learning?

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

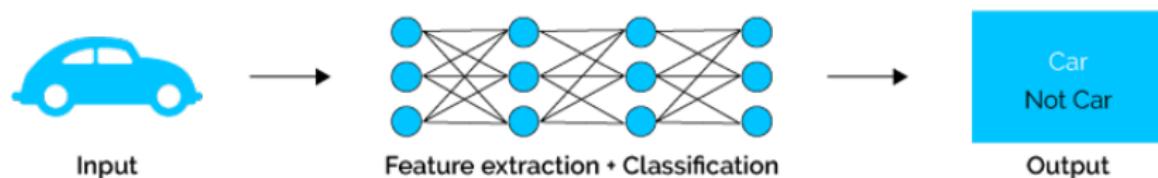
ML vs DL: What's the difference?

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 ...

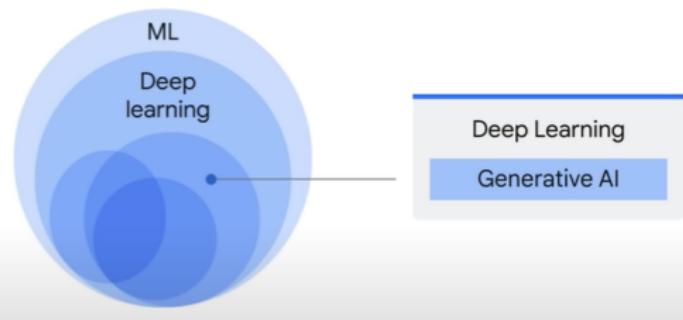
- ▶ You have lots of data (about 10k+ examples)
- ▶ The problem is “complex” - speech, vision, natural language
- ▶ The data is unstructured
- ▶ Techniques to model ‘ANY’ function given ‘ENOUGH’ data.

(Ref: Introduction to TensorFlow 2.0 - Brad Miro)



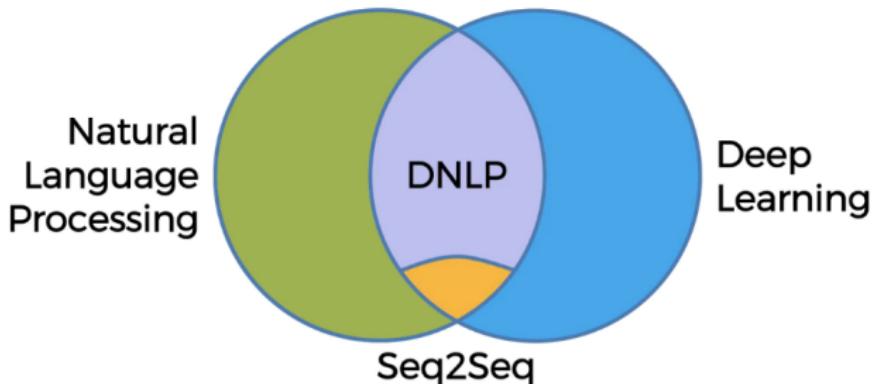
What is Gen AI wrt AI, ML, DL

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

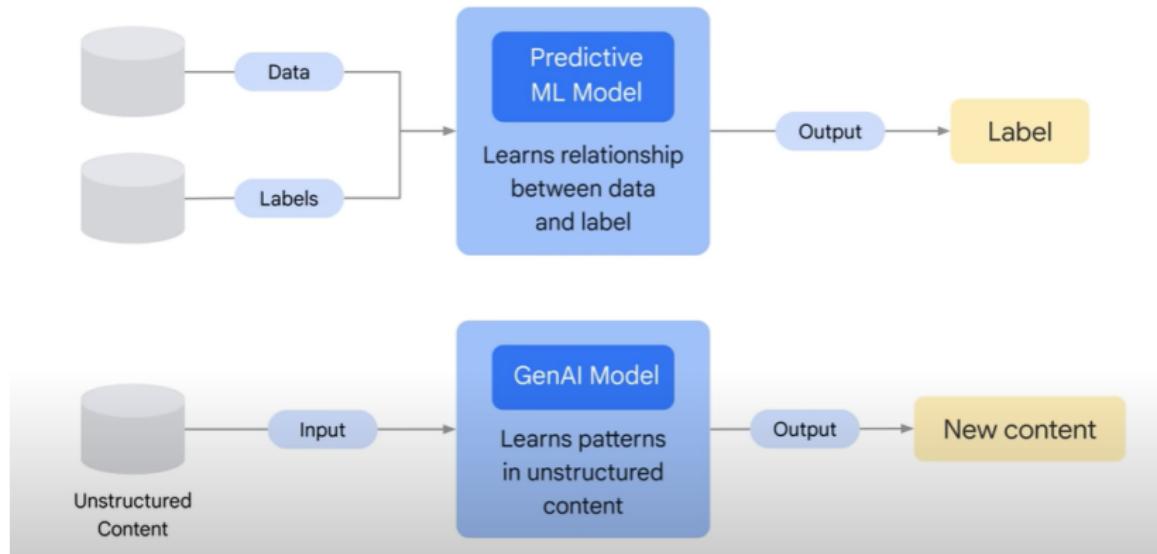


Generative

- 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

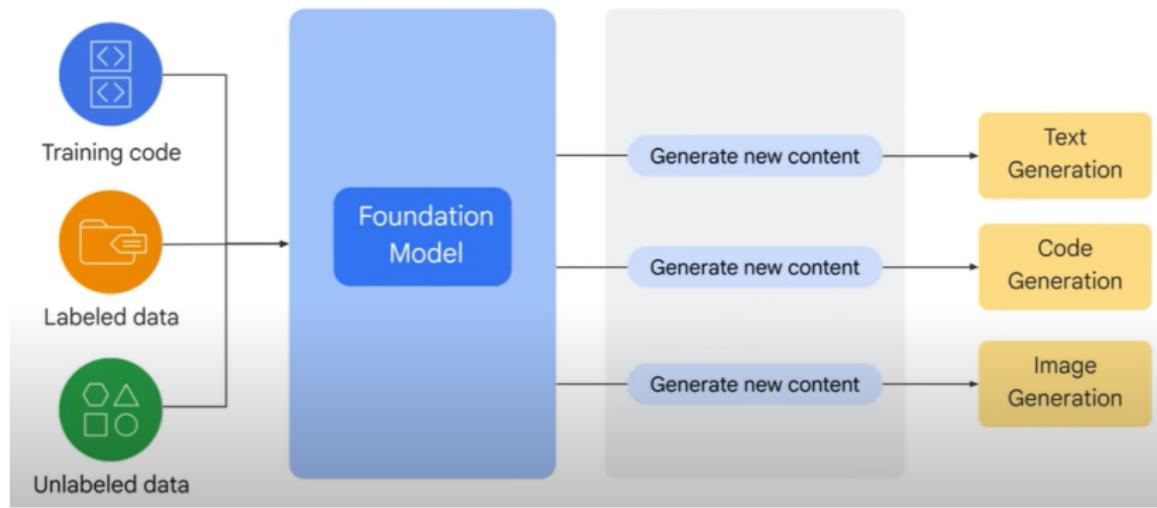
(Ref: Introduction to Generative AI - Google Cloud Tech)

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

YHK

Difference across technologies, old to new

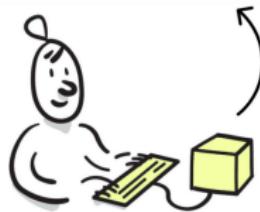
Lets see how the solutions to the problem of detecting a cat from images using traditional programming, deep learning, and generative AI, respectively.



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.

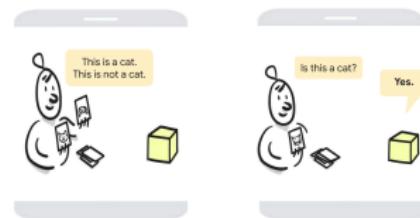
```
cat:  
  type: animal  
  legs: 4  
  ears: 2  
  fur: yes  
  likes: yarn, catnip
```



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

Deep Learning

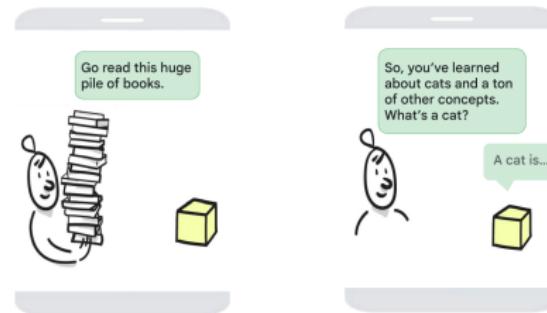
- ▶ Deep learning utilizes neural networks to automatically learn features for cat detection.
- ▶ Convolutional Neural Networks (CNNs) are particularly effective for image classification tasks.
- ▶ Large labeled datasets of cat images are used to train the network.
- ▶ The network learns to identify 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

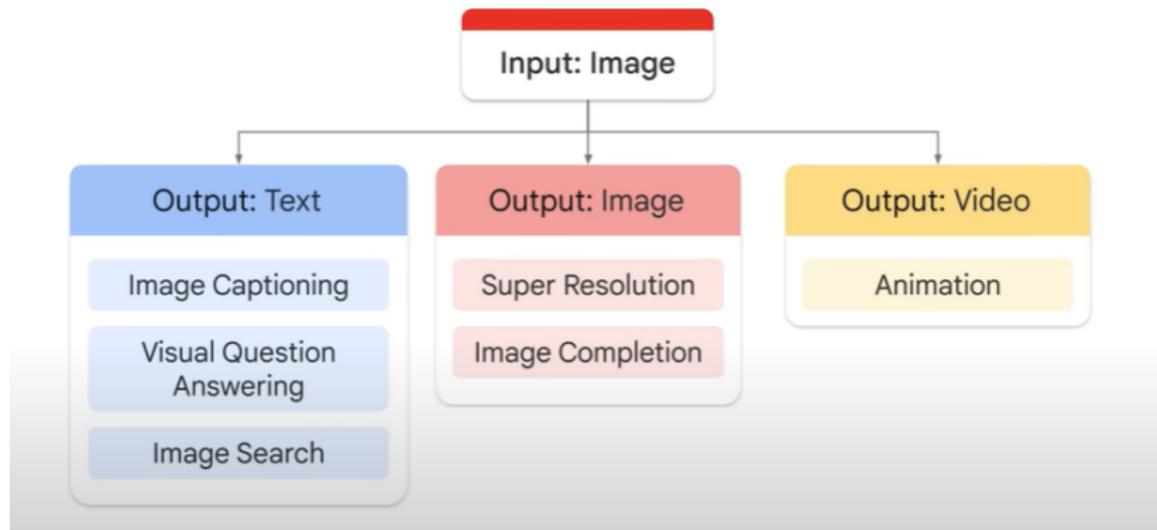
- ▶ Generative AI focuses on generating new data, including images of cats.
- ▶ Generative Adversarial Networks (GANs) are used to generate realistic cat images.
- ▶ The GAN consists of a generator and a discriminator that compete against each other.
- ▶ The generator learns to generate increasingly 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

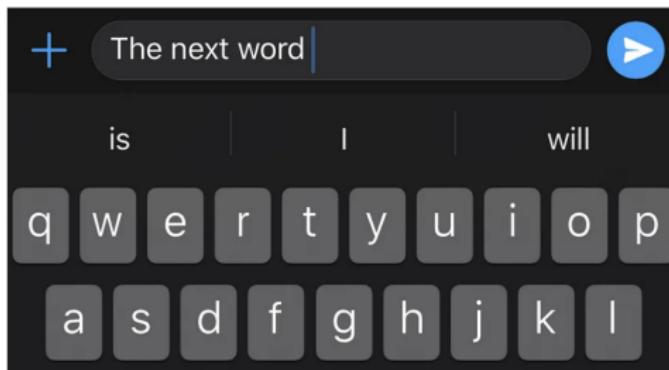


(Ref: Introduction to Generative AI - Google Cloud Tech)

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.



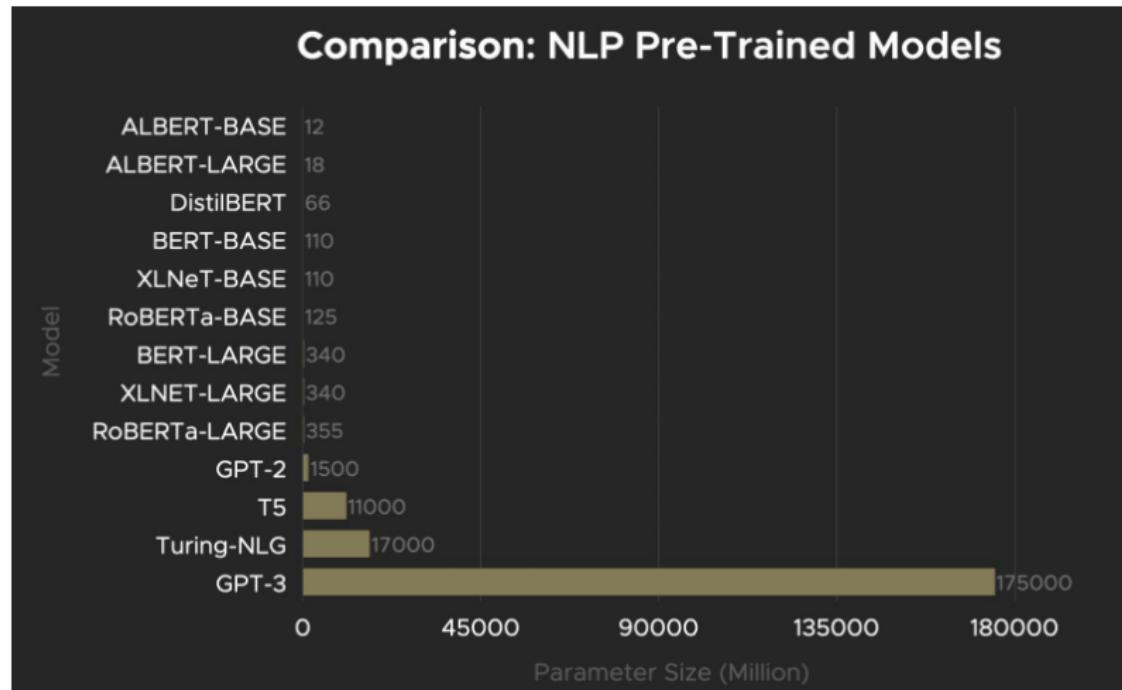
Next word prediction using language modeling in keyboards(Mandar Deshpande)

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

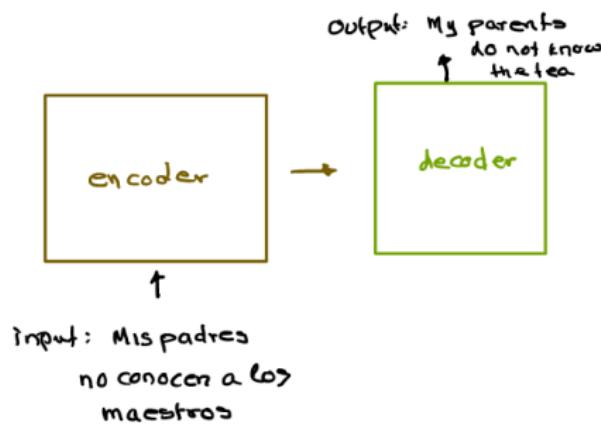
Why they are called Large? Parameters



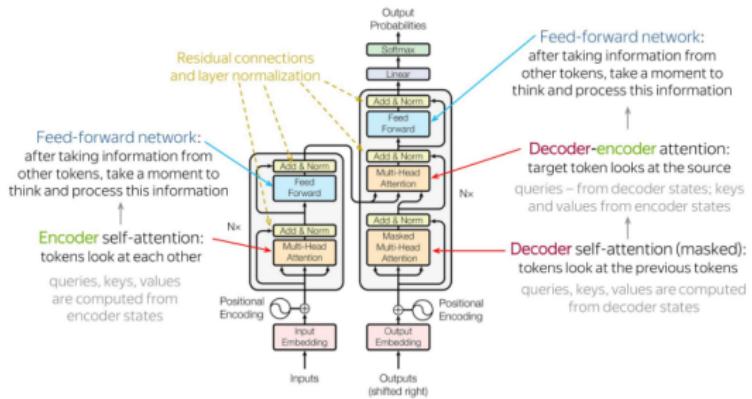
(Ref: Deus.ai <https://www.deus.ai/post/gpt-3-what-is-all-the-excitement-about>)

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

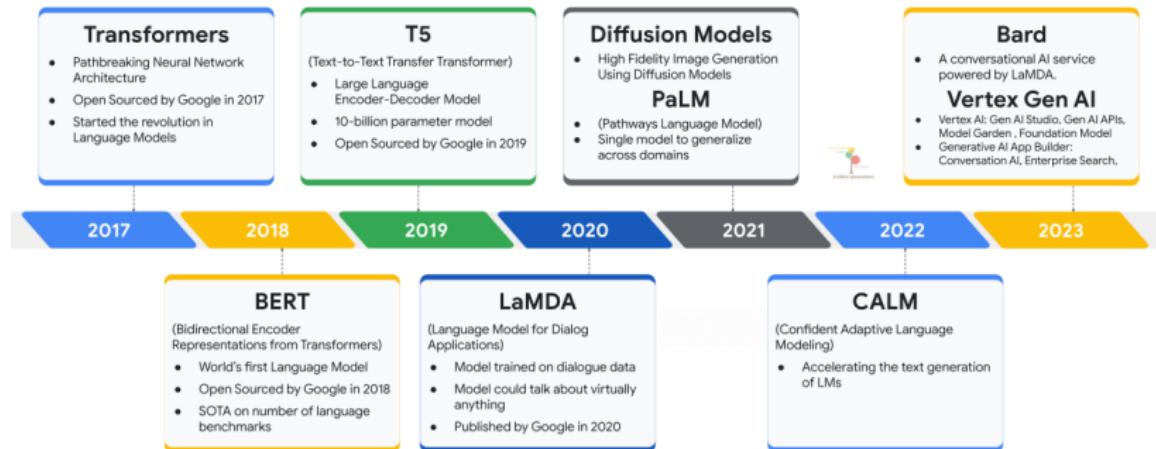


(Ref: The Complete Prompt Engineering for AI Bootcamp (2023))

- ▶ 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

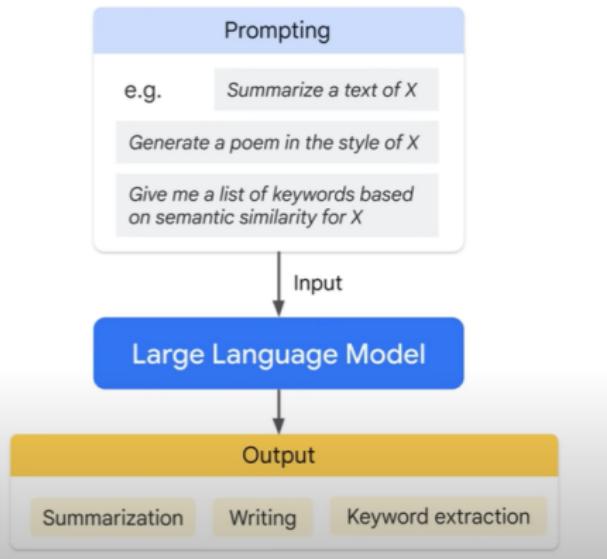
The Progress of Models ...



(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 . . .

YHK

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

The screenshot shows a user interface for generating product names. At the top, there's a header with a square icon, three dots, and a search bar containing 'Fit Shaker: Fast and Healthy'. Below the search bar are two buttons: 'Chat' and 'Edit Chat'. A purple sidebar on the right contains the following text:
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

Below the sidebar, a light gray box displays a list of suggested product names based on seed words:

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

At the bottom left is a text input field with placeholder text 'Write a message...' and a blue send button with a white arrow icon. At the bottom right is a small circular profile picture.

- ▶ 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

Making of a Sandwich

YHK

Basic

Explain how to make a peanut butter and jelly sandwich

YHK

Adding Roles

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

YHK

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

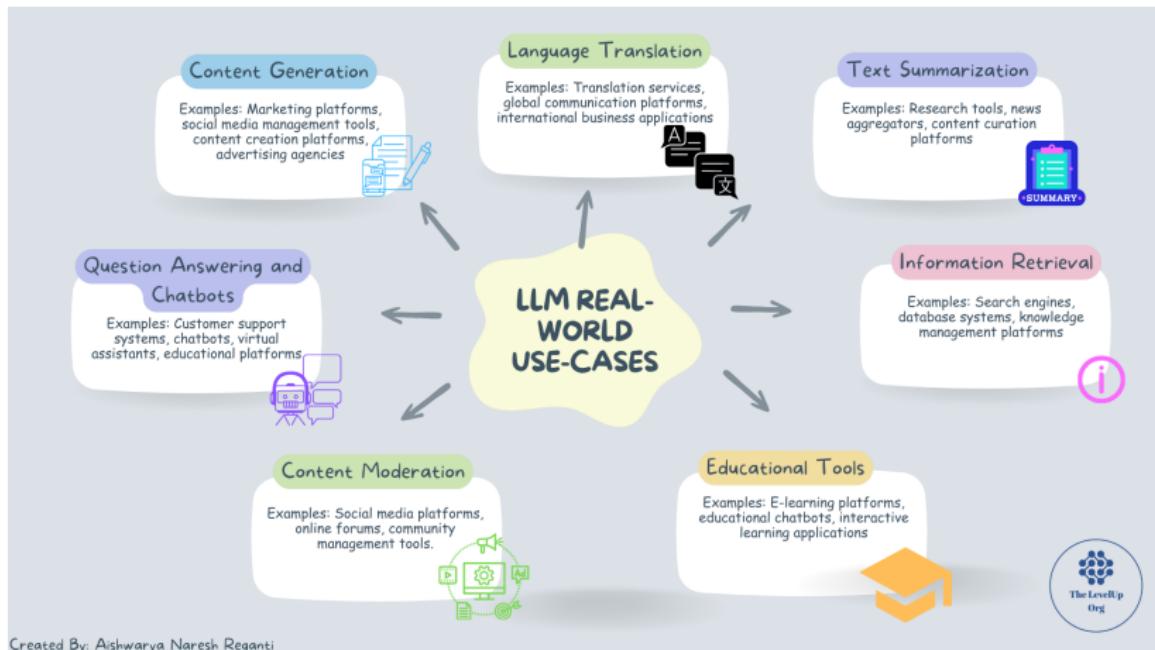
Here are recipes of multiple sandwiches. Sandwich 1: recipe 1. Sandwich 2: recipe 2. Explain how to make a peanut butter jelly sandwich.



Conclusions of Generative AI

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

LLM Real World Use Cases



Created By: Aishwarya Naresh Reganti

(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")

