Is Llama 3 or GPT-4 better?

Llama 3 and GPT-4 are two of the most advanced large language models (LLMs) available to the public.

Let's see which LLM is better by comparing both models in terms of multimodality, context length, performance, and cost.

Multimodal

The release of GPT-4o finally brought initial information showing that GPT-4 has multimodal capabilities. You can now access these multimodal features by interacting with ChatGPT using the GPT-4o model. As of June 2024, GPT-4o does not have any built-in way to generate video and audio. However, it is capable of generating text and images based on video and audio inputs.

Llama 3 is also planning to offer an intermodal model for the upcoming Llama 3 400B. It will most likely integrate similar technologies with CLIP (Contrast Language-Imager Pre-Training) to generate images using Zero-shot Learning techniques. But since Llama 400B is still in training, the only way for the 8B and 70B models to generate images is to use extensions like LLaVa, Visual-LLaMA, and LLaMA-VID. As of now, Llama 3 is purely a language-based model that can take text, images, and audio as input to generate text.

Context length

Context length refers to the amount of text a model can process at once. This is an important factor when considering LLM capabilities because it determines the amount of context in which the model can operate when interacting with users. In general, higher context length makes LLM better because it provides a higher degree of coherence and continuity and can reduce the repetition of errors during the interaction.

Model	Description of training data	Parameters	Context length	GQA	Number of tokens	Limited knowledge
Llama 3	Incorporates publicly available online data	8B	8k	Have	15T+	March, 2023
Llama 3	Incorporates publicly available online data	70B	8k	Have	15T+	December 2023

Llama 3 models have an effective context length of 8,000 tokens (about 6,400 words). This means that the Llama 3 model will have a contextual memory of about 6,400 words during the interaction. Any words that exceed the 8,000 token limit will be forgotten and will not provide any additional context during the interaction.

Model	Describe	Context window	Training data
GPT-4o	Multi-modal model, cheaper and faster than GPT-4 Turbo	128,000 tokens (API)	Up to Oct 2023
GPT-4-Turbo	Model GPT-4 Turbo is streamlined with visibility.	128,000 tokens (API)	Up to Dec 2023
GPT-4	The first GPT-4 model	8,192 tokens	Up to Sep 2021

In contrast, GPT-4 now supports significantly larger context lengths of 32,000 tokens (about 25,600 words) for ChatGPT users and 128,000 tokens (about 102,400 words) for those using API endpoints. This gives the GPT-4 model advantages in managing extended conversations and the ability to read long documents or even read entire books.

Efficiency

Compare performance by looking at the Llama 3 benchmark report April 18, 2024 from Meta AI and GPT-4 May 14, 2024, OpenAI's GitHub report. Here are the results:

Model	MMLU	GPQA	MATH	HumanEval	DROP
GPT-4o	88.7	53.6	76.6	90.2	83.4
GPT-4 Turbo	86.5	49.1	72.2	87.6	85.4
Llama3 8B	68.4	34.2	30.0	62.2	58.4
Llama3 70B	82.0	39.5	50.4	81.7	79.7
Llama3 400B	86.1	48.0	57.8	84.1	83.5

Here's what each criterion evaluates:

1. **MMLU (Massive Multitask Language Understanding)** : Assesses the model's ability to understand and answer questions on a variety of academic topics.
2. **GPTQA (General Purpose Question Answering)** : Evaluates the model's skill in answering real-life questions in the open domain
3. **MATH** : Test the model's ability to solve problems.
4. **HumanEval** : Measures the model's ability to generate correct code based on a given human programming prompt.
5. **DROP (Discrete Reasoning Over Paragraphs)** : Evaluates the model's ability to perform discrete reasoning and answer questions based on text passages.

Recent benchmarks highlight the performance differences between the GPT-4 and Llama 3 models. Although the Llama 3 8B model appears to lag significantly behind, the 70B and 400B models show lower results but similar to both the GPT-4o and GPT-4 Turbo models in terms of academic and general knowledge, reading and comprehension, reasoning and logic, and coding. However, no Llama 3 model has yet achieved the performance of GPT-4 in purely mathematical terms.

Price

Cost is an important factor for many users. OpenAI's GPT-4o model is available for free to all ChatGPT users with a limit of 16 messages every 3 hours. If you need more, you'll have to subscribe to ChatGPT Plus for $20/month to expand the GPT-4o's message limit to 80, while also gaining access to other GPT-4 models.

On the other hand, both the Llama 3 8B and 70B models are free and open source, which can be a significant advantage for developers and researchers looking for a cost-effective solution without compromising to performance.

Accessibility

GPT-4 models are widely accessible through OpenAI's Generative AI ChatGPT chatbot and through its API. You can also use GPT-4 on Microsoft Copilot, which is a way to use GPT-4 for free. This wide availability ensures that users can easily take advantage of its capabilities in various use cases. In contrast, Llama 3 is an open source project that provides model flexibility and encourages broader experimentation and collaboration within the AI ??community. This open access approach could democratize AI technology, making it available to a wider audience.

While both models are available, GPT-4 is much easier to use because it's integrated into popular productivity tools and services. On the other hand, Llama 3 is mainly integrated into research and business platforms such as Amazon Bedrock, Ollama, and DataBricks (except for Meta AI chat support), which fails to attract the larger market of not technically savvy.

Is GPT-4 or Llama 3 better?

So which LLM is better? GPT-4 is a better LLM. GPT-4 excels in multimodality with advanced capabilities in handling text, image and audio input, while similar features of Llama 3 are still in development. GPT-4 also provides much larger context lengths and better performance, and is widely accessible through popular tools and

services, making GPT-4 more user-friendly .

However, it is important to emphasize that the Llama 3 models have performed very well for a free and open source project. As a result, Llama 3 remains an outstanding LLM, popular with researchers and businesses for its free and open source nature, while also offering impressive performance, flexibility, and notable security features. trust. While general consumers may not find an immediate use for Llama 3, it remains the most viable option for many researchers and businesses.

In summary, while GPT-4 stands out for its advanced multimodal capabilities, greater context length, and seamless integration into widely used tools, Llama 3 offers a viable alternative. value with its open source nature, allowing for more customization and cost savings. So, in terms of applications, GPT-4 is ideal for those looking for ease of use and comprehensive features in one model, while Llama 3 is well suited for developers and researchers. are looking for flexibility and adaptability.