LLM OLYMPICS 2024

This blog is an attempt to mention most of the major developments in the LLM race, *as the world is calling it,* the center of the new tech revolution. You will soon find out that it's not just a running race but rather a leaping race forward. Buckle up and let us begin.

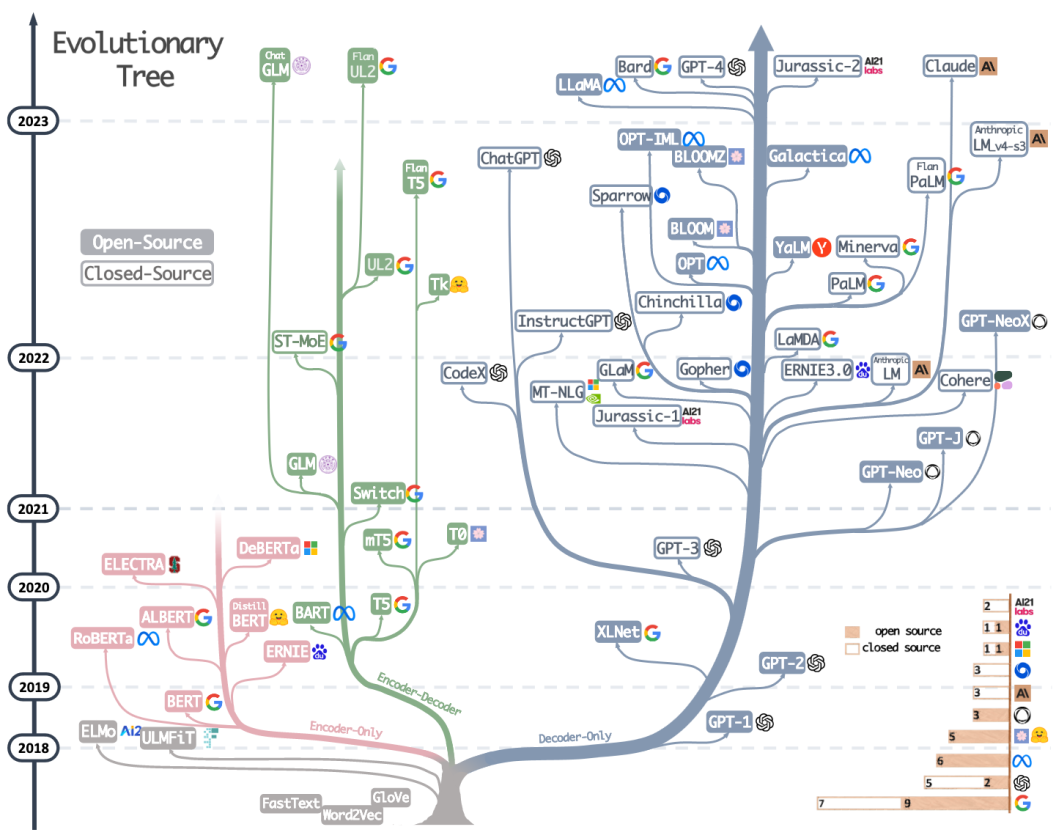
*Please feel free to correct me in any way, I am also learning and writing as I go. Looking forward to your constructive feedback.*

*Figure 1: (Left) OpenAI’s Sora generating video of a woman walking down the streets of Tokyo, (Right) Claude 3 Model family*

Starting with ChatGPT, the generative AI frenzy has triggered global wide accessibility of LLMs and its emergent capabilities. Now, not only being limited to the tech giants but rather, it has already started being incorporated to numerous consumers facing enterprise solutions/products making things ease, efficient, instant and diverse with retrieving documents from piles, generating interactive content, analyzing data, inferring business decisions and even replacing humans.

“***Compute, Not Fiat or Bitcoin, Will Be The 'Currency of the Future' and will be the most expensive commodity, Says Sam Altman as Nvidia's Jensen Huang Highlights $100 Trillion AI Opportunity***”.

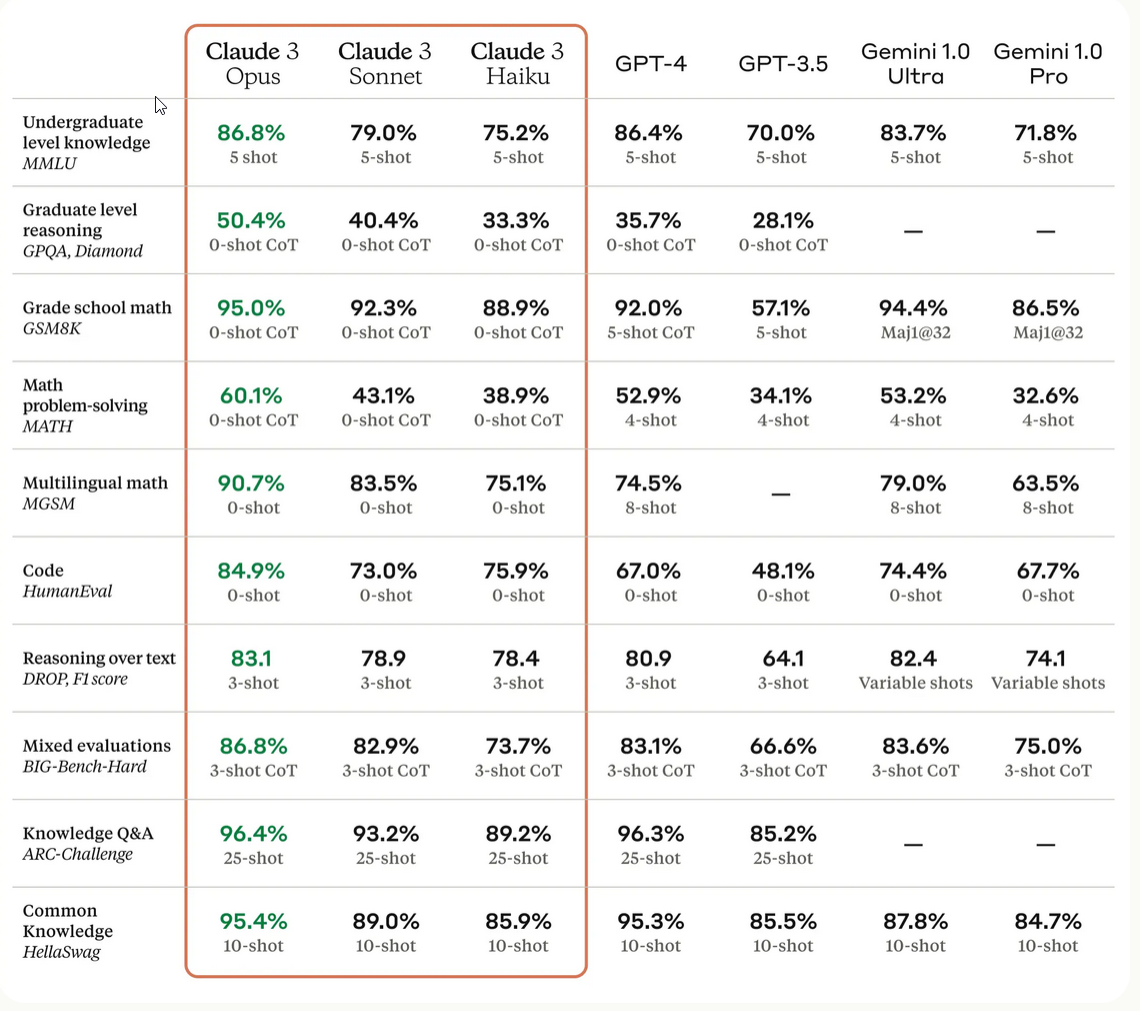
It is no longer the case that we are limited to text, as the past was, the advancement includes ALL the data you can ever imagine. With multimodality, think about images, videos, tables, graphs, audios, flowcharts, designs, networks, graphs, everything can be processed, understood, referenced, retrieved, analyzed and generated that are far more superior to what a human ever will be able to. Think about readings and understanding pages of books, generating software with thousands of lines of code, generating diverse identities with images, sketches & 3D avatars, creating real experiences, scenes, movies with videos, expressing sentiments with audios and now potentially exploring the physical worlds with robots, all is not just possible, but is already being done, and is leaping forward like never.



*Figure 2: The evolutionary tree of modern LLMs from "Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond"*

Will 2024 be the year when LLMs surprise human intelligence, break the MMLU benchmark, undergraduate level knowledge, from 86.85% with Claude-3 to all the way 100% or will 2024 be the year where humanity finally gets the wake-up call and a taste of the future to say Stop it, we have had enough? When do we say, we need a pause because it is not going to stop.

OpenAI is already planning for Artificial General Intelligence (AGI), Microsoft and Google have already started integrating all its suite's products with LLMs, Elon Musk at X is pouring billions of dollars into its own, businesses all around the world have a LLM based revenue and subscription model, we already have an AI software developer among us and so much more. The only question for me is, I have seen enough, what else will surprise me more? But just after me saying that, I will be in the coming days, very soon.



*Figure 3: C*omparison of the Claude 3 models to those of our peers on multiple benchmarks

I think that it will be enough of me contextualizing the hyperbolic advancements in LLMs, rather we in this blog we will look at the timelines of the developments of LLMs recently from our participants of the race.

**Introducing the Olympics 2024** of bigger, better and human-like LLM we have: OpenAI/Microsoft, Google (DeepMind), Anthropic, Amazon, Mistral, Cohere, Hugging face, NVIDIA, Inflection, Baidu, now X (Twitter) and sometime later Apple*.*

I will structure the announcements, facts, features, news and anything I can find in the form of a timeline as shown below, almost all are referenced below for more further information if you need.

*You can find the complete timeline* [*here : [LLM OLYMPICS 2024]*](https://excalidraw.com/#json=soCGVTwZ3fcK1KRq_7Yt1,307awQFVaSe8cRyuF2kCTg) *as well as.*

[2017] :

[X]: Transformer Model Architecture, Attention is all you need *(The Almighty Supreme)*

*“LSTMs are dead, Long Live Transformers”*

[2018]

[06] : OpenAI Generative Pre-Trained Transformers – 1.5B params

[08] : Google’s Bidirectional Encoder Representations for Transformers (BERT)

[2019]

[02] : OpenAI GPT-2 with 8.3B params, *Generates human-like texts*

[03] : Baidu Enhanced Representation through knowledge Integration (ERNIE) with 4.5B params

[06] : Microsoft invest $1Billion in OpenAI

[2020]

[05] : OpenAI GPT-3 with 175B params, *largest ever language model*

[2021]

[01] : OpenAI introduces DALL-E with 12 B params, *Generate images from text*

[05] : Google Introduces LaMDA, *intended for dialogue-based agents*  
 *Google plans to integrate LaMDA into its main search portal, its voice assistant, and Workplace, its collection of cloud-based work*  *software that includes Gmail, Docs, and Drive. But the eventual goal, said Pichai, is to create a conversational interface that allows*  *people to retrieve any kind of information—text, visual, audio—across all Google’s products just by asking*.

[06] : Github introduces Copilot, AI pair-programmer for Coding

[2022]

[01] : OpenAI’s InstructGPT, *a model better at following instructions and less likely to hallucinate*

[04] : Google introduces PaLM with 540B params

[07] : Huggingface introduces BLOOM 176B params, *beating the previous largest model*

[11] : OpenAI Introduces ChatGPT, generative AI chatbot

[2023]

[01] : Microsoft invests $10 billions more on OpenAI (rumoured)

[02] : Google introduces BARD, Meta Introduces LLaMA, Microsoft includes AI in its search engine, Bing

[03] : OpenAI introduces APIs for developers, launches GPT-4, *likely a trillion params multimodal model*

Amazon introduces its LLM, BedRock, NVIDIA showcases NEMO

*-------------------------------------- Quantization Techniques -------------------------------------------*

[05] : Falcon tops OpenLLM Leaderboard with 40B params

[07] : Anthropic Claude 2, *200K token context length*

Meta releases LLaMA 2 with 7, 13, 70B params, *includes commercial license and 40% better performance*

[12] : Mixtral introduces 8x7B, *introduces mixture of models (MoEs) in its architecture*

Microsoft introduces Phi-2 with 2.7B params, *surprises with the power small LLMs, beats Mixtral, LLaMA2*

[2024]

[02] : OpenAI introduces Sora, *text to video model capable of generating realistic and imaginative scenes*

Google introduces Gemini 1.5 (pro) with MoE architecture, *context size of up to 10M tokens (GPT4 has 128K),*

*can incorporate 1 hour of video, 11 hours of audio, >30K lines of code, >700K words*

Microsoft publishes the paper “The Era of 1-bit LLMs”, *representing weight parameters as either {-1, 0,1} not floating-point numbers (16 bits)*

OpenAI collaborates with FigureAI, *to develop humanoid robots with $675B params in Series B funding*

[03]: Cognition Labs introduces world’s first fully autonomous AI Software Developer called Devin, *capable of*   
 *developing an entire software itself, and setting a new state of the art on the SWE-bench coding benchmark*

Anthropic introduces Claude-3 (Opus (most intelligent), Sonnet and Haiku) with 175B params, *beating all*  *previous models in almost all evaluation metrics of text*

*XAI (previously twitter AI)* open sources Grok-1, 314 B params with MoE architecture*, doesn’t beat all the*  *models but Elon Musk is just getting started.*  
 *A unique and fundamental advantage of Grok is that it has real-time knowledge of the world via the X platform.*

[...] : OpenAI is rumored to be not training GPT-5 and *expanding GPT-4 by connecting it to Internet and moreeee...*  
 *Sam Altman says that this technology affects everyone, so it is important to involve everyone in discussions.*

Google and Apple might collaborate to add Gemini to iPhone, Samsung Galaxy and Pixel phones

The DeFacto understanding or the general belief in the world of LLMs is that the larger the model size, the bigger the model means the better the model. So, it can be well predicted that the athletes in this race will continue beefing up their flagship models with even more parameters, instructions and use cases.

What does it mean for humanity to already start developing models better than GPT-4 or even GPT-5 when we still don’t understand how GPT-4 works and what it does? How will the world evolve and how will the job market shape itself, when LLMs are connected to the internet and proxy the users on their behalf, who will we be talking to? What skills will be relevant and what won’t?

LLM generating text and amusing images, might be only a tiny bit of what's really important, what if LLMs are able to teach themselves (self-learning capabilities), learn to differentiate better from good, and apply betterness in itself? Or AI is still very expensive in resources and won’t change much? We are just hallucinating.

With all these, let us not talk about ethical and privacy issues, we will be here all day.

Nevertheless, we are all watching with our bated breath.

Thank you for reading till here, Much Appreciated.

# References

1. Announcing Grok, <https://x.ai/blog/grok>
2. Better language models and their implications, <https://openai.com/research/better-language-models>
3. Language models are few-shot learners, <https://openai.com/research/language-models-are-few-shot-learners>
4. Evaluating large language models trained on code, <https://openai.com/research/evaluating-large-language-models-trained-on-code>
5. WebGPT: Improving the factual accuracy of language models through web browsing, <https://openai.com/research/webgpt>
6. Mixtral of experts, <https://mistral.ai/news/mixtral-of-experts/>
7. 2024 Outlook for Language Models by Ubaid Dhiyan, <https://www.linkedin.com/pulse/2024-outlook-language-models-ubaid-dhiyan-t2uoc/>
8. The Claude 3 Model Family: Opus, Sonnet, Haiku,   
   <https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf>
9. Video generation models as world simulators, <https://openai.com/research/video-generation-models-as-world-simulators>
10. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, <https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf>
11. The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits, <https://arxiv.org/abs/2402.17764>
12. Introduction to Quantization cooked in 🤗 with 💗🧑‍🍳, <https://huggingface.co/blog/merve/quantization>
13. This week, [@xAI](https://twitter.com/xai)will open source Grok, <https://twitter.com/elonmusk/status/1767108624038449405>
14. Overview of LLM Quantization Techniques & Where to Learn Each of Them?, <https://yousefhosni.medium.com/overview-of-llm-quantization-techniques-where-to-learn-each-of-them-0d8599acfec8>
15. Our next-generation model: Gemini 1.5, <https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/>
16. [bistandbytes 4-bit quantization blogpost](https://huggingface.co/blog/4bit-transformers-bitsandbytes) - This blogpost introduces 4-bit quantization and QLoRa, an efficient finetuning approach.
17. [Merve's blogpost on quantization](https://huggingface.co/blog/merve/quantization) - This blogpost provides a gentle introduction to quantization and the quantization methods supported natively in transformers.
18. OpenAI's GPT-5 release could be as early as this summer, <https://sea.mashable.com/tech/31750/openais-gpt-5-release-could-be-as-early-as-this-summer>
19. The History of Large Language Models, <https://synthedia.substack.com/p/the-history-of-large-language-models>
20. Introducing Devin, the first AI software engineer, <https://www.cognition-labs.com/introducing-devin>
21. A Conversation with the Founder of NVIDIA: Who Will Shape the Future of AI? <https://www.youtube.com/watch?v=8Pm2xEViNIo>
22. The Race is On: Google and Microsoft Compete Over Large Language Models with ChatGPT, <https://medium.com/@suruchi.hr/the-race-is-on-google-and-microsoft-compete-over-large-language-models-with-chatgpt-5a63165f24c8>
23. Apple is reportedly exploring a partnership with Google for Gemini-powered feature on iPhones, <https://techcrunch.com/2024/03/17/apple-is-reportedly-exploring-a-partnership-with-google-for-gemini-powered-feature-on-iphones/>
24. Outrageously Large Neural Networks: The Sparsely Gated Mixture-of-Experts Layer, <https://arxiv.org/abs/1701.06538>