Llama 2 Vs ChatGPT 4

Llama 2 has fewer parameters than ChatGPT, still it can run on a single gpu making It an accessible choice for various applications.

Llama 2 supports 20 languages while ChatGPT supports 26 languages

Llama 2 is free for use, and it contains 40 % more data than llama 1, it is a free alternative of ChatGPT 4, while ChatGPT on the other hand costs 20$ per month

Llama 2 is trained on lesser tokens then chatgpt4

According to Meta's research paper, it acknowledges that it is less powerful than ChatGPT 4.

Benchmark (shots)	GPT-3.5	GPT-4	PaLM	PaLM-2-L	LLAMA 2
MMLU (5-shot)	70.0	**86.4**	69.3	78.3	68.9
TriviaQA (1-shot)	–	–	81.4	**86.1**	85.0
Natural Questions (1-shot)	–	–	29.3	**37.5**	33.0
GSM8K (8-shot)	57.1	**92.0**	56.5	80.7	56.8
HumanEval (0-shot)	48.1	**67.0**	26.2	–	29.9
BIG-Bench Hard (3-shot)	–	–	52.3	**65.7**	51.2

Table 4: Comparison to closed-source models on academic benchmarks. Results for GPT-3.5 and GPT-4 are from OpenAI (2023). Results for the PaLM model are from Chowdhery et al. (2022). Results for the PaLM-2-L are from Anil et al. (2023).

For coding tasks ChatGPT 4 with code interpreter and specialized models like starcoder should be ahead of llama 2 according to benchmarks.

Power and Performance: META acknowledges that LLAMA 2 is less powerful than GPT-4 and PaLM 2. It falls slightly behind in performance benchmarks compared to its rivals.

Training Data: LLAMA 2 was trained on fewer "tokens" (text used for training) compared to its competitors. It trained on two million tokens, while Google's PaLM 2 trained on 3.6 million tokens.

Language Support: LLAMA 2 supports fewer languages than PaLM 2 and GPT-4. It covers 20 languages, whereas PaLM 2 supports 100 and GPT-4 supports 26. Google's Bard, which utilizes PaLM 2, even supports nine Indian languages.

However, Llama-2 is weak in coding.

It is not better than GPT-3.5 (**48.1**) level or GPT-4 (**67**) when it comes to coding. Although it MMLU (Massive Multitask Language Understanding) benchmark is good, HumanEval shows coding capability is quite a bit lower compared to StarCoder (33.6) or many other models specifically designed for coding.

> When it comes to writing, Llama-2 and GPT-4 are very different, too.
>
> When asked to write a poem, both had a different approach. ChatGPT seems to have more intentional word choices which are more focused on the way words sound, a more sophisticated poet with a wider vocabulary. While Llama-2 uses a more obvious rhyming word selection, like a high school poem.

Llama 2 is open source, whereas GPT 4 is not open source but can be accessed through APIs.

OpenAI describes GPT-4 as 10 times more advanced than its predecessor, GPT-3.5. This enhancement enables the model to better understand the context and distinguish nuances, resulting in more accurate and coherent responses. Also, llama 2 ‘s performance is equivalent to GPT3.5, thus making GPT 4 better than llama 2.

The LLM offers three tiers of parameters (factors that AI systems can learn from training data) reviewed by human evaluators:

- 7 billion parameters
- 13 billion parameters
- 70 billion parameters

While this falls short of GPT 3.5's 175 billion parameters, when it comes to Massive Multitask Language Understanding (MMLU), a scoring system used to assess the problem-solving capabilities of language models, the gap is much narrower.

For instance, Llama 2 has an MMLU score of 68.9, which is just behind GPT 3.5's 70.0. Although this is a long way off from GPT4's 86.4 rating, it is close enough to position Llama 2 as a viable open-source competitor to GPT 3.5.

Although Llama 2 isn't in a position to unseat GPT4, so far, it has demonstrated that it can be competitive against GPT 3.5 in certain areas.

1. Model Architecture:

- Llama 2: Llama 2 is an auto-regressive language model that uses an optimized transformer architecture.
- ChatGPT-4: ChatGPT-4 is based on eight models with 220 billion parameters each, connected by a Mixture of Experts (MoE).

2. Parameter Sizes:

- Llama 2: Llama 2 comes in a range of parameter sizes, including 7 billion, 13 billion, and 70 billion.
- ChatGPT-4: ChatGPT-4 boasts a significantly larger parameter size with approximately 1.76 trillion parameters (eight models with 220 billion parameters each).

3. Language Support:

- Llama 2: Llama 2 is intended for use in English.
- ChatGPT-4: The provided information doesn't specify the language support for ChatGPT-4.

4. Availability and Accessibility:

- Llama 2: Llama 2 is open-source and freely available for commercial and research use, making it accessible to startups, established businesses, and lone operators without cost.
- ChatGPT-4: The information provided states that ChatGPT-4 is a paid system, suggesting that it may have a commercial licensing model.

In addition to open-source models, we also compare LLAMA 2 70B results to closed-source models. As shown in Table 4, LLAMA 2 70B is close to GPT-3.5 (OpenAI, 2023) on MMLU and GSM8K, but there is a significant gap on coding benchmarks. LLAMA 2 70B results are on par or better than PaLM (540B) (Chowdhery et al., 2022) on almost all benchmarks. There is still a large gap in performance between LLAMA 2 70B and GPT-4 and PaLM-2-L.

We also analysed the potential data contamination and share the details in Section A.6.

Benchmark (shots)	GPT-3.5	GPT-4	PaLM	PaLM-2-L	LLAMA 2
MMLU (5-shot)	70.0	**86.4**	69.3	78.3	68.9
TriviaQA (1-shot)	–	–	81.4	**86.1**	85.0
Natural Questions (1-shot)	–	–	29.3	**37.5**	33.0
GSM8K (8-shot)	57.1	**92.0**	56.5	80.7	56.8
HumanEval (0-shot)	48.1	**67.0**	26.2	–	29.9
BIG-Bench Hard (3-shot)	–	–	52.3	**65.7**	51.2

Table 4: Comparison to closed-source models on academic benchmarks. Results for GPT-3.5 and GPT-4 are from OpenAI (2023). Results for the PaLM model are from Chowdhery et al. (2022). Results for the PaLM-2-L are from Anil et al. (2023).