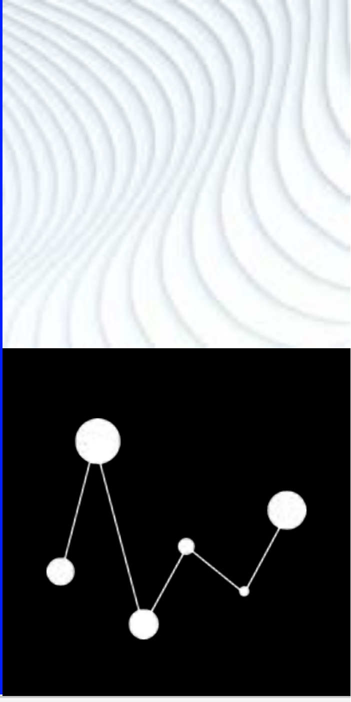
Comparing Large Language Models

How to decide which is the best state-of-the-art model



LSEG DATA & ANALYTICS

AUTHORS

Stanislav Chistyakov
Data Scientist, LSEG Analytics
Stanislav.Chistyakov@lseg.com

David Oliver
Director, Data Scientist,
LSEG Analytics
David.Oliver@lseg.com

The authors wish to thank Will Cruse for assistance in researching this publication.

LSEG Analytics has a focus on AI in the Financial Services Industry. Through our research and product development we have grown a significant body-of-knowledge that we wish to share for the benefit of the wider community.

In this series of technical blogs, we discuss some of the most useful and interesting aspects of AI from our work. In this blog, we explain how the quality of state-of-the-art models is reported, in particular the use of leaderboards. Understanding the ranking of state-of-the-art models allows us to stay on top of significant developments in the fast-changing field of AI.

The field of AI has a rapidly growing set of LLMs available...

A recent trend in AI is an increase in number of new large language models (LLMs) being released. These are being developed by both major technology companies (Google, Meta) and up-and-coming startups (Anthropic, Mistral). For example, in only the last six months, OpenAI released new models including GPT-4o, GPT-4o mini, o1-preview, o1-mini, and furthermore, updated these models and added new features [1].

At LSEG Analytics, like most any other group developing AI applications, we are faced with the dilemma of deciding which LLM to use, when to upgrade and when to decommission older LLMs. Of upmost importance in this decision is understanding the ability of state-of-the-art models to perform certain tasks; we'll refer to this ability as the "quality" of the LLM. Quality is non-trivial to measure, both due to the versatility of tasks LLMs can now achieve, as well as the difficulty in the criteria used to assess performance in these tasks. Note that in LSEG Analytics we are very aware there are other considerations beyond just quality that factor into which LLM to choose, such as cost, latency and data privacy. Given these are heavily dependent on the details of the application under consideration they won't be discussed further here.

In this technical blog we discuss one of the most popular approaches to evaluating LLMs, the **Chatbot Arena**. This leaderboard uses a crowdsourced approach; collecting a large amount of feedback in asking users to perform a pairwise comparison of two responses generated by two different LLMs.

Analysis of this leaderboard also allows for insights into the current state of AI. Maybe of most interest is that even though LLM quality is continually improving there is **evidence of diminishing differentiation in quality between new state-of-the-art LLMs**. This indicates the choice of LLM is becoming less dependent on overall quality; it may be that any of the top state-of-the-art models will be performant enough for a given task. Furthermore, for the top state-of-the-art models, **Open LLMs - although not yet surpassing Proprietary LLMs in the leaderboards - are demonstrating notable quality.** A consequence of this is Proprietary models may need offer more than just higher quality to remain the most attractive choice of LLM.

***Proprietary versus Open LLMs**: For the purposes of this publication, we classify models as **Proprietary** if access to them is fully controlled by companies that produced them, and which are typically available via an API or an application. **Open** LLMs are any models that are not "Proprietary" and can differ by their licenses and the information that is available about the data, training process, and model architecture.*

Understanding how the Chatbot Arena leaderboard is constructed provides an appreciation of what makes the latest-and-greatest model superior to previous models. The leaderboard is NOT a marketing platform to promote LLMs, meaning when the next GPT model from OpenAI becomes available, it will be quantitatively demonstratable how the new model is an improvement over the previous generations of LLMs. Ultimately – and most importantly – this understanding of the leaderboard enables better decisions as to which LLMs to choose for an AI application.

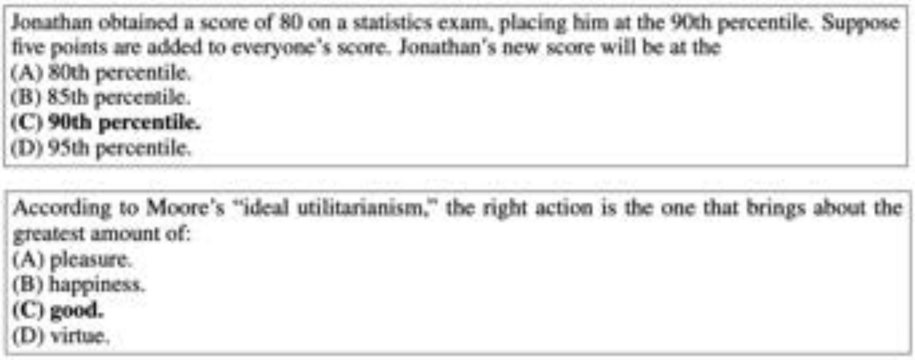
Static Benchmarks: An intuitive, but nowadays limited, method to evaluate LLMs

To provide the context of recent approaches to LLM evaluation, such as the Chatbot Arena leaderboard, it is first worthwhile providing a brief overview of ***ground-truth-based*** benchmarks [2]. These are about as straightforward as it gets; there are a clear set of expected ***labels*** to which the LLM predictions can be explicitly compared. This type of benchmarking is very popular and intuitive. The ground-truth-based benchmarks are typically ***static*** – the data does not change. Using static data ensures evaluation is the same when comparing different models or the same model over time.

Over the last decade, a comprehensive set of static benchmarks were developed to assess an LLMs ability to solve a variety of **Natural Language Processing** (NLP) tasks such as Question-and-Answering, Sentiment Analysis, and Sentence Classification. These static benchmarks preceded the current AI boom by several years.

A major static benchmark is known as **GLUE** – General Language Understanding Evaluation – released in 2018. GLUE consists of nine English language understanding tasks, designed to be sufficiently difficult and diverse to test a model's generic language understanding ability rather than its performance on a specific NLP task [3]. The performance of each model can be summarised as a single number; an arithmetic mean of scores of individual tasks. This provided a straightforward method to compare models, which contributed to GLUE's popularity; NLP researchers were able to claim that "Model A" achieved *X* on GLUE, while "Model B" achieved *Y*, with the confidence this would be readily understood by a wide audience.

However, as NLP models inevitably advanced, the scores on benchmarks such as GLUE became less informative - there was a need for more challenging tasks to assess quality. This led to the development of one of the most common benchmarks still currently in use; **MMLU** (Massive Multitask Language Understanding) [4]. MMLU combines approximately 16,000 multiple-choice questions, from elementary to professional level, across 57 academic subjects, such as mathematics, law, medicine, history, and ethics. Examples of such questions are shown on **Figure 1**. MMLU was designed to require advanced problem-solving ability and extensive world knowledge to achieve a high score.



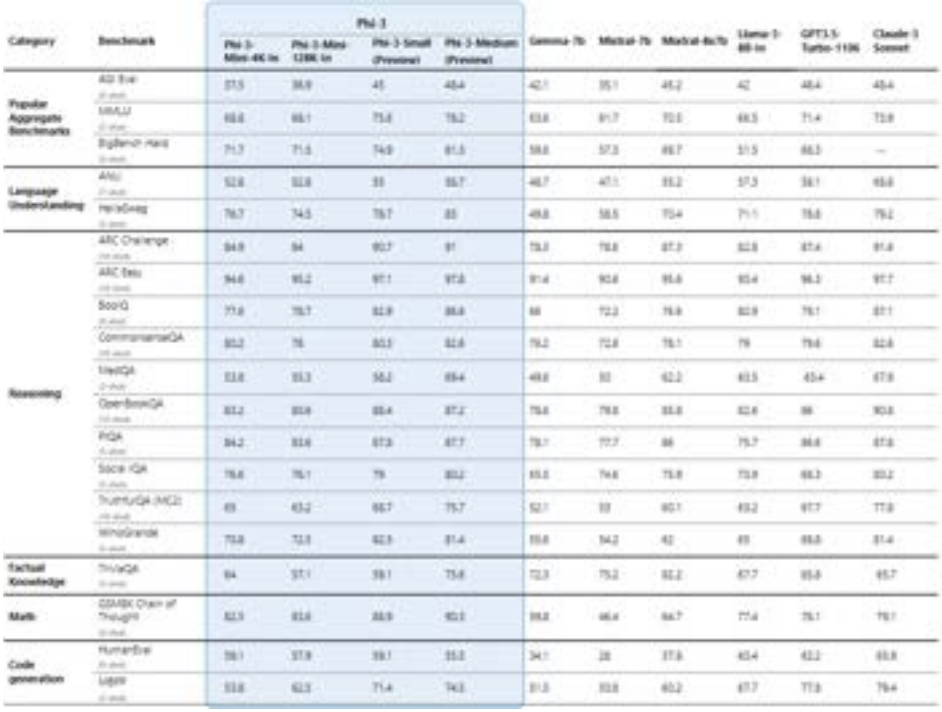
Jonathan obtained a score of 80 on a statistics exam, placing him at the 90th percentile. Suppose five points are added to everyone's score. Jonathan's new score will be at the
(A) 80th percentile.
(B) 85th percentile.
(C) 90th percentile.
(D) 95th percentile.

According to Moore's "ideal utilitarianism," the right action is the one that brings about the greatest amount of:
(A) pleasure.
(B) happiness.
(C) good.
(D) virtue.

Figure 1: Two examples from the MMLU set, which comprises of approximately 16,000 questions [4]

The pace of development of AI is observed in the significant increase in MMLU scores over the past few years. For reference, human expert performance is estimated to be at 89.8%. When MMLU was first introduced in 2020 the state-of-the-art model at the time, GPT-3 X-Large, scored 43.9%. In December 2023, Gemini 1.0 Ultra from Google DeepMind achieved 90.0%, thus becoming the first model to surpass expected human performance [5].

Nowadays, most LLMs are expected to perform a larger variety of tasks than previous generations of models. Some of these new tasks include; code understanding and code generation, complex reasoning, solving mathematical problems, and in the case of multimodal models, image, video, audio understanding. As those capabilities require their own testing, so model developers typically go beyond using a single benchmark such as MMLU and report results on multiple datasets.



Category	Benchmark	Phi-3				Gemma-7b	Mistral-7b	Mixtral-8x7b	Llama-3-8B-in	GPT3.5-Turbo-1106	Claude-3 Sonnet
		Phi-3-Mini-4K-in	Phi-3-Mini-128K-in	Phi-3-Small (Preview)	Phi-3-Medium (Preview)						
Popular Aggregate Benchmarks	AGI Eval 0-shot	37.5	36.9	45	48.4	42.1	35.1	45.2	42	48.4	48.4
	MMLU 5-shot	68.8	68.1	75.3	78.2	63.6	61.7	70.5	66.5	71.4	73.9
	BigBench Hard 0-shot	71.7	71.5	74.9	81.3	59.6	57.3	69.7	51.5	68.3	—
Language Understanding	ANLI 7-shot	52.8	52.8	55	58.7	48.7	47.1	55.2	57.3	58.1	63.6
	HellaSwag 5-shot	76.7	74.5	76.7	83	49.8	58.5	70.4	71.1	78.8	79.2
Reasoning	ARC Challenge 10-shot	84.9	84	90.7	91	78.3	78.6	87.3	82.8	87.4	91.6
	ARC Easy 10-shot	94.6	95.2	97.1	97.8	91.4	92.6	95.6	93.4	96.3	97.7
	BoolQ 0-shot	77.6	78.7	82.9	86.6	66	72.2	76.6	80.9	79.1	87.1
	CommonsenseQA 10-shot	80.2	78	80.3	82.6	76.2	72.6	78.1	79	79.6	82.6
	MedQA 2-shot	53.8	55.3	58.2	69.4	49.6	50	62.2	60.5	63.4	67.9
	OpenBookQA 10-shot	83.2	80.6	88.4	87.2	78.6	79.8	85.8	82.6	86	90.8
	PIQA 5-shot	84.2	83.6	87.3	87.7	78.1	77.7	86	75.7	86.6	87.8
	Social IQA 5-shot	76.6	76.1	79	80.2	65.5	74.6	75.9	73.9	68.3	80.2
	TruthfulQA (MC2) 10-shot	65	63.2	68.7	75.7	52.1	53	60.1	63.2	67.7	77.8
	WinoGrande 5-shot	70.8	72.5	82.5	81.4	55.6	54.2	62	65	68.8	81.4
Factual Knowledge	TriviaQA 5-shot	64	57.1	59.1	73.9	72.3	75.2	82.2	67.7	85.8	65.7
Math	GSM8K Chain of Thought 0-shot	82.5	83.6	88.9	90.3	59.8	46.4	64.7	77.4	78.1	79.1
Code generation	HumanEval 0-shot	59.1	57.9	59.1	55.5	34.1	28	37.8	60.4	62.2	65.8
	MBPP 3-shot	53.8	62.5	71.4	74.5	51.5	50.8	60.2	67.7	77.8	79.4

Figure 2: Example of how model results are typically reported; this is from when Microsoft released Phi-3. Each column is a different language model, and each row a different benchmark. Note that MMLU results are shown in the "Popular Aggregate Benchmarks" section, on the second row [6].

Static benchmarking remains popular due to the inherent simplicity allowing for quick and easy comparisons between models and tracking improvements over time.

However, there are downsides to using static benchmarks:

- Due to significant quality advancements in new models, **benchmark saturation** is often observed when the reported scores become too high and may no longer represent meaningful improvements in quality. For example, MMLU-Pro was proposed as a more challenging alternative to MMLU after the scores on the latter plateaued, highlighting the need for constant updates to benchmarks as models evolve [7].
- The versatility of LLMs now allows for them to be applied to more creative tasks, however, **obtaining ground-truth data for these tasks can be challenging, sometimes impossible**. Consequently, most static benchmarks may not be able to capture the nuanced differences in quality between different models for creative tasks [2].
- Also, there is a **chance of data from popular benchmarks entering the training data for new models**, this can undermine the trust in the reported results.

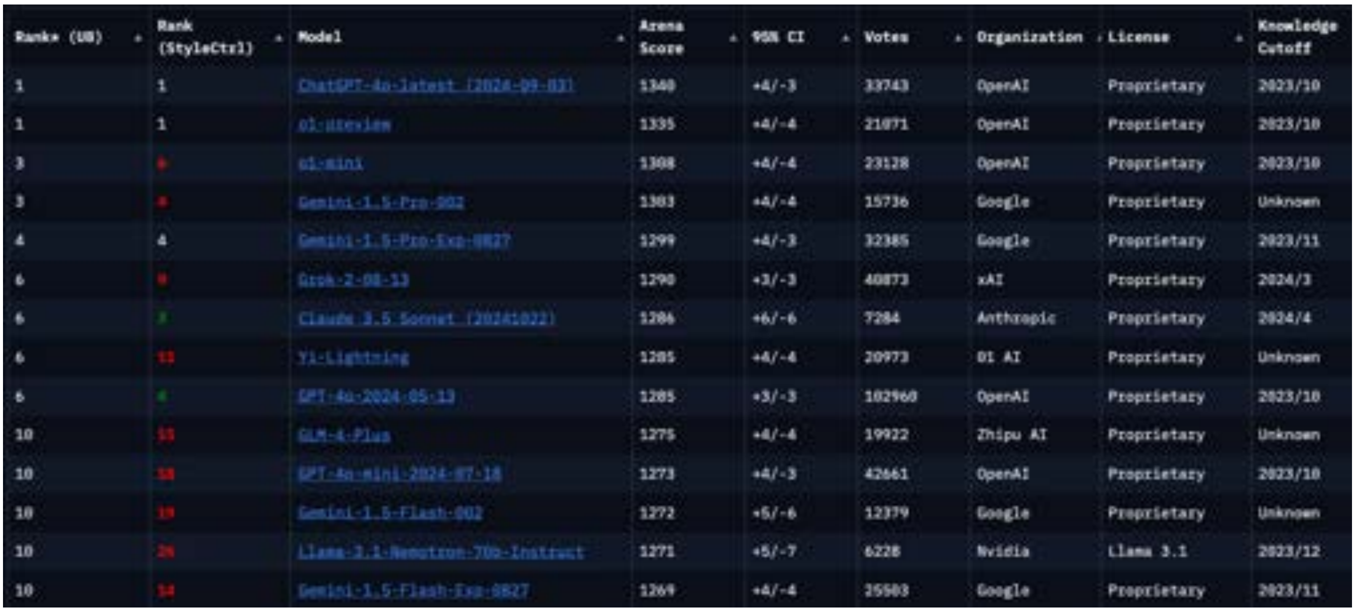
Chatbot Arena: A crowdsourced approach to LLM evaluation

As a solution to the challenges associated with static benchmarks, LMSYS (The Large Model Systems Organization) built a crowdsourced platform for collecting anonymous feedback on LLMs – Chatbot Arena [2]. As of November 2024, for the 169 models submitted, more than 2,200,000 human pairwise comparisons have been collected to build a leaderboard. The leaderboard is for both Proprietary and Open models. Anyone can contribute to scoring the LLMs by going to https://lmarena.ai/, entering a query and choosing the better response out of two options. The user does not know which models were used to generate those responses. After the collection of human preferences, the ranking is derived by calculating "Arena Scores" that indicate the relative "strength" of models when compared with others. The scoring system is based on Elo ratings, which is often used in games like chess to estimate the relative skill levels of two players [8].

The score should be interpreted as follows; *the difference in scores between two models is a predictor of the likely outcome of the comparison between their responses.* The bigger the difference, the higher the probability the higher ranked model will generate "better" output.

For a detailed explanation of the conversion from scores to probabilities, we refer the interested reader to this https://lmsys.org/blog/2023-12-07-leaderboard/ [9] *authored by the creators of Chatbot Arena.* The conversion is not obvious when looking only at scores and probabilities. For example, if the difference between two models' ratings is 50, the probability of the model with a higher score winning is 57%, while if the difference was 400, the probability would be 91%. Furthermore, it is important to note that a single score cannot be interpreted in absolute terms, a difference in scores is required. This is different to interpreting static benchmarks, which are absolute.

Since its inception in May 2023, Chatbot Arena has become the most popular venue to observe top-ranked models.



Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	1	ChatGPT-4o-latest (2024-09-03)	1340	+4/-3	33743	OpenAI	Proprietary	2023/10
1	1	o1-preview	1335	+4/-4	21071	OpenAI	Proprietary	2023/10
3	6	o1-mini	1308	+4/-4	23128	OpenAI	Proprietary	2023/10
3	4	Gemini-1.5-Pro-002	1303	+4/-4	15736	Google	Proprietary	Unknown
4	4	Gemini-1.5-Pro-Exp-0827	1299	+4/-3	32385	Google	Proprietary	2023/11
6	8	Grok-2-08-13	1290	+3/-3	40873	xAI	Proprietary	2024/3
6	3	Claude 3.5 Sonnet (20241022)	1286	+6/-6	7284	Anthropic	Proprietary	2024/4
6	11	Yi-Lightning	1285	+4/-4	20973	01 AI	Proprietary	Unknown
6	4	GPT-4o-2024-05-13	1285	+3/-3	102960	OpenAI	Proprietary	2023/10
10	15	GLM-4-Plus	1275	+4/-4	19922	Zhipu AI	Proprietary	Unknown
10	18	GPT-4o-mini-2024-07-18	1273	+4/-3	42661	OpenAI	Proprietary	2023/10
10	19	Gemini-1.5-Flash-002	1272	+5/-6	12379	Google	Proprietary	Unknown
10	26	Llama-3.1-Nemotron-70b-Instruct	1271	+5/-7	6228	Nvidia	Llama 3.1	2023/12
10	14	Gemini-1.5-Flash-Exp-0827	1269	+4/-4	25503	Google	Proprietary	2023/11

Figure 3: Screenshot of the Chatbot Arena Leaderboard, in November 2024 [10]**.**

The leaderboard has the following columns:

- **Rank (UB)** – model's ranking defined as "*1 + the number of models that are statistically better than the target model*".
- **Rank (StyleCtrl)** – model's ranking that takes into account "*style control*", which factors in various characteristics of the responses such as length and markdown usage.
- **Model** – model name. In some cases, it contains a date to specify the exact version of the model.
- **Arena Score** – score that is derived from the results of pairwise comparisons between models.
- **95% CI** – 95% confidence intervals for the Arena Score.
- **Votes** – the number of votes used in calculating the score.
- **Organization** – company or research lab that created the model.
- **License** – name of the license associated with the model. "Proprietary" if access to the model is fully controlled by the company that created it.
- **Knowledge Cutoff** – date when the training data was last updated.

Note: Many Proprietary models are only available via an API, with the same model having different version changing with time. So, when possible, model names on the leaderboard try to capture as much information about the model checkpoint as possible, for example the date (e.g, "*Claude 3.5 Sonnet (20241022)*").

Overview of current state-of-the-art models seen at the top of the Chatbot Arena Leaderboard:

- OpenAI models [1]:
    - **GPT-4o**: LLM released in May 2024, which was designed to be able to handle multiple types of both inputs and outputs (text, audio, image, videos), while matching the performance of the previous generation of models, GPT-4-Turbo on text in English and code.
    - **o1-preview**: LLM that was released in September 2024 and was designed for reasoning through complex tasks by "thinking" longer before generating the response.
    - **o1-mini**: a smaller version of "o1-preview".
- Gemini models by Google DeepMind [11]:
    - **Gemini 1.5 Pro**: the best model for general performance across various tasks by Google DeepMind, which, similar to GPT-4o, can process multiple types of inputs, but can only return text. It has a very long context window, i.e., the technical characteristic of LLMs that specifies the maximum length of input that they can process in a single call.
    - **Gemini 1.5 Flash**: a fast and versatile version of Gemini 1.5 models that also supports multiple types of inputs.
- Grok models by xAI [12]:
    - **Grok-2**: the frontier LLM by xAI that, according to the company's announcement, demonstrates state-of-the-art reasoning capabilities.
- Claude models by Anthropic [13]:
    - **Claude 3.5 Sonnet**: the most performant model by Anthropic with a particular focus on safety, showing state-of-the-art performance in vision tasks such as visual reasoning, interpreting charts and graphs, as well as general text and code understanding and generation tasks.
- Yi models by 01.AI [14]:
    - **Yi-Lightning**: the latest high-performance model by 01.AI, which is a leading Chinese developer of LLMs that offers both Proprietary and Open models that are amongst the best on the Chatbot Arena leaderboard.
- GLM models by Zhipu AI [15] :
    - **GLM-4-Plus**: a Proprietary model developed by the Beijing-based company Zhipu AI that demonstrates similar performance to the best models by OpenAI, Google Deepmind, Anthropic and others on both benchmarks and Chatbot Arena.
- Llama-3.1-Nemotron models by NVIDIA [16]:
    - **Llama-3.1-Nemotron-70B-Instruct**: at the time of writing this publication, this was the best open LLM that was fine-tuned based on the Llama-3.1 model developed by Meta.
- Llama 3.1 models by Meta [17]:
    - **Llama 3.1-405B-Instruct**: the first Open LLM that approaches state-of-the-art performance across most tasks and capabilities.
    - **Llama 3.1-70B-Instruct**: a smaller version of Llama 3.1 family of models.

Disclaimer: *Stating the obvious – AI is a fast-changing field and therefore the leaderboard is expected to change. This is both as there are new models, but also, there are new metrics being introduced to expand the leaderboard's information. This change can happen overnight, as the authors experienced several times during the writing of this blog! The snapshot of the leaderboard for the below analysis was taken on 4th November 2024.*

Chatbot Arena Leaderboard: How understanding the leaderboard allows for insights in the field of AI

We wish to demonstrate that understanding the Chatbot Arena leaderboard, and the knowledge of how LLMs quality is being judged, provides information relevant to important issues in the field of AI.

To provide this insight, and for reference in the follow sections, we display the top 5 models from the Chatbot Arena leaderboard when it was launched in May 2023. At this point in time there were two dominant state-of-the-art models; GPT-4 (OpenAI) and Claude v1 (Anthropic).

This dominance is displayed in **Figure 4**, with win rates ranging between 67 - 92% against three of the remaining top 5 models. The 4th and 5th best overall models, being the two best Open models, Vicuna-13B and Koala-13B, could only manage win rates of 18% and 8% against GPT-4, respectively.

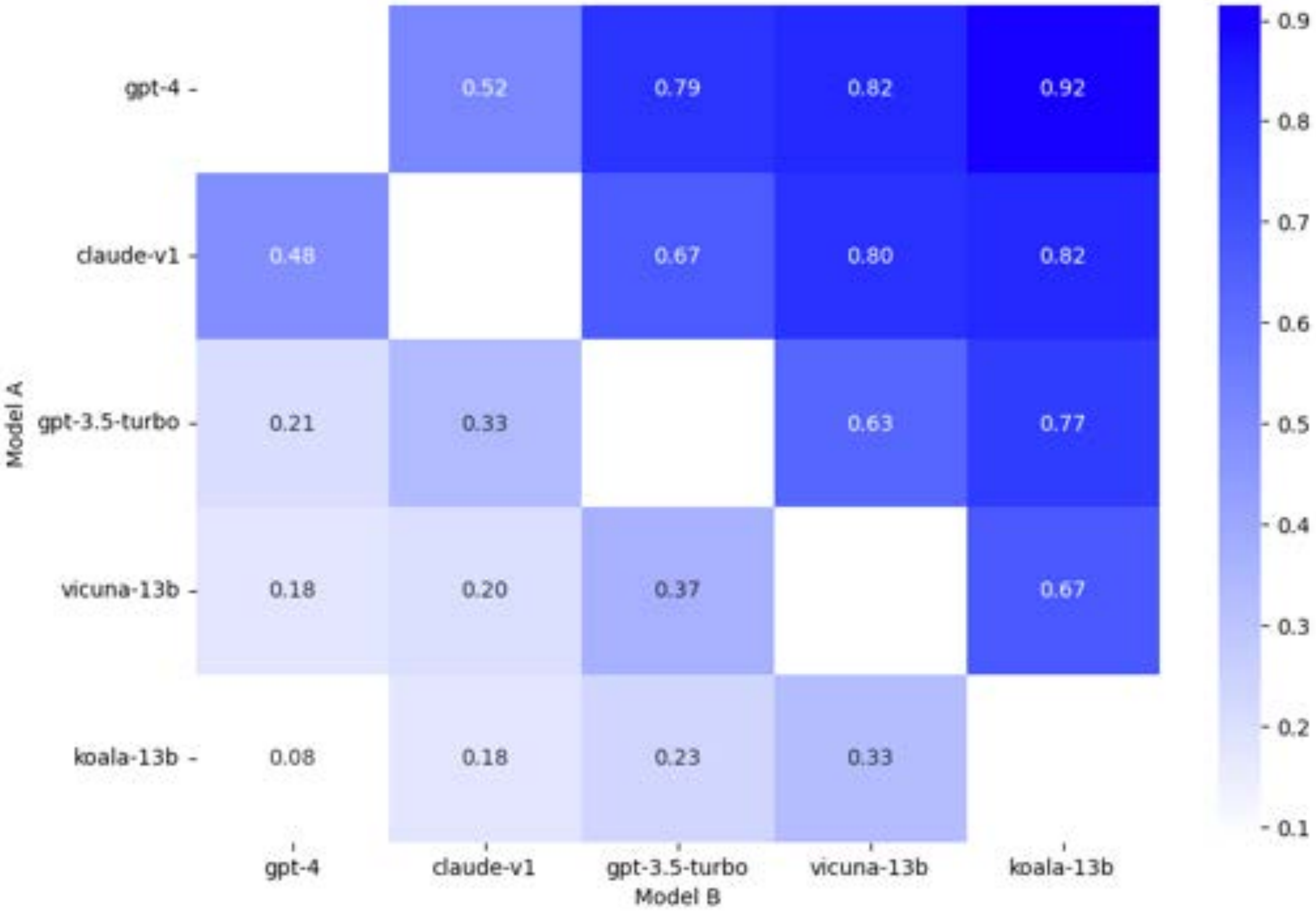


Figure 4: Top 5 models on Chatbot Arena in May 2023. There are two dominant models, GPT-4 and Claude v1. Shown in the top right corner are these two models win rates against the other three models, ranging from 67% - 92%.

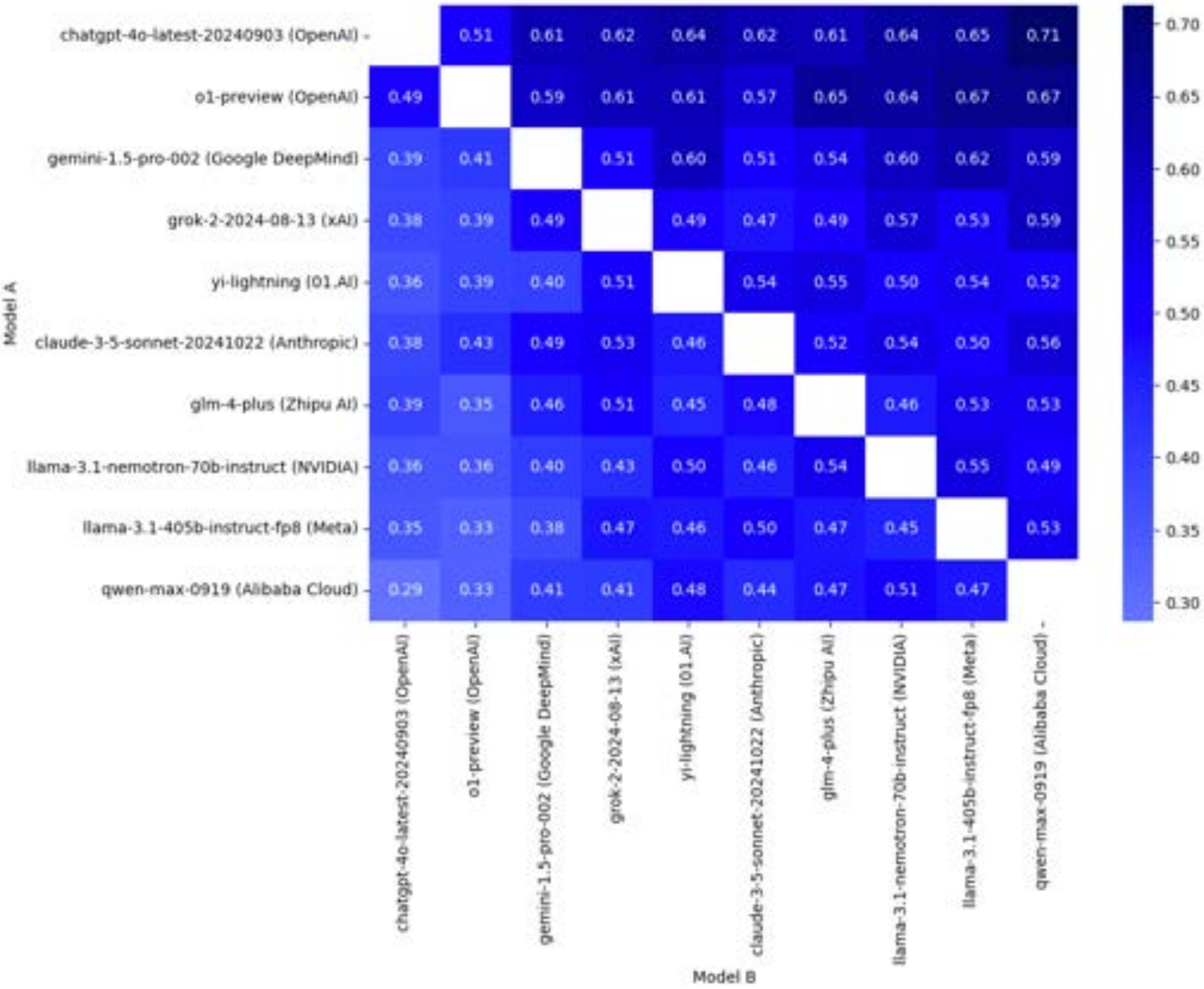
Comparing the current leaderboard from November 2024 with that from May 2023 allows for insight into how the field of AI is developing. We'll briefly expand on the following two points:

1. Although LLM quality is continually improving, there is evidence of diminishing differentiation in quality between new state-of-the-art LLMs.
2. Open models – although yet to surpass Proprietary models in terms of overall quality – are seen to nowadays perform comparatively well. For a more specific tasks, such as coding, the Chatbot Arena leaderboard rankings indicates that Open models are within reach of gaining parity with Proprietary models.

Insight 1: LLM quality is continually improving, however, there is evidence of diminishing differentiation in quality between new state-of-the-art LLMs

The number of LLMs exhibiting state-of-the-art performance has dramatically increased with more companies and research labs now producing LLMs of similar perceived quality than before. Over the past 18 months, models released by Google DeepMind (*Gemini* models), xAI (*Grok* models), Meta (*Llama* models), 01.AI (*Yi-Lightning*), Zhipu AI (*GLM-4-Plus*) and others have demonstrated competitive performance.

Figure 5 displays the win rates of the top models in November 2024. Compared with the data from May 2023 in **Figure 4**, there are now more LLM providers offering state-of-the-art models of similar quality. Unlike in May 2023, there are no longer two models dominating the win rate, instead, many of the top models perform well against each other. For this blog, we do not wish to be overly specific in our analysis. Instead, we wish to draw attention to the most obvious feature of this comparison; most win rates in the top right of the table are currently only around 50 - 71%, whereas, compared to May 2023 these numbers were much higher, up to 92%. Only one pair is currently observed to have a win rate exceeding 70%, “chatgpt-4o-latest-20240903” versus “qwen-max-0919”.



	chatgpt-4o-latest-20240903 (OpenAI)	o1-preview (OpenAI)	gemini-1.5-pro-002 (Google DeepMind)	grok-2-2024-08-13 (xAI)	yi-lightning (01.AI)	claude-3-5-sonnet-20241022 (Anthropic)	glm-4-plus (Zhipu AI)	llama-3.1-nemotron-70b-instruct (NVIDIA)	llama-3.1-405b-instruct-fp8 (Meta)	qwen-max-0919 (Alibaba Cloud)
chatgpt-4o-latest-20240903 (OpenAI)		0.51	0.61	0.62	0.64	0.62	0.61	0.64	0.65	0.71
o1-preview (OpenAI)	0.49		0.59	0.61	0.61	0.57	0.65	0.64	0.67	0.67
gemini-1.5-pro-002 (Google DeepMind)	0.39	0.41		0.51	0.60	0.51	0.54	0.60	0.62	0.59
grok-2-2024-08-13 (xAI)	0.38	0.39	0.49		0.49	0.47	0.49	0.57	0.53	0.59
yi-lightning (01.AI)	0.36	0.39	0.40	0.51		0.54	0.55	0.50	0.54	0.52
claude-3-5-sonnet-20241022 (Anthropic)	0.38	0.43	0.49	0.53	0.46		0.52	0.54	0.50	0.56
glm-4-plus (Zhipu AI)	0.39	0.35	0.46	0.51	0.45	0.48		0.46	0.53	0.53
llama-3.1-nemotron-70b-instruct (NVIDIA)	0.36	0.36	0.40	0.43	0.50	0.46	0.54		0.55	0.49
llama-3.1-405b-instruct-fp8 (Meta)	0.35	0.33	0.38	0.47	0.46	0.50	0.47	0.45		0.53
qwen-max-0919 (Alibaba Cloud)	0.29	0.33	0.41	0.41	0.48	0.44	0.47	0.51	0.47	

Figure 5: Top providers on Chatbot Arena in November 2024. When compared directly with Figure 4, several of the top state-of-the-art models are currently seen to be performing well against each other. Only one pairing in the top right corner exceeds 70%.

Insight 2: Open models – although yet to surpass Proprietary models in terms of overall quality – are nowadays observed to perform relatively well.

Whether Proprietary models will continue to dominate the leaderboards is an important issue for organisations that are looking to build AI applications. It is interesting to observe that as of November 2024 there are several Open models of notable quality; for example, Meta (Llama-3.1-405B-Instruct), NVIDIA (Llama-3.1-Nemotron-70B-Instruct), Mistral (Mistral Large), Nexusflow (Athene-V2), DeepSeek-AI (DeepSeek-v2.5) and Alibaba Cloud (Qwen-Max).

This illustrate Open model quality **Figure 6** displays the win rates of two Open models, “llama-3.1-nemotron-70b-instruct” by NVIDIA and “llama-3.1-405b-instruct-fp8” by Meta; these are the two highest ranked Open models in November 2024. Their win rates against the top 7 Proprietary models, although not yet at parity with the top 4 models, GPT-4o, o1-preview, Gemini 1.5 Pro, and Grok-2, indicate that they are at parity with the 0.1.AI, Anthropic and Zhipu AI models.

For comparison, using the data from **Figure 4**, the win rate range of Open versus Proprietary models observed in May 2023 was **8 - 37%**, whereas currently, the range is **33 - 54%**. This appears to be significant improvement for Open models in overall quality.

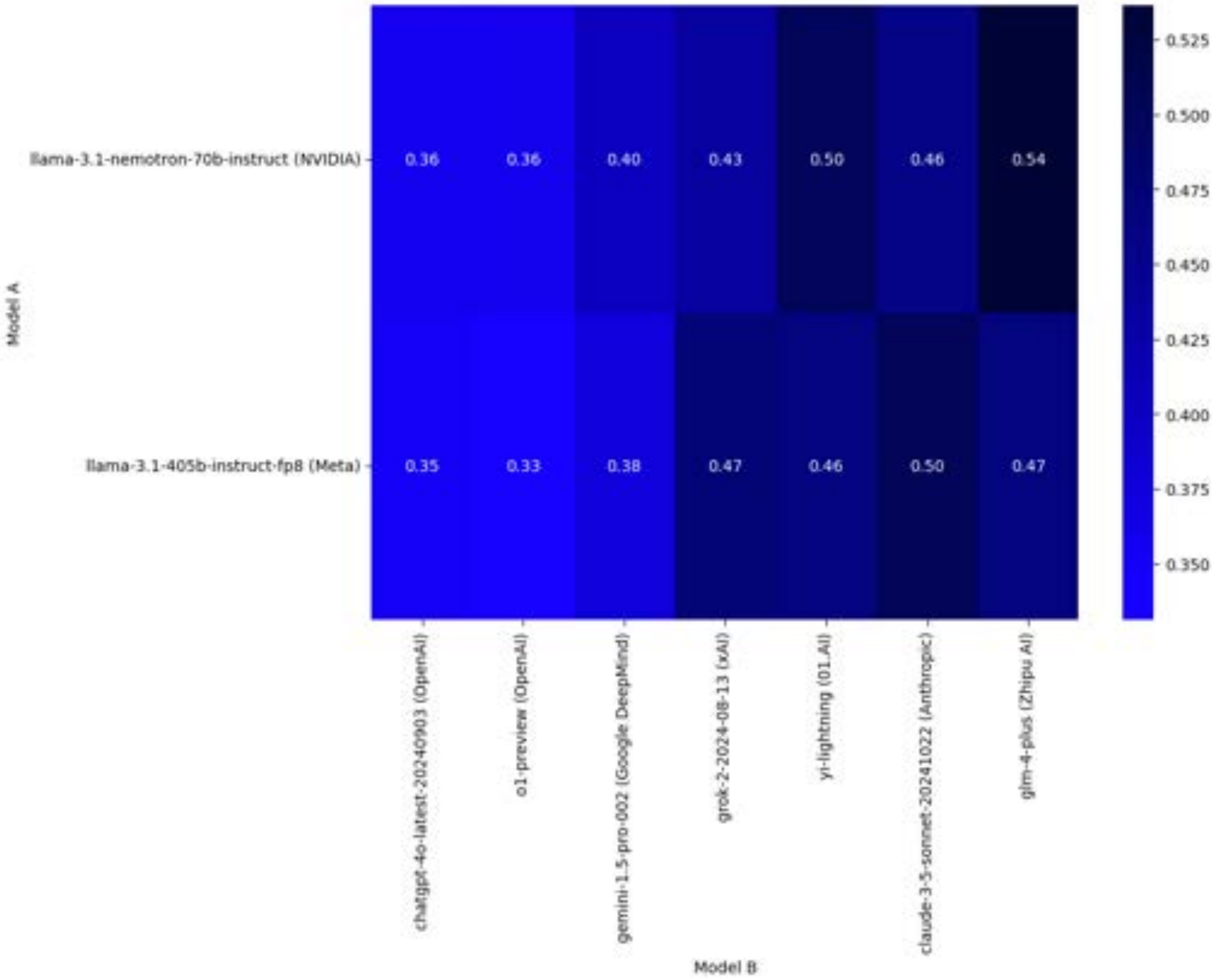
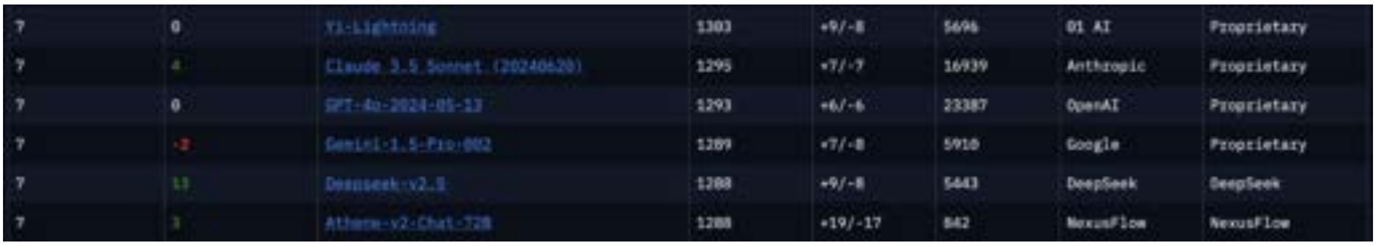


Figure 6: The top 2 Open models win rates against some of the best Proprietary models. Note that there are three pairs of 50% or higher.

Finally, we wish to bring attention to how the Chatbot Arena leaderboard also supports viewing different categories of prompts, for example for code-related tasks. This allows for identification of Open models that excel in specific tasks.

DeepSeek-V2.5 [18] and Athene-v2-Chat-72B [19] are two such Open models, that have advanced coding abilities. As of November 2024, both models surpassed many major Proprietary models in code-specific tasks, and, have a higher leaderboard ranking than popular Proprietary models such as GPT-4o ("2024-08-06" version), and GPT-4-Turbo ("2024-04-09" version), and share the same ranking with Gemini 1.5 Pro ("002" version) and Claude 3.5 Sonnet ("20240620" version). Note that at the time of writing, our opinion is that an AI development team would be forgiven for defaulting to GPT-4o, Gemini 1.5 Pro or Claude 3.5 Sonnet, simply due to those models' overall performance and general popularity.



7	0	Yi-Lightning	1303	+9/-8	5696	01 AI	Proprietary
7	4	Claude 3.5 Sonnet (20240620)	1295	+7/-7	16939	Anthropic	Proprietary
7	0	GPT-4o-2024-05-13	1293	+6/-6	23387	OpenAI	Proprietary
7	-2	Gemini-1.5-Pro-002	1289	+7/-8	5910	Google	Proprietary
7	13	Deepseek-v2.5	1288	+9/-8	5443	DeepSeek	DeepSeek
7	3	Athene-v2-Chat-72B	1288	+19/-17	842	NexusFlow	NexusFlow

Figure 7: Chatbot Arena Leaderboard ranking for the specific task of Coding. All the models shown have been allocated the same overall ranking, of 7th. Note that the final two shown, Deepseek-v2.5 and Athene-v2-Chat-72B, are Open models, and are performing comparable to popular models such as GPT-4o-2024-05-13 and Gemini-1.5-Pro-002.

Although we view the issue of whether Open models will achieve parity with Proprietary models of being of importance to the field of AI, the aim of this blog is not to conclusively prove or disprove this point. Instead, we wish to demonstrate that understanding the Chatbot Arena leaderboard, and the knowledge of how LLMs quality is being judged, provides insights into important issues, such as Open versus Proprietary model selection.

Final Thoughts: Will LLMs be used to evaluate themselves?

In this technical blog, we've discussed how the quality of state-of-the-art LLMs is currently being reported. The use of static benchmarks, whilst intuitive, are potentially limited to the more advanced and nuanced tasks being offered by recent LLMs. Leaderboards, such as Chatbot Arena based on direct human-pairwise comparisons, have emerged as a popular alternative to static benchmarks. At LSEG Analytics we use the information within these arenas to gain insights into broader AI trends, such as the diminishing difference between the top state-of-the-art models.

Given the Chatbot Arena's success, it is worth considering whether an organisation might benefit from setting up an arena internally, to compare between candidate models for a more specific task. This would enable employees to crowdsource the best performing LLMs for their products. However, there are two main drawbacks to this approach; a workforce may not be qualified to provide appropriate feedback and scaling to a wider audience can be expensive. Given human feedback is inherent to the current arena setup, such labour costs might be unavoidable if large amounts of feedback are required.

LLM-as-a-Judge

Given the obvious issue of the significant manual effort to crowdsourcing results for an arena, and also, given the dramatic increase in quality of LLMs, it is natural to ask the question - ***can LLMs be used to evaluate themselves?*** If an LLM can generate human-like written content, can it also assess the quality of generated text like a human?

Such a technique is commonly referred to as "*LLM-as-a-Judge*" [20] [21]. Whilst promising, the quality of this approach isn't yet at the point where LLM's can be routinely relied upon to replace humans for quality evaluation. As early as Dec 2023 - which is a significant time ago in the fast-changing field of AI - researchers behind the Chatbot Arena claim to achieve over 80% agreement between LLM judges and human preferences. This roughly matches the level of agreement between humans [21]. Given that LLM-as-a-Judge has not been routinely adopted yet goes to show the difficulty in automating evaluation of current state-of-the-art models.

LSEG Analytics looks forward to sharing a follow up publication discussing the use LLM-as-a-judge, an aspect we think clearly is one to keep an eye on in the field of AI.

Figures

	Description	Reference
	MMLU Benchmark Question Examples	"Measuring Massive Multitask Language Understanding" [4] https://arxiv.org/abs/2009.03300
	Phi-3 Model Announcement Benchmark Results	"Introducing Phi-3: Redefining what's possible with SLMs" [6] https://azure.microsoft.com/en-us/blog/introducing-phi-3-redefining-whats-possible-with-slms/
	Chatbot Arena Leaderboard Screenshot, early November 2024	Screenshot of https://lmarena.ai/ [10] taken in early November 2024
	Top-5 Models on Chatbot Arena (May 2023), Fraction of Model A Wins	Figure plotted using publicly available data from https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard
	The Best Models by Top LLM Providers on Chatbot Arena (November 2024), Fraction of Model A Wins	Figure plotted using publicly available data from https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard
	The Best "Open" Models vs. The Best "Proprietary" Models on Chatbot Arena (November 2024), Fraction of Model A Wins	Figure plotted using publicly available data from https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard
	DeepSeek-V2.5 and Athene-v2-Chat-72B Open Models with Advanced Coding Abilities	Screenshot of https://lmarena.ai/ [10] taken in early December 2024

References

[1] "OpenAI Model Release Notes," [Online]. Available: https://help.openai.com/en/articles/9624314-model-release-notes.

[2] W. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. I. Jordan, J. E. Gonzalez and I. Stoica, "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference," [Online]. Available: https://arxiv.org/pdf/2403.04132.

[3] A. Wang, A. Singh, J. Michael, F. Hill, O. L. Levy and S. R. Bowman, "GLUE: A Multi-Task Benchmark And Analysis Platform For Natural Language Understanding," [Online]. Available: https://arxiv.org/pdf/1804.07461.

[4] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song and J. Steinhardt, "Measuring Massive Multitask Language Understanding," [Online]. Available: https://arxiv.org/pdf/2009.03300.

[5] "Gemini Ultra," [Online]. Available: https://deepmind.google/technologies/gemini/ultra/.

[6] B. Misha, "Introducing Phi-3: Redefining what's possible with SLMs," [Online]. Available: https://azure.microsoft.com/en-us/blog/introducing-phi-3-redefining-whats-possible-with-slms/.

[7] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue and W. Chen, "MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark," [Online]. Available: https://arxiv.org/pdf/2406.01574.

[8] "Elo Rating System," [Online]. Available: https://en.wikipedia.org/wiki/Elo_rating_system.

[9] "Chatbot Arena: New models & Elo system update," [Online]. Available: https://lmsys.org/blog/2023-12-07-leaderboard/.

[10] "Chatbot Arena (formerly LMSYS)," [Online]. Available: https://lmarena.ai/ .

[11] "Google DeepMind Gemini," [Online]. Available: https://deepmind.google/technologies/gemini/.

[12] "Grok-2 Beta Release," [Online]. Available: https://x.ai/blog/grok-2.

[13] "Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku," [Online]. Available: https://www.anthropic.com/news/3-5-models-and-computer-use.

[14] "Yi Models by 01.AI," [Online]. Available: https://www.01.ai/.

[15] "GLM-4-Plus Model," [Online]. Available: https://bigmodel.cn/dev/howuse/glm-4.

[16] "Llama-3.1-Nemotron-70B-Instruct," [Online]. Available: https://huggingface.co/nvidia/Llama-3.1-Nemotron-70B-Instruct.

[17] "Introducing Llama 3.1: Our most capable models to date," [Online]. Available: https://ai.meta.com/blog/meta-llama-3-1/.

[18] "DeepSeek-Coder-V2: Breaking the Barrier of Closed-Source Models in Code Intelligence," [Online]. Available: https://huggingface.co/deepseek-ai/DeepSeek-Coder-V2-Instruct .

[19] "Introducing Athene-V2," [Online]. Available: https://nexusflow.ai/blogs/athene-v2.

[20] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu and C. Zhu, "G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment," [Online]. Available: https://arxiv.org/pdf/2303.16634.

[21] L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez and I. Stoica, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," [Online]. Available: https://arxiv.org/pdf/2306.05685.

[22] T. Li, W. Chiang, E. Frick, L. Dunlap, T. Wu, B. Zhu, J. E. Gonzalez and I. Stoica, "From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and BenchBuilder Pipeline," [Online]. Available: https://arxiv.org/pdf/2406.11939.

Visit **lseg.com** | @LSEGplc LSEG