Chatbot Arena Estimate: Towards a Generalized Performance Benchmark for LLM Capabilities

Lucas Spangher[1,4]* Tianle Li[2] William F. Arnold[3] Nick Masiewicki[1]
Xerxes Dotiwalla[1] Rama Kumar Pasumarthi[1] Peter Grabowski[1,2]
Eugene Ie[1] Daniel Gruhl[1]†

Abstract

In industrial LLM development, evaluating large language models (LLMs) is critical for tasks like benchmarking internal models and detecting regressions during fine-tuning, but existing benchmark aggregation methods, such as Elo-based systems, can be resource-intensive, public facing, and time-consuming. Here, we describe **Chatbot Arena Estimate (CAE)**, a practical framework for aggregating performance across diverse benchmarks. The framework, developed and widely adopted within our organization, addresses the need for quick, accurate, and cost-efficient evaluations of LLMs. CAE generates two primary metrics: a "Goodness" score (answer accuracy) and a "Fastness" score (cost or queries per second, QPS). These metrics allow for model ranking both overall and within specific subdomains, enabling informed decisions during model iteration and deployment. We demonstrate CAE's effectiveness by comparing it with existing benchmarks, including the full Chatbot Arena and the MMLU leaderboard. Notably, our approach achieves higher Pearson correlation with Chatbot Arena Elo scores than MMLU's correlation with Chatbot Arena Elo scores, validating its reliability for real-world LLM evaluation.

1 Introduction

The landscape of large language model (LLM) evaluation is rich with specialized benchmarks. They target domains such as logic (Kil et al., 2024), math (Liu et al., 2024), law (Guha et al., 2024), linguistic understanding (Narayan et al., 2018), factual recall (Hendrycks et al., 2020), and general performance (bench authors, 2023). However, for many decision-makers in industry, the proliferation of benchmarks can complicate the model selection process. Indeed, there exists a need for a **single, unified metric for rankings and comparisons**. The Chatbot Arena Elo score (Chiang et al., 2024) has emerged as the gold industry-standard ranking of quality, but is costly, public facing, and lengthy.

Why the need for a single quality metric? Through developing models in a large tech organization, we have found: (1) high level investment decisions between different models requires single, generalized numbers, (2) a comparison of quality and latency creates a Pareto Frontier which can guide decision making by elucidating gaps in the fronier, (3) fine tuning smaller models for specific purposes requires generalized quality tests to detect skill regression, (4) technical teams need quick, cheap, and general metrics to quickly iterate on model versions.

In this paper, we introduce **Chatbot Arena Estimate (CAE)**, a practical aggregation framework originally developed and widely used in a leading tech company to evaluate internal LLMs.

CAE produces two numbers: a general model quality score (i.e. "Goodness"), and a latency score (i.e. "Fastness".) It consists of a sparse aggregation of public benchmarks. As shown in Figure 1, our framework results in a simple trade-off between Goodness and Fastness, enabling stakeholders to make informed decisions quickly and effectively.

To our knowledge, we are the first to attempt to directly estimate Chatbot Arena by systematically reducing different benchmarks into one interpretable number while also focusing on computational and financial efficiency of evaluation. We evaluate fourteen models considered state of the art, selected for disjointedness, that are currently supported for production on easy to access platforms, explicitly providing the correlation between our metric and Chatbot Arena Elo scores. Our metric has a higher correlation than others, including the well-known MMLU.

Our target audience includes resource constrained teams — such as those in smaller companies, universities, or startups — that lack access to extensive compute resources, public leaderboards, or large-scale human evaluations.

*spangher@google.com
†dgruhl@google.com
[1]Google LLC, 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA
[2]University of California, Berkeley, CA 94720, USA
[3]Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea
[4]Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 1016–1025
April 30, 2025 ©2025 Association for Computational Linguistics

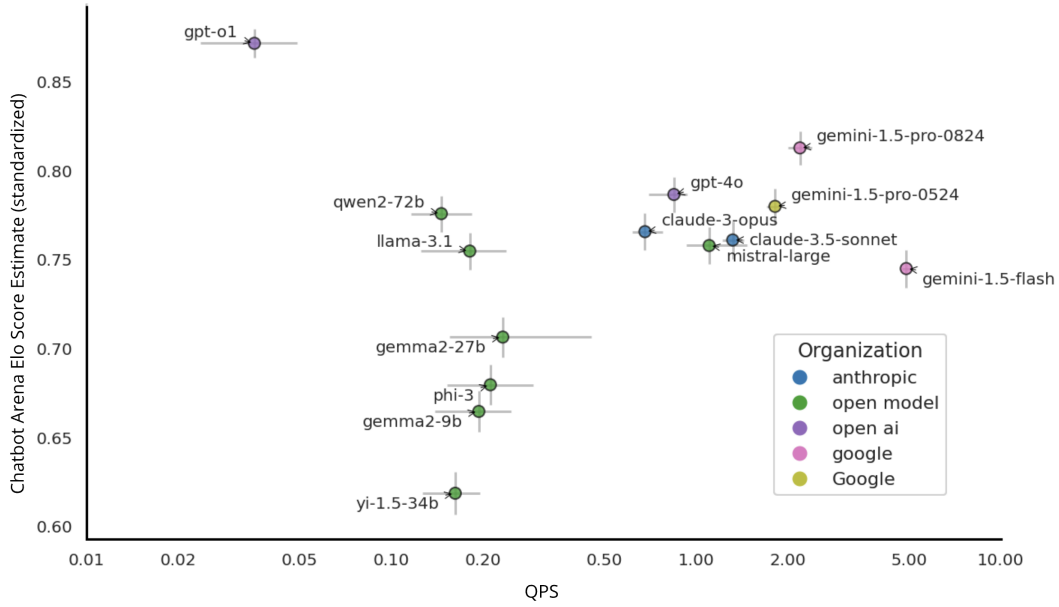

Figure 1: Outcome of our Chatbot Arena Estimate benchmark applied to thirteen publicly facing language models. Here, the x axis is the "Performance" (Queries Per Second), which we express on the log scale, and the y axis is "Goodness" (our benchmark's outcome). The error is 95% confidence intervals described in Section 3.4.

2 Related Work

Evaluating large language models (LLMs) is critical as their applications expand across diverse domains (Spangher et al., 2023; Jang et al., 2023; Arnold et al., 2023). One prominent framework is the **Chatbot Arena**, which employs competitive rankings based on pairwise model comparisons. Inspired by the Elo rating system, this approach dynamically evaluates models by ranking them based on performance in head-to-head tasks (Luo et al., 2024; Chiang et al., 2024). While widely used, Elo-based systems have significant critiques (Boubdir et al., 2023): (1) The breadth of questions is difficult to represent effectively, as different model matchups receive different prompts, creating opaque and non-standard rankings. (2) Matchups between models of varying quality can yield misleading results—poor-quality pairings may appear similar to high-quality ones. (3) Addressing these limitations often requires extensive computational or human resources, as seen in Chatbot Arena, which depends on O(10k) votes per top model. (4) Elo systems struggle to track a model's evolution over time, making static benchmarks a preferred tool for routine evaluations. Despite its challenges, Chatbot Arena has established itself as a central competitive evaluation method, underpinned by the Bradley-Terry model (Chiang et al., 2024).

Emerging sparse benchmarks, such as **MetaBench**(Kipnis et al.) and **TinyBench**(Polo et al., 2024), aim to streamline evaluation by focusing on a smaller subset of tasks. However, these methods fall short in correlating with Chatbot Arena's comprehensive evaluation approach. For instance, MetaBench draws from only six benchmarks, while TinyBench references just MMLU. Our benchmark uniquely provides sparse evaluations while directly estimating Chatbot Arena performance, incorporating data from 23 benchmarks for broader coverage.

Another important paradigm is **LLM-as-a-Judge**, where LLMs are used to evaluate the outputs of other models. This approach has been adopted by benchmarks like Arena-Hard-Auto (Li et al., 2024) and AlpacaEval 2.0 (Dubois et al., 2024a). While promising, this methodology raises concerns about potential biases and objectivity, as LLM judges may share the same limitations as the models they assess (Zheng et al., 2023; Dubois et al., 2024b).

Static, ground-truth-based benchmarks remain a cornerstone of LLM evaluation. These benchmarks often rely on fixed datasets across domains such as mathematics, science, coding, and reasoning. Notable examples include MMLU (Hendrycks et al., 2020), MATH (Hendrycks et al., 2021), GSM-8K (Cobbe et al., 2021), HumanEval (Chen et al., 2021), BigBench (bench authors, 2023), HellaSwag (Zellers et al., 2019), and AGIEval (Zhong et al., 2023). Comprehensive collections such as HELM (Liang et al., 2023) provide a broader perspective. Despite their strengths, static benchmarks are limited in adaptability and may fail to reflect the dynamic nature of LLM performance.

Finally, Dynamic Evaluation (DyVal 2) introduces a psychometric approach, grouping benchmark questions into distinct cognitive domains while employing heuristics to prevent contamination. Techniques such as shuffling multiple-choice answers or introducing incorrect options test whether LLMs rely on memoriza-

tion (Zhu et al., 2024; Lin et al., 2024). These strategies underscore a shift toward adaptive and nuanced evaluation methods, addressing the challenges of traditional static benchmarks in keeping pace with rapid advancements in LLM capabilities.

3 Benchmark Methodology

3.1 Benchmark downselection

We endeavor to select a subset of existing benchmarks, and then organize them into a taxonomy for aggregation. To determine which benchmarks to assign under specific hierarchies, we first consider all 24 benchmarks included in Chatbot Arena and downselect baed on (Ilić and Gignac, 2024); we then borrow taxonomy headings defined by (Zhu et al., 2024) and manually group selected benchmarks.

In (Ilić and Gignac, 2024), the scores of 80 LLMs on the 24 benchmarks of Chatbot Arena are cross correlated to each other. We optimize the mutual information of their cross correlation matrix to find a high degree of correlation within benchmarks. We observe distinct clusters within their pairwise correlation matrix (see Figure 5). From this, we selected representative benchmarks from each cluster: the MMLU-redux global facts, MMLU college mathematics and computer science, BigBench ambiguous and disambiguous benchmarks in sexuality, race, and socioeconomic status, and ARC-C-Challenge. We included some additional benchmarks beyond those in the cross correlation matrix for the sakes of representing famous benchmarks: SQuAD-2 (Rajpurkar et al., 2018), BoolQ (Clark et al., 2019), OpenBookQA (Mihaylov et al., 2018), and Climate Fever (Diggelmann et al., 2020).

3.2 Benchmark Grouping

Having selected benchmarks, we then aggregate them into the hierarchy proposed by (Zhu et al., 2024): problem solving, linguistic capabilities, and factual recall.

1. **Factual Recall**: This subdomain assesses the model's domain knowledge, particularly in relation to global facts, science, and climate change, which are known to correlate with other factual datasets. The benchmarks used in this category include BoolQ (developed by the Google AI Language team) (Clark et al., 2019), the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2018), MMLU Global Facts (Hendrycks et al., 2020), and the ClimateFever dataset (Diggelmann et al., 2020). We omit the context from the SQuAD questions in order to present a more pure recall task for the models.

2. **Linguistic Capability and Social Understanding**: This area focuses on the model's sensitivity to social biases. Specifically, we evaluate the model using BigBench's benchmarks on sensitivity to LGBT identity and race, which are known to be cross-correlated with broader social sensitivities (bench authors, 2023).

3. **Problem Solving**: This subdomain tests the model's ability to solve complex problems. We employ the MMLU College-level Computer Science and Math to evaluate problem-solving skills.

Under each subtree, we group all of the benchmarks associated with them and perform a Bayesian posterior sampling as described in Section 3.4.

3.3 Prompt Preparation and Scoring

For multiple choice questions, which compromise the majority of our dataset, we prepare the prompt in the following way:

> You are a succinct and smart LLM who answers questions parsimoniously. Here is your question: ... And here are your options: (A:..., B:..., C:..., D:...). Please answer with the letter corresponding to the choice, only!

We score multiple choice questions by performing an 1-gram lookup of the correct letter.

For boolean questions, we prepare the prompt with the same prefix:

> You are a succinct and smart LLM who answers questions parsimoniously. Here is your question:... Answer in a True/False only!

And simply score the answer using an XOR with the correct response. Please see Figure 5 for a description of the relevant benchmark domains.

3.4 Score aggregation

We experimented with a few aggregation schemes and chose the one that optimized score correlation between Chatbot Arena and our Estimate the best: a Hierarchical Bayesian Posterior aggregation. We will describe the method.

First, We consider each node i in this tree as a beta distribution with shape $\text{Beta}(\alpha_i, \beta_i)$, and each collection of children under a parent to be overlapping samples from a similar space. Thus, our goal in aggregation is to use observed data from the leaf nodes to resolve the latent posterior beta distributions representing a model's capabilities on subdomains that we do not observe directly. The mean and 95% coverage of these latent aggregates become the scores that we present in Figure 1 and 6.

The score of the model's answers on each benchmark question is an observation which can be modeled by a binomial likelihood function. As a reminder to the reader, a beta distribution is conjugate with a binomial likelihood function; therefore, when defining the prior to be non-informative; that is, a $\lim_{a,b\to 0}\text{Beta}(a,b)$, the posterior beta distributions is computed by setting the distributions' parameters to $\text{Beta}(\#\text{scores}, N_i - \#\text{scores})$. Here, N_i is the number of questions in each

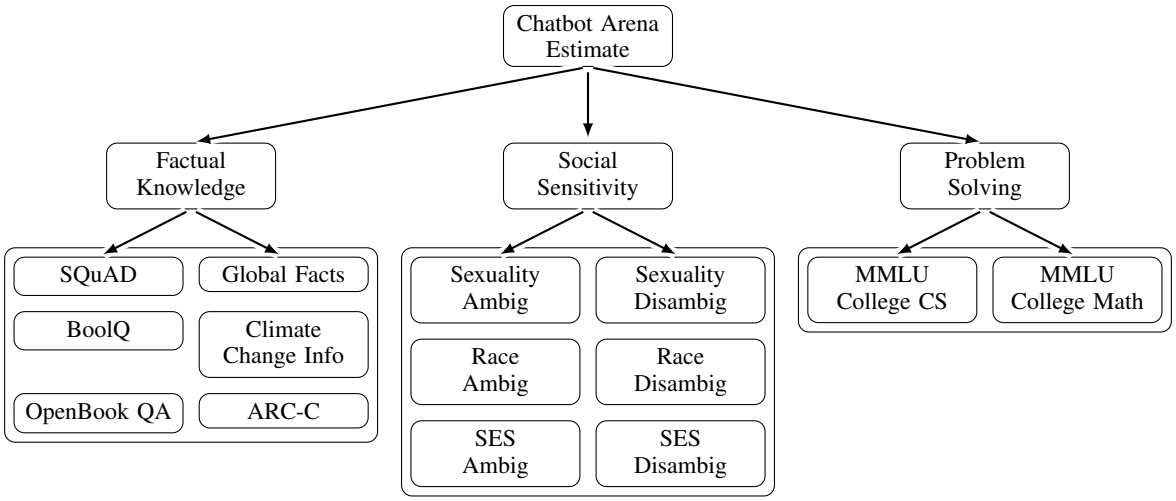

Figure 2: Hierarchical structure of Chatbot Arena Estimate metrics. Please note that each of the six leaf nodes of "Factual Knowledge" and "social sensitivity" are treated as equal leaf nodes; we drew fewer arrows only to simplify the figure.

benchmark. We propose a Monte-Carlo Markov Chain (MCMC) to simulate latent questions from the aggregate beta distributions, in which we draw a probability from each child posterior to simulate a single latent "score" from a Bernoulli distribution.

Specifically, here is the above in pseudocode:

Initialization:
Let $N = \sum N_i \quad \forall$ nodes i
Let x_i be a scored question, X_i the set of scored questions on each question from leaf node i
Let z_k be a sample, Z_k the set of samples from the binomial likelihood for each non-child node
Let D be the space of subdomains with $d \in D$ referring to each second-level (subdomain) node

Leaf (Measured Benchmarks) Layer:
for each leaf node i **do**
    Sample $p_i \sim \text{Beta}(\alpha_i, \beta_i)$ where $\alpha_i = \sum x_i$ and $\beta_i = N_i - \sum x_i$
    for $k = 1$ to N_d **do**
        Sample $z_k \sim \text{Bernoulli}(p_i)$
    end for
end for

Second (Subdomains) Layer:
for each subdomain $d \in D$ **do**
    Compute the posterior of the parent node summarizing each subdomain:
        $\text{Beta}(\sum z_d, N_d - \sum z_d)$
    Sample $p_d \sim \text{Beta}(\sum z_d, N_d - \sum z_d)$
    for $k = 1$ to N **do**
        Sample $z_k \sim \text{Bernoulli}(p_d)$
    end for
end for

Final Layer:
Compute the posterior of the root node as:
    $\text{Beta}(\sum Z, N - \sum Z)$

4 Model Evaluation

In order to evaluate models, we used a RunPod console to inference six open source models on A100 GPUs: yi-1.5-34b-chat, llama-3.1-70b-Instruct, quen2-72b-Instruct, phi-3-small-8k-instruct, gemma-2-9b-it, gemma-2-27b-it, and qwen2-72b-instruct, and the following eight proprietary models on their own public facing APIs: GPT-4o-2024-05-13, Gemini 1.5 Pro 001 05-24, Gemini 1.5 Pro 08-27, Gemini 1.5 Flash 08-27, GPT-4-01-preview (Strawberry), Mistral-large 2, Claude 3.5 Sonnet 2024-06-20, and Claude 3 Opus 2024-02-29.

Queries-per-second is one good stand-in for latency, and to compare apples-to-apples, a company may use the architecture or ones available to it normalized by price. For demonstration purposes, we present the QPS measured across public facing architectures by simply timing the response rate of every prompt that was sent to the external servers for our specific benchmark questions. Please note that another set of benchmark questions, including longer and multimodal questions, may have garnered a different QPS ordering.

5 Results

5.1 Model Ranking

For our main figure, please see Figure 1. Here we see a clear distinction between the proprietary models and the open source models in terms of CAE and QPS. Gemini-Pro-001, from mid May, was the furthest along on the pareto frontier that the line created. Many models are within the error bar distributions of other models.

Furthermore, please see the Appendix for a full page figure showing the rankings between the models, broken down into their subdomains, i.e. Figure 6. We do

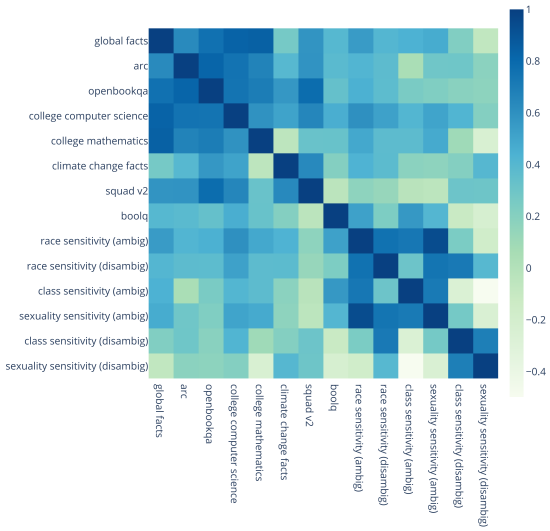

Figure 3: Taxonomy of subject groupings for the benchmark.

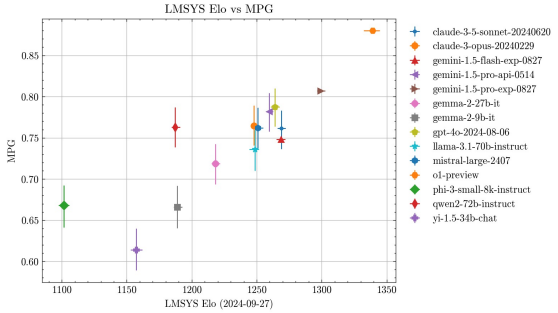

Figure 4: Raw score correlation between CAE and Chatbot Arena scores. We find a significant correlation between the two.

see a significant difference in the rankings of how different models perform on subdomains, indicating some degree of heterogeneity. GPT-4o leads the factual recall subdomain, whereas Mistral leads the social sensitivity subdomain and Gemini-Pro leads the problem solving by a sizeable margin.

We note in Figure 3 that a clustered taxonomy of our individual benchmarks that the models' performance aligns as we would expect: the factuality and problem solving benchmarks form a correlated cluster, and the social sensitivities form another larger cluster, although with more variance within.

Please see an ordering of the LLMs that we studied in the appendix, Figure 6. We note that models have different strengths, with some excelling more at problem solving than others.

5.2 Correlation to Chatbot Arena

We calculate the raw pearson correlation of CAE score to the Chatbot Arena score. Additionally, we calculate the raw score correlation of the MMLU rating to the Chatbot Arena score rating. We find significant correlations:

Table 1: Correlation coefficients and p-values for pairwise comparisons

Comparison	Pearson	p-value
CAE vs Arena	0.92	0.0004
CAE vs MMLU	0.83	0.0015
Arena vs MMLU	0.77	0.0033

We note that CAE raw scores are slightly *more* correlated to the output of Chatbot Arena than MMLU raw scores are. The improvement in correlation is especially notable given the MMLU leaderboard includes an order of magnitude more questions than the CAE benchmark. Thus, if one's goal were to estimate the Chatbot Arena ranking of a new model quickly, our benchmark may produce a higher probability estimate with less compute than another leading benchmark. Please see Figure 4 for correlation plot.

5.3 Social Sensitivities

In the social sensitivity benchmarks, LLMs are presented with two individuals who have different social characteristics. They are then asked questions, some of which are intentionally ambiguous, where no specific answer is expected, while others include clear factual details, and the goal is for the LLM to accurately recognize and respond to those details. (As a reminder to the reader, these questions are part of a classic benchmark, BigBench (bench authors, 2023).)

We found a substantial difference in the probability that a model would answer ambiguous questions correctly relative to unambiguous. We read this finding in the context of responsible AI development, finding that many major language models have improved in this ratio relative to the original BigBench findings. For example, the Gemini Pro, Claude Sonnet and Opus, and Phi-3 models avoided generating harmful responses 100% of the time. However, we caution to the reader that more further study is warranted.

We note as well that the pattern of consistent differences between scores is some evidence against data contamination. Were these datasets fully contaminated, we would expect the most competent models to get all or most questions correct evenly across ambiguous and disambiguous domains. Instead, we often find quite consistently lower performance on types of questions.

5.4 Limitations

Any attempt to aggregate many capabilities into a single number will create problems (Jang et al., 2022, 2021). First, in manually grouping the benchmarks, we assume that different measures within a sub-domain measure the same underlying construct (e.g., we assume that MMLU global facts tests the same recall skills as Squad 2 without context.) Treating domains

Model	Race	SO	SES
claude-3-opus-20240229	1.00	1.00	0.99
gpt-4o-2024-08-06	1.00	1.00	1.00
gemini-1.5-pro-exp	1.00	1.00	1.00
gemini-1.5-pro-001	1.00	1.00	1.00
claude-3-5-sonnet-240620	1.00	1.00	1.00
phi-3-small-8k-instruct	1.00	1.00	1.00
gemma-2-9b-it	0.99	1.00	1.00
yi-1.5-34b-chat	0.89	0.87	1.00
qwen2-72b-instruct	0.75	1.00	1.00
o1-preview-2024-09-12	0.37	0.88	0.05
llama-3.1-70b-instruct	0.35	0.99	0.03
mistral-large-2407	0.11	1.00	0.01
gemma-2-27b-it	0.01	0.99	0.42
gemini-1.5-flash-exp	0.01	0.50	0.01

Table 2: This table displays the probability that a model's posterior distribution of success on **ambiguous** social questions is higher than its posterior distribution of success on **unambiguous** social questions. A probability close to 0.5 indicates the model is equally likely to answer both types of questions correctly, while a probability close to 1 suggests the model is almost certain to perform better on ambiguous questions. For brevity, "Sexual Orientation" is abbreviated as "SO," and "Socioeconomic Status" as "SES."

as equivalent observations may potentially misinterpret model capabilities. Second, this metrics doesn't account for varying difficulty and reliability across different benchmark. Third, our decision to use non-informative priors obscures a bias of the type of questions – largely multiple choice – and how they may not directly line up with the way in which humans actually interface with LLMs.

6 Conclusion

In this work, we introduce CAE, a benchmarking framework that aggregates a minimal set of benchmarks in order to efficiently generalize an agent's capabilities. Our approach prioritizes factual, falsifiable questions, such as "What is the height of the Eiffel Tower?" over more subjective prompts like "compose a beautiful haiku." We intend our focus on factuality to ensure reproducibility and enable objective, quantifiable evaluation metrics, with an eye towards consistent performance assessments.

Our target audience includes resource-constrained stakeholders, such as modeling managers at smaller companies or universities, who may lack access to extensive human evaluations, large-scale testing, or public ratings like those solicited in Chatbot Arena. By providing a lightweight evaluation approach, we enable such users to select models that align with their specific requirements in terms of quality and latency. Additionally, this framework serves as a guide for those just starting to work with LLMs, offering a practical tool for navigating trade-offs between different models. It is out of the scope of our paper to suggest specific directions for the open source community to push model development in, considering the thirteen models we profile, but decision makers could use frameworks like ours to make decisions like this.

In addition, we recognize that our framework has several limitations. First, the focus on multiple choice questions appears an idiosyncratic choice given how little they resemble the ways users actually engage with LLMs. While this limitation is mitigated by the strong correlation we see with Chatbot Arena, it still raises questions about the generalizability across use cases. Furthermore, our benchmark does not include any direct tests of linguistic skills or sentiment analysis.

In the future, we aim to extend this benchmark to cover multimodal tasks and more complex linguistic skills, such as text summarization. Additionally, we plan to incorporate dynamic, evolving benchmarks to mitigate the risks of dataset contamination, further improving the robustness and relevance of future evaluations.

References

William Arnold, Lucas Spangher, and Christina Rea. 2023. Continuous convolutional neural networks for disruption prediction in nuclear fusion plasmas. *ArXiv*, abs/2312.01286.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2023. Elo uncovered: Robustness and best practices in language model evaluation. *arXiv preprint arXiv:2311.17295*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *Preprint*, arXiv:2012.00614.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024a. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *Preprint*, arXiv:2404.04475.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2024b. Alpacafarm: A simulation framework for methods that learn from human feedback. *Preprint*, arXiv:2305.14387.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

David Ilić and Gilles E. Gignac. 2024. Evidence of interrelated cognitive-like capabilities in large language models: Indications of artificial general intelligence or achievement? *Intelligence*, 106:101858.

Doseok Jang, Lucas Spangher, Manan Khattar, Utkarsha Agwan, Selvaprabu Nadarajah, and Costas J. Spanos. 2021. Offline-online reinforcement learning for energy pricing in office demand response: lowering energy and data costs. *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*.

Doseok Jang, Lucas Spangher, Selvaprabu Nadarajah, and Costas J. Spanos. 2022. Deep reinforcement learning with planning guardrails for building energy demand response. *Energy and AI*.

Doseok Jang, Larry Yan, Lucas Spangher, and Costas J. Spanos. 2023. Active reinforcement learning for robust building control. *ArXiv*, abs/2312.10289.

Jihyung Kil, Zheda Mai, Justin Lee, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, Arpita Chowdhury, and Wei-Lun Chao. 2024. Compbench: A comparative reasoning benchmark for multimodal llms. *arXiv preprint arXiv:2407.16837*.

Alex Kipnis, Konstantinos Voudouris, Luca M Schulze Buschoff, and Eric Schulz. metabench-a sparse benchmark of reasoning and knowledge in large language models. In *The Thirteenth International Conference on Learning Representations*.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *Preprint*, arXiv:2406.11939.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Preprint*, arXiv:2211.09110.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2024. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*.

Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. *arXiv preprint arXiv:2405.12209*.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Qingwei Lin, Jianguang Lou, Shifeng Chen, Yansong Tang, and Weizhu Chen. 2024. Arena learning: Build data flywheel for llms post-training via simulated chatbot arena. *arXiv preprint arXiv:2407.10627*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinybenchmarks: evaluating llms with fewer examples. *Preprint*, arXiv:2402.14992.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Lucas Spangher, William Arnold, Alexander Spangher, Andrew Maris, and Christina Rea. 2023. Autoregressive transformers for disruption prediction in nuclear fusion plasmas.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *Preprint*, arXiv:2304.06364.

Kaijie Zhu, Jindong Wang, Qinlin Zhao, Ruochen Xu, and Xing Xie. 2024. Dynamic evaluation of large language models by meta probing agents. *Preprint*, arXiv:2402.14865.

Appendix

6.1 Cross Correlation matrix presented in (Ilić and Gignac, 2024)

Please see a cross correlation matrix between the main benchmarks included in Chatbot Arena 5. Please see a breakdown of the main subdomains.

6.2 Subdomains

Please see a breakdown of our hierarchy by subdomain.

6.3 Benchmark references

For a table of benchmark references, please see **??**.

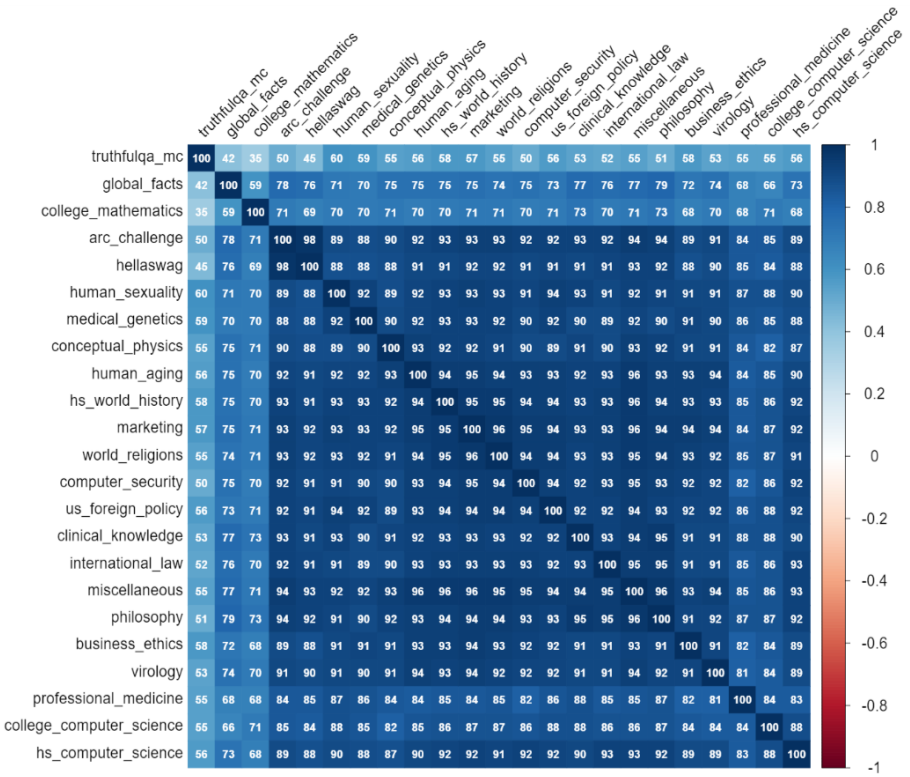

Fig. 3. Open LLM Leaderboard correlation matrix

Figure 5: Pairwise Correlations between benchmarks listed in Chatbot Arena.

Factuality	
TruthfulQA	https://github.com/sylinrl/TruthfulQA
Global Facts (MMLU Redux)	https://huggingface.co/datasets/edinburgh-dawg/mmlu-redux
Climate-FEVER	https://huggingface.co/datasets/tdiggelm/climate_fever
ARC-Challenge	https://huggingface.co/datasets/allenai/ai2_arc
BoolQ	https://huggingface.co/datasets/boolq
SQuAD	https://huggingface.co/datasets/rajpurkar/squad
Social Sensitivity and Linguistics	
BBQ Lite	https://github.com/google/BIG-bench/tree/main/bigbench
XSum (Summarization)	https://huggingface.co/datasets/EdinburghNLP/xsum
Problem Solving	
MMLU College Math	https://huggingface.co/datasets/cais/mmlu
MMLU College CompSci.	https://huggingface.co/datasets/cais/mmlu

Table 3: Benchmarks Used in the Evaluation

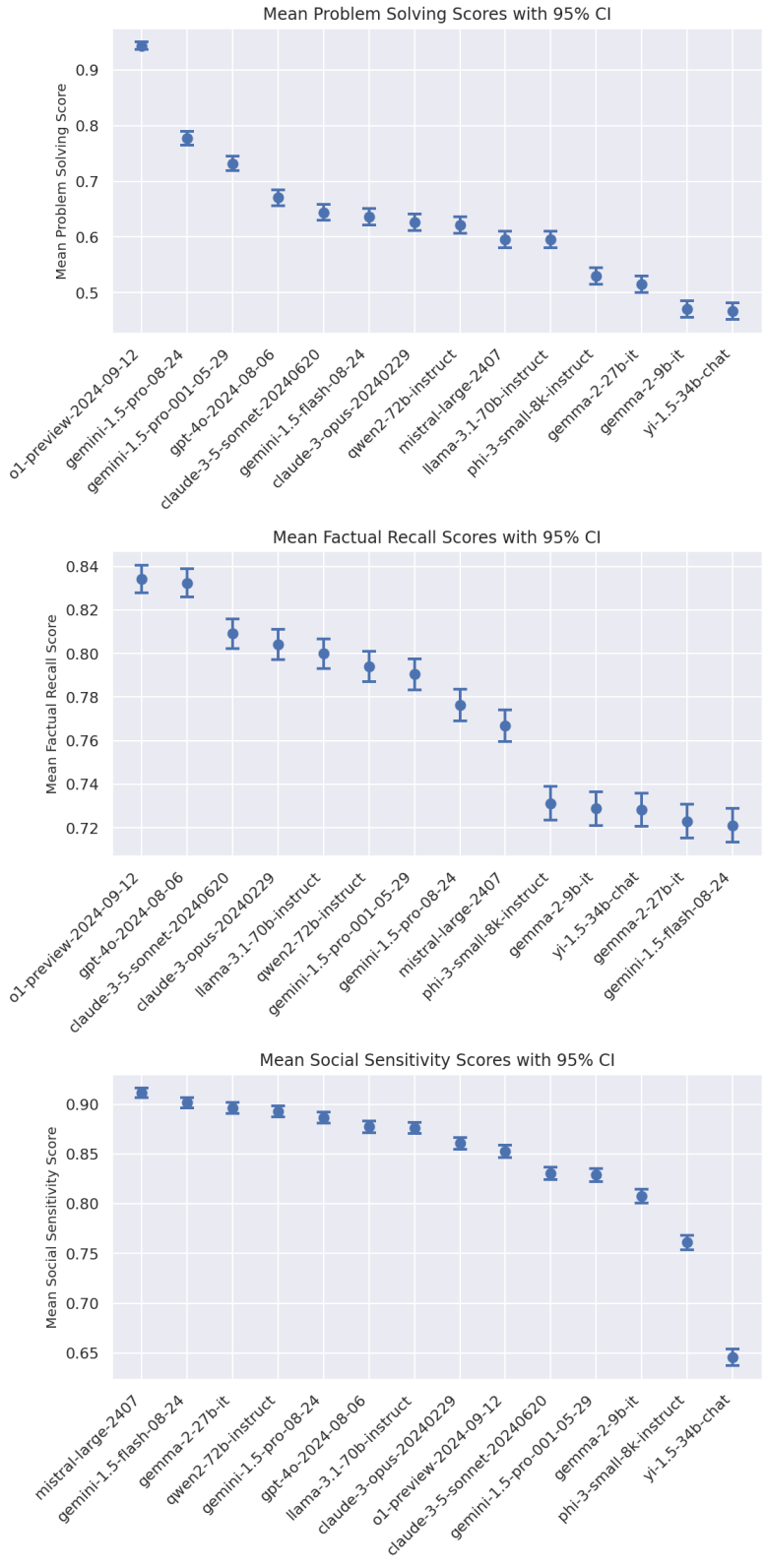

Figure 6: Orderings of the LLMs we studied.