Comparison of Core Abilities of Four Large Language Models: GPT-3, GPT-4, LLaMA 2, and PaLM 2

Guoyu Xu

{ CST2309177@xmu.edu.my }

School of Computing Science and technology, Xiamen University Malaysia, Sepang, Malaysia

Abstract. Different large language models vary dramatically in their resource requirements, with some costing millions to train while others can run on a laptop. This paper examines four models that represent different design philosophies: GPT-3, GPT-4, LLaMA 2, and PaLM 2. GPT-3, with its 175 billion parameters, first showed that scaling up could unlock few-shot learning. GPT-4 went further by adding image understanding, though OpenAI never published the architecture. Meta took the opposite approach with LLaMA 2, releasing weights publicly so anyone could experiment. Google's PaLM 2 focused on languages beyond English, training on over 100 languages. After reviewing how each model works and where it performs best, the paper discusses what problems remain unsolved, including that models still make up facts, training still costs too much, and no one fully understands why certain prompts fail.

Keywords: Large Language Models, Transformer, Natural Language Processing, Model Evaluation

1 Introduction

Three years ago, few people outside machine learning research had heard of GPT. Now these models write emails, debug code, and even pass law exams. The shift happened fast. In 2020, GPT-3 surprised researchers by answering questions it was never trained to answer, just from seeing a few examples in the prompt. By 2023, GPT-4 could look at a photo and explain what was funny about it. Meanwhile, Meta released LLaMA 2 for free, and suddenly university labs could run experiments that previously required corporate budgets. Google countered with PaLM 2, which handles Hindi and Swahili nearly as well as English.

There are two things that changed and made this possible. Firstly, hardware got faster. Training GPT-3 required mass-scale coordination across thousands of GPUs, which was expensive, but no longer impossible. Secondly, the Transformer architecture turned out to scale better than anyone expected [1]. Earlier neural networks struggled with long documents because information had to pass through many steps. Transformers use attention mechanisms that let any word look directly at any other word, no matter how far apart. When Vaswani et al.

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365089

published this idea in 2017, they tested it on translation. Nobody guessed it would eventually power systems that write poetry and solve math problems.

The models keep finding new uses. Customer service bots handle complaints without human agents. GitHub Copilot suggests code as programmers type. Students use ChatGPT for homework help, sometimes too much help, which worries teachers. Lawyers draft contracts faster. Doctors get second opinions on diagnoses. Not all applications work well yet. The models confidently state wrong facts, a problem researchers call hallucination. They can be tricked into saying harmful things. Running them costs real money, both in electricity and in cloud computing fees. These limitations matter as much as the capabilities.

Researchers have studied these models from many angles. Brown et al. built GPT-3 and documented how performance improves with scale [2]. The GPT-4 technical report showed gains on standardized tests but left out most architectural details [3]. Touvron et al. at Meta argued that open models benefit science and released LLaMA 2 weights for download [4]. Anil et al. described how PaLM 2 handles multilingual tasks [5]. Liang et al. pointed out that different papers use different benchmarks, making comparison hard, and proposed HELM as a unified framework [6]. Each paper tells part of the story, but none compares all four models side by side.

That comparison is what this paper attempts. GPT-3 and GPT-4 come from the same company, so tracking their differences shows how the field moved in three years. LLaMA 2 shows what happens when weights become public, as researchers fine-tune it for medicine, law, and dozens of other domains. PaLM 2 shows a different priority, which is linguistic diversity over English dominance. Section 2 explains how each model works under the hood. Section 3 compares benchmark scores. Section 4 covers problems that remain open. Section 5 wraps up with conclusions.

2 Technical Foundations and Model Architectures

All four models trace back to the same 2017 paper on Transformers [1], but they diverge in size, training data, and what their creators chose to reveal. Before diving into specifics, it helps to understand what these models share. All four use the decoder-only variant of the Transformer, meaning they generate text one token at a time, left to right. All four learn by predicting the next word in massive text corpora. The differences lie in scale, training recipes, and business decisions about openness. Some labs publish everything; others guard their secrets. Table 1 gives the quick version before the details.

Table 1. Overview of the four models

Model	Developer	Release	Parameters	Architecture	Open	Key Focus
GPT-3	OpenAI	2020.06	175B	Decoder-only, 96 layers	No	Few-shot learning
GPT-4	OpenAI	2023.03	Undisclosed	Multimodal, RLHF	No	Reasoning, Safety
LLaMA 2	Meta	2023.07	7B-70B	RMSNorm, SwiGLU, RoPE	Yes	Openness, Efficiency
PaLM 2	Google	2023.05	Undisclosed	Compute-optimal	No	Multilingual, Coding

2.1 GPT-3

OpenAI trained GPT-3 on about 570 gigabytes of text including books, Wikipedia, web pages filtered from Common Crawl [2]. The model itself has 175 billion parameters spread across 96 layers, with 96 attention heads per layer and a hidden dimension of 12,288. For comparison, GPT-2 had 1.5 billion parameters [7]. The jump was over 100x. Training required distributing computation across thousands of GPUs, a coordination challenge that only a few organizations could manage at the time.

The training objective sounds simple, which is to predict the next word. Given "The cat sat on the", the model learns to predict "mat" or "floor" or whatever comes next in the training data. After doing this billions of times, the model picks up grammar, facts, reasoning patterns, and even some arithmetic. The surprise with GPT-3 was few-shot learning. When given three examples of English-to-French translation in the prompt, the model translates a fourth sentence without any fine-tuning. Nobody designed this feature, as it emerged from scale.

GPT-3 has real weaknesses. It makes up facts that sound plausible but are wrong. Small changes to how a question **is** phrased can flip the answer. The model has no way to say, "I don't know", so it always produces something. Moreover, running it costs money. API calls add up. These problems did not go away in later models, though some got better.

2.2 GPT-4

OpenAI published the GPT-4 technical report in March 2023, but left out most details [3]. The report included no parameter count and no architecture diagram. The company cited competition and safety as reasons. What they did share is that GPT-4 scores in the 90th percentile on the bar exam, up from the 10th percentile for GPT-3.5. It handles images, not just text. Users can upload a photo of their fridge and ask what to cook for dinner.

One known ingredient is RLHF, which stands for reinforcement learning from human feedback [8]. The process works as follows. First, generate several responses to a prompt, have humans rank them from best to worst, train a reward model on those rankings, then fine-tune the language model to maximize the reward. OpenAI also ran red-teaming exercises where external experts tried to make the model produce harmful content. Findings went back into training.

Compared to GPT-3, GPT-4 handles longer inputs and makes fewer reasoning errors on multi-step problems. Vision opens new applications: reading handwritten notes, interpreting charts, describing photos for blind users. The downside is opacity. Outside researchers cannot check the weights, run ablations, or verify claims. When the model fails, no one outside OpenAI can diagnose why. This secrecy sparked debate. Some researchers argue that safety requires caution because publishing weights lets bad actors fine-tune models for harm [16, 17]. Others counter that science depends on reproducibility, and closed models cannot be properly audited [18]. The tension has no easy resolution. OpenAI chose commercial interests and safety concerns over openness; whether that trade-off benefits society remains contested.

2.3 LLaMA 2

Meta released LLaMA 2 in July 2023 with a license that allows most research and commercial use [4]. Anyone can download the weights. This matters. With GPT-4, users send queries to an API and get responses back. With LLaMA 2, researchers run the model on their own hardware.

They can look inside, change things, and fine-tune on their own data. Three sizes exist: 7 billion, 13 billion, and 70 billion parameters. The smallest runs on a gaming GPU.

The architecture tweaks the standard Transformer recipe. RMSNorm replaces LayerNorm for faster training. SwiGLU activation replaces ReLU. Rotary position embeddings (RoPE) encode where each token sits in the sequence. Training used 2 trillion tokens with a context window of 4,096 tokens, which is twice the original LLaMA. Meta also released chat-tuned versions that went through supervised fine-tuning and RLHF.

The open release sparked fast iteration. Within weeks, people posted quantized versions that fit in less memory, instruction-tuned variants, and specialized versions for medical and legal text. On hard reasoning tasks, LLaMA 2 still trails GPT-4. Its multilingual coverage is narrower than PaLM 2. But for researchers who need to understand what the model is doing, open weights change everything.

2.4 PaLM 2

Google announced PaLM 2 in May 2023 [5]. The headline feature is that training data spans over 100 languages, with effort to include lower-resource languages that most datasets underrepresent. This shows in benchmarks. On translation tasks involving Tamil, Swahili, or Tagalog, PaLM 2 beats models trained mostly on English and Chinese.

The training philosophy follows the Chinchilla paper [9]. That study argued that most language models are too big for how much data they see, and they would perform better if researchers used smaller models trained on more tokens. PaLM 2 takes this advice. Google did not publish the parameter count, but the model reportedly matches larger predecessors while costing less to run.

PaLM 2 powers Bard, Google's chatbot. Specialized versions exist for medicine (Med-PaLM 2) and security (Sec-PaLM). Like GPT-4, the weights stay closed. Researchers cannot download them or verify claims independently. The choice makes business sense for Google but frustrates scientists who want reproducibility.

3 Performance Analysis

Comparing language models is messy because each lab picks benchmarks that favor its model, uses different prompting tricks, and sometimes tests on different versions without clear labels. The HELM project tried to fix this by testing many models on the same tasks with the same prompts [6]. Under HELM, GPT-4 usually wins. PaLM 2 and LLaMA 2-70B trade second place depending on the task. GPT-3 falls noticeably behind.

MMLU tests knowledge across 57 subjects including history, physics, law, and medicine [10]. On this benchmark, GPT-4 leads while LLaMA 2-70B comes closer than its parameter count would suggest, which hints that data quality and training efficiency matter as much as size. On reasoning tasks such as GSM8K for grade school math and MATH for competition problems, GPT-4 pulls ahead by a wider margin. GPT-3 tends to lose track of intermediate steps, and LLaMA 2 handles easy problems but stumbles on longer chains. PaLM 2 performs well on math but shines brightest on multilingual tasks.

For coding, HumanEval asks models to write Python functions from docstrings [11]. GPT-4 and

PaLM 2 post the best pass rates. GPT-3 often writes code that runs but gives wrong answers. Code Llama, a LLaMA 2 variant tuned on code, closes some of the gap. The overall pattern is that GPT-4 leads on most metrics but stays closed. LLaMA 2 offers transparency at some cost in performance. PaLM 2 wins on languages beyond English. One pattern stands out across all benchmarks: gains come from both scale and data quality. GPT-4 likely has more parameters than LLaMA 2-70B, but LLaMA 2 closes much of the gap through careful data curation. PaLM 2 achieves strong results with reportedly fewer parameters by following compute-optimal scaling laws. Raw size matters less than the research community once assumed. Training longer on cleaner data often beats training briefly on a larger model.

4 Challenges and Future Directions

4.1 Current Challenges

Hallucination tops the list of unsolved problems [12]. Ask a language model about a made-up court case, and it may cite fake precedents with confident legal language. Users without expertise cannot tell the difference. Medical chatbots giving wrong diagnoses could hurt people. Retrieval-augmented generation helps because the model looks up facts in a database before answering, but it adds complexity and latency.

Cost blocks many researchers. Training GPT-3 required mass-scale GPU clusters running for weeks [13]. Even inference adds up when millions of users send queries. Environmental concerns follow: all that computation burns electricity. Only a handful of companies can afford frontier research, which concentrates power in ways that worry some observers.

Alignment remains tricky [8]. RLHF helps models refuse harmful requests, but clever prompts still break through. Different user groups want different things, and what counts as helpful varies by culture, age, and context. Interpretability lags behind capability [14]. When a model fails, engineers often cannot explain why. That makes deployment risky in regulated industries like healthcare and finance.

4.2 Future Directions

Retrieval-augmented generation keeps gaining ground [15]. The idea is to pair a language model with a search system that fetches relevant documents before generating answers. This reduces hallucination and lets models access information newer than their training data. Efficiency research also progresses. Mixture-of-experts routes each query through only part of the network. Quantization shrinks models to run on smaller hardware. LLaMA 2 showed that careful design beats brute-force scaling in some cases.

Multimodal expansion continues. GPT-4 reads images; future models may handle audio, video, and sensor data. Standardized benchmarks that test safety and bias, not just accuracy, would help the field track real progress instead of cherry-picked results. The open-source movement may reshape the landscape. LLaMA 2 proved that public weights attract contributors who improve models faster than any single company could. Mistral, Falcon, and other open models followed. If this trend continues, closed models may lose their lead as open alternatives catch up. Companies like OpenAI will need to decide whether to compete on secrecy or join the open ecosystem. Neither path guarantees success.

5 Conclusion

The four models examined in this paper represent four different bets on what matters most in language model development. GPT-3 bet that scale would unlock new abilities, and it won. GPT-4 bet that multimodal input and RLHF would push performance higher, and it did, though at the cost of transparency. LLaMA 2 bet that open weights would accelerate research, and community uptake proved them right. PaLM 2 bet on linguistic diversity, and it now leads on non-English tasks.

Each bet reflects a different theory about what matters most. OpenAI believed scale and alignment would win users. Meta believed openness would win researchers. Google believed multilingual coverage would win global markets. Three years from now, the rankings may shift, but these underlying philosophies will likely persist in whatever models come next.

Several problems remain unsolved, as models still hallucinate, training still costs millions. Alignment still breaks under adversarial prompts. Interpretability still lags. These gaps matter more as language models move into medicine, law, and education. The tension between commercial secrecy and scientific openness will not resolve soon. Neither will the tension between pushing capabilities and ensuring safety.

This comparison shows that no single model wins on every axis. Which model fits best depends on what the task demands. GPT-4 suits applications where accuracy matters most and API costs are acceptable. LLaMA 2 makes more sense when researchers need to look inside the model or fine-tune it themselves. PaLM 2 stands out for languages like Tamil or Swahili that other models handle poorly. New models appear every few months, and today's rankings will not last. Still, open versus closed, efficient versus large, English-first versus multilingual, which are the core tensions this paper highlights, are unlikely to disappear anytime soon.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. Advances in Neural Information Processing Systems. vol. 30, pp. 5998-6008 (2017)

[2] Brown, T., Mann, B., Ryder, N., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems. vol. 33, pp. 1877-1901 (2020)

[3] OpenAI.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

[4] Touvron, H., Martin, L., Stone, K., et al.: LLaMA 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)

[5] Anil, R., Dai, A., Firat, O., et al.: PaLM 2 technical report. arXiv preprint arXiv:2305.10403 (2023)

[6] Liang, P., Bommasani, R., Lee, T., et al.: Holistic evaluation of language models. arXiv preprint arXiv:2211.09110 (2022)

[7] Radford, A., Wu, J., Child, R., et al.: Language models are unsupervised multitask learners. OpenAI Technical Report (2019)

[8] Ouyang, L., Wu, J., Jiang, X., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems. vol. 35, pp. 27730-27744 (2022)

[9] Hoffmann, J., Borgeaud, S., Mensch, A., et al.: Training compute-optimal large language models. Advances in Neural Information Processing Systems. vol. 35, pp. 30016-30030 (2022)

[10] Hendrycks, D., Burns, C., Basart, S., et al.: Measuring massive multitask language understanding.

International Conference on Learning Representations (2021)
[11] Chen, M., Tworek, J., Jun, H., et al.: Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374 (2021)
[12] Ji, Z., Lee, N., Frieske, R., et al.: Survey of hallucination in natural language generation. ACM Computing Surveys. vol. 55, no. 12, pp. 1-38 (2023)
[13] Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for deep learning in NLP. Annual Meeting of the Association for Computational Linguistics. pp. 3645-3650 (2019)
[14] Zhao, W., Zhou, K., Li, J., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 (2023)
[15] Lewis, P., Perez, E., Piktus, A., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems. vol. 33, pp. 9459-9474 (2020)
[16] Seger, E., Dreksler, N., Moulange, R., et al.: Open-sourcing highly capable foundation models: an evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. arXiv preprint arXiv:2311.09227 (2023)
[17] Lermen, S., Rogers-Smith, C., Ladish, J.: LoRA fine-tuning efficiently undoes safety training in Llama 2-Chat 70B. arXiv preprint arXiv:2310.20624 (2023)
[18] Kapoor, S., Bommasani, R., Klyman, K., et al.: On the societal impact of open foundation models. arXiv preprint arXiv:2403.07918 (2024)