Generative AI

A brief overview on Large Language Models

RAG: Retrieval Augmented Generation

Surya Pradeep Kumar

CONTENTS

01 | What is a Language Model?

02 | What makes a model “Large”?

03 | Limitations of LLMs

04 | RAG: Retrieval & Generation

What is a Language Model?

Imagine you're typing a message on your phone, and it suggests the next word you might want to use.
That's a language model!

How would you complete this sentence: "I woke up early, got ready, and made a ..." ?

01 "cup of coffee"

02 "plan for the day"

03 "glass of orange juice and toast"

04 "to-do list"

05 "plate of bacon and eggs"

➢ Large Language models leverage massive amounts of data and becomes incredibly good at predicting the next word in any sequence.

➢ Prective task for the model: **Language Modeling** → Learning to predict next word

Predict the next word using the input sequence:
"I woke up early, got ready, and made a ..." ?

→ Language Model →

Word	Probability
cup	**0.085**
plan	0.071
glass	0.002
…	…
plate	0.005

Auto-regressively predicts for next word / token:

"I woke up early, got ready, and made a cup ..." ?

❖ Transformers and Self-Attention: Magic Behind LLMs

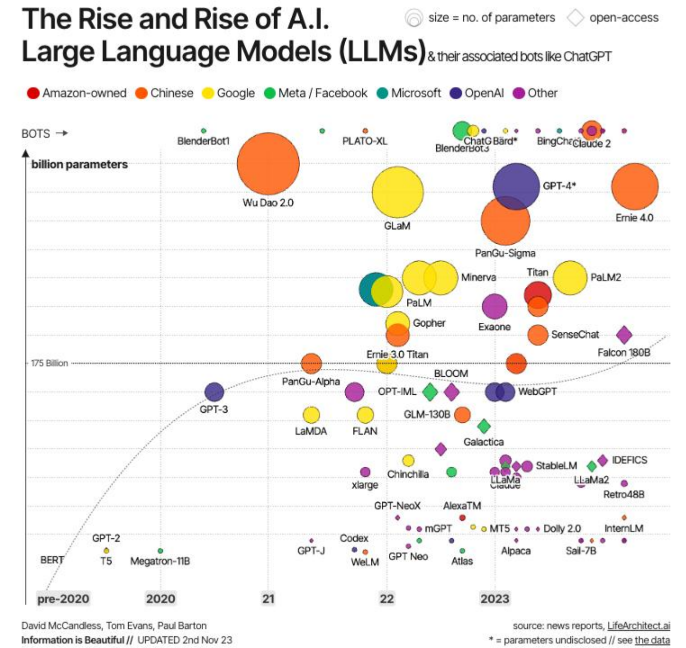
- A large language model is a trained deep-learning model that understands and generates text in a human-like fashion. Behind the scenes, it is a large **transformer** model that does all the magic.
- Leveraging massive training data, they are able to understand natural language and generate it. They learn by analyzing patterns in the text to learn how words and sentences are structured and how they relate to each other.
- They are shown to have a meaningful understanding of general language and that they are able to (re)produce information related to the information that was present in their training data
- Examples: GPT-3 and GPT-4 from OpenAl, LLaMA from Meta, and Gemini from Google.

- Crux of the transformer is the **self-attention mechanism** – Think of it as your brain's spotlight! When reading a sentence, we focus on important words to understand the context.
- The self-attention mechanism allows the inputs to interact with each other ("self") and find out who they should pay more attention to ("attention"). The outputs are aggregates of these interactions and attention scores.
- Self-attention enables the model to weigh the importance of different words in an input sentence and dynamically adjust their influence on the output. In other words, It considers the entire input sequence and determines the importance of each word to the current word based on their semantic and syntactic relationship.
- Thanks to Transformers, LLMs can do cool stuff like translating languages, answering questions, writing code, and even poems and entire stories!

➢ 2017, Vaswani et al. published a paper, "Attention is All You Need," which establishes transformer model. https://arxiv.org/abs/1706.03762

❖ What makes a model "Large"?

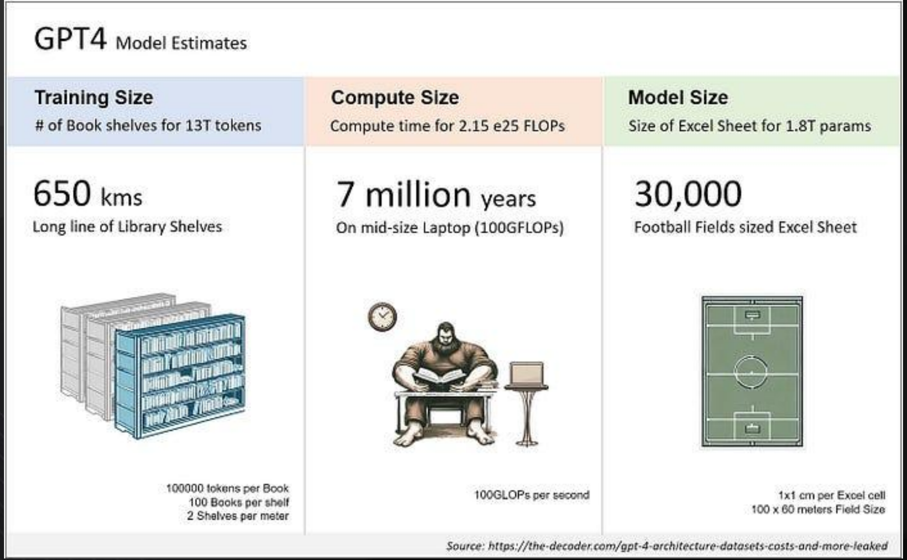
- The definition of large is a bit fuzzy: But think of models like BERT (with 110 million parameters) or GPT4 (up to 1.76 trillion parameters). These models are pre-trained on datasets raining from billion tokens up to trillion tokens scale.
- Parameters are like the model's "knowledge weights." The more parameters, the more patterns it could remember and the more knowledge it could assimilate in its weights.



The Rise and Rise of A.I. Large Language Models (LLMs) & their associated bots like ChatGPT

David McCandless, Tom Evans, Paul Barton
Information is Beautiful // UPDATED 2nd Nov 23

source: news reports, LifeArchitect.ai
* = parameters undisclosed // see the data



GPT4 Model Estimates

Training Size	**Compute Size**	**Model Size**
# of Book shelves for 13T tokens	Compute time for 2.15 e25 FLOPs	Size of Excel Sheet for 1.8T params
650 kms	7 million years	30,000
Long line of Library Shelves	On mid-size Laptop (100GFLOPs)	Football Fields sized Excel Sheet
100000 tokens per Book 100 Books per shelf 2 Shelves per meter	100GLOPs per second	1x1 cm per Excel cell 100 x 60 meters Field Size

Source: https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked

- ➢ https://www.datacamp.com/blog/what-is-an-llm-a-guide-on-large-language-models
- ➢ https://medium.com/@georgeanil/visualizing-size-of-large-language-models-ec576caa5557

What does GPT mean ?

Generative

Generates sequences through auto-regressive **Next Word Prediction**

Pre-trained

LLM is **pretrained on massive amounts of text** from internet (common crawl ...), books, etc., and other sources

Transformer

Neural Network Architecture introduced in 2017 for language translation task. Crux of the model is **Attention Mechanism**

Limitations of LLMs

01 Hallucinations

As models are desinged to produce coherent text, they can “hallucinate” and generate text that is incorrect, but seems plausible. So these models might gaslight the users into believing wrong info.

02 Learning

Large Language Model as they exist now, don’t have memory i.e, they are stateless and don’t ‘remember’ or ‘learn’ from previous interactions.

ChatAGPT’s conversation memory is all done with prompt trickery: previous prompts are injected into the stateless model to make it seem the model remembers previous messages!

03 Custom Data

Large Language Models don’t provide responses based on our Custom Domain Data (databases, content, ...).
So, the model only draws from it’s own pre-existing knowledge based on it’s static training which can grow stale and **Out of Date** over time

04 Token Limits

Models are limited by the TOKEN_LIMIT, and most models can process, at best, a few pages of total input (prompt context) / output (completion). This means we can’t just feed a model an entire document in the prompt, and ask for a summary or extract information from the document.

05 Transparency

Even if the model generates a correct answer, we may not be able to provide transparency / explainibility to what source content helped generate the answer as an LLM by itself, functions like a black box

06 Ehical Concerns

Potential bias, hate, abuse, harm, ethical concerns, etc: sometimes, answers generated by an LLM can be outright harmful

07 Training Cost

A 70B parameter model like LLAMA2 might need ~2048 A100 GPUs for a month to train, adding up to $20–40M training cost, not to mention what it takes to download and store the data

Even fine-tuning the models, although acheivable, is a considerably expensive endeavour

Enter RAG: Retrieval Augmented Generation

Rather than rely on model’s old knowledge or fine-tune the model on our knowledge base or pass the entire context possible with each prompt, RAG adds an **information retrieval** mechanism to **augment** the user prompt with relevant context to **generate** a final structured response from the LLM

01

Real-Time Context

Dynamic data can be used to infuse the LLM with latest, relevant knowledge

02

Private Data

Our own custom, private information can be used to enable the LLM to work in custom domains

03

Grounding the LLM

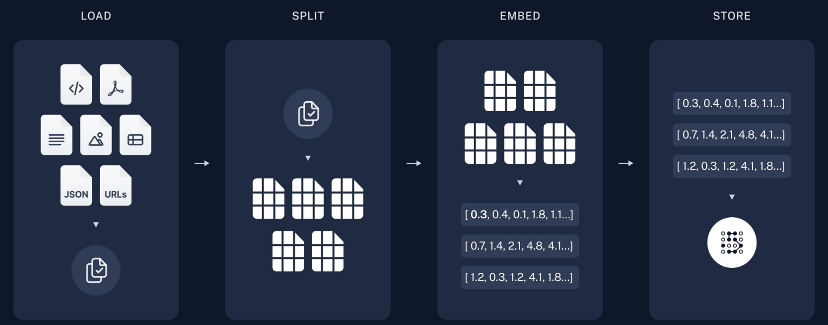
Using our own knowledge, we ground the large language model and are able to mitigate hallucinations, bias and also provide validation through provididng exact source content where information is drawn from

04

Performance & Latency

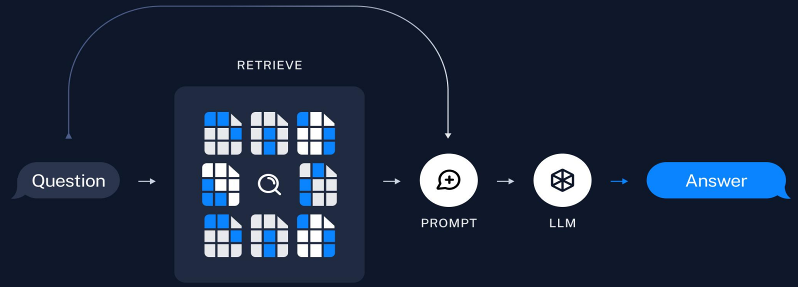
No need to train / fine-tune the large language model, we just use the RAG paradigm on any knowledge base enabling production-grade Gen AI based applications which don’t require heavy costs and compute and also get responses faster

RAG Flow



❑ **Indexing:** Pipeline for ingesting data from a source and indexing it. This usually happens offline.

❑ **Retrieval and Generation:** the actual RAG chain, which takes the user query at run time and retrieves the relevant data from the index, then passes that to the model.



➢ https://python.langchain.com/docs/use_cases/question_answering/