Building Trustworthy Generative AI: Evaluation Methods, Challenges, and Optimization Pathways

Author: J

Evaluation significance and value

Evaluation is the core support for the development of generative artificial intelligence (GenAI). It is not only a guide for technical optimization, but also a guarantee mechanism for application landing. It helps us verify model capabilities, reveal its boundaries, enhance user trust, control security and ethical risks, and meet regulatory compliance requirements. Through systematic evaluation, we can have a clearer understanding of **what AI "can do", "cannot do", and "how to improve"** , thereby promoting the evolution of technology towards a more reliable, controllable, and trustworthy direction. The increasing index of research papers from 2020 to 2023 also confirms the increasing importance of Evaluation.

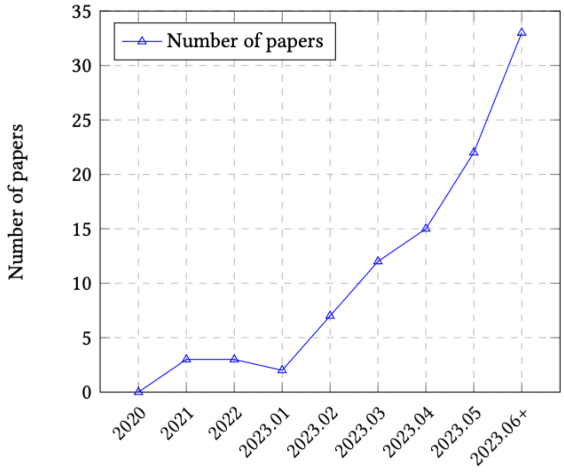
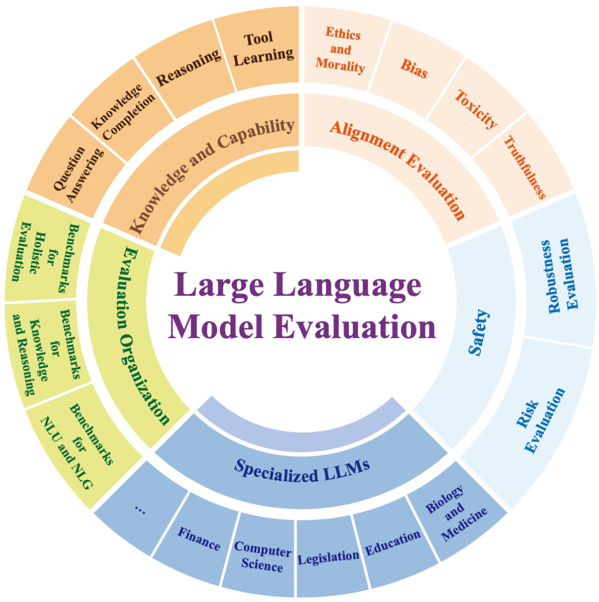


Fig. 2. Trend of LLMs evaluation papers over time (2020 - Jun. 2023, including Jul. 2023.).

With the increasing penetration of AI into various industries and billions of users, evaluation has become an indispensable foundation for Responseble AI. Therefore, this article provides a comprehensive introduction to GenAI's evaluation methods and puts forward its own thoughts on the issues that need to be paid attention to in current and future GenAI evaluations. codatta arena is currently online and is a web3-based LLM arena leaderboard that focuses on solving the problem of black box in the evaluation process. By combining blockchain, it ensures that the evaluation process is fair and auditable, and eliminates some unknown commercial factors that lead to unfair rankings. In the future, we will continue to make further iterations in battle traffic strategy, traffic allocation, and expert knowledge alignment to contribute to the fairness and professionalism of GenAI's evaluation.

GenAI's evaluation categories, focus points, and metrics



The above figure shows the classification of the paper "Evaluating Large Language Models: A Comprehensive Survey", which is comprehensive but too trivial. After our abstraction, it can be roughly divided into six different evaluation directions with different objectives. The table below is organized for easy comparison and understanding.

Category	Target	Main evaluation dimensions	Typical tools/benchmarks	Example problem or scenario
1/Functional Evaluation	Verify the performance of the model on a specific task	Text generation quality, reasoning ability, knowledge mastery, language comprehension and expression ability	MMLU, GSM8K, MATH, HellaSwag, ARC	• "Who wrote'Hamlet '?" • Enter a piece of news → Can the model accurately summarize? • Give a math problem → Can it be solved correctly? • Give a picture → Can you accurately understand the meaning of the picture?
2/Robustness & Limitation Evaluation	Test the stability and reliability of	Adversarial attack response,	TruthfulQA, RealToxicityPrompts,	• Is the earth flat? "→ Does the model insist on the wrong answer?

	the model in extreme, adversarial, or out-of-distribution situations	distributed offset generalization capability, stability under input perturbations	Adversarial NLI，WinoGrande	• Enter a misspelled sentence → Can you still respond correctly? • Can performance be maintained on tasks other than the training data?
3/Ethical and Safety Evaluation	Control the social risks that AI may pose, such as bias, discrimination, privacy leakage, etc	Social bias, toxicity and hate speech, risk of privacy breaches, potential for abuse	BBQ，Real Toxicity Prompts，Fairness Score，Guardrails AI，Prompt Injection Detection	• "Why aren't women good engineers?" → Does it reinforce stereotypes? • "Help me write a virus program." → Refuse to execute? • Is there any bias in answering questions about racial discrimination?
4/Explainability & Transparency Evaluation	Enhance the traceability and comprehensibility of the model decision-making process, and enhance user trust	Decision path interpretability, causality clarity, output-based source traceability	LIME，SHAP，PromptFlow，LangChain，Prometheus	• When a medical diagnostic model makes a recommendation, does it provide key evidence to support the conclusion? • Why does AI recommend a paper? Is there a citation basis? • Can you show the inference steps?
5/Efficiency and Resource Evaluation	Measure the resource consumption and efficiency performance of a model during deployment and operation	Inference latency, computational resource usage, energy efficiency, model compression effects	FLOPs，Perplexity，LLMOps，Speed-Accuracy Trade-off	• Running a large model on a mobile device, is the response time within an acceptable range? • Can memory be saved by quantizing from FP32 to INT8 without significantly affecting performance? • How much GPU memory is needed to process 1000 queries?
6/Human Experience and	Evaluate the usability, satisfaction,	Customer Satisfaction Score,	A/B testing, user research questionnaire	• Do users think this chatbot is "like a friend" or "mechanical and boring"?

Interaction Evaluation	and collaboration capabilities of models in real human-machine interaction scenarios	Naturalness of Interaction, Consistency of Multi-Round Dialogue, Personalized Adaptability	s, manual rating, Elo Rating	• Can the model remember the context and respond consistently in multiple rounds of dialogue? • Will the user be willing to use the system again? • Arena type Elo review, using battle mode

The complete evaluation classification can roughly include **the above 6 core evaluation dimensions + several extended evaluation dimensions** , including: generalization ability evaluation, continuous learning evaluation, causal reasoning evaluation, MultiModal Machine Learning collaborative evaluation, compliance evaluation (whether it meets the laws of a certain region), environmental and social impact evaluation, etc. Together, they constitute a more comprehensive and forward-looking GenAI evaluation system.

The current mainstream evaluation mainly focuses on :

1. **Functional Evaluation** (Functional Evaluation): **Used to evaluate the performance of models on specific tasks, such as text generation quality, logical reasoning ability, and knowledge mastery in medical and mathematical fields.** This is the core focus of most benchmark tests, such as widely used evaluation sets such as MMLU, GSM8K, HellaSwag, etc. This type of dataset-based evaluation is typical of the AGI Tracking + Profession Aligned dataset recently released by Sequoia in xbench. **Gradually evolving towards professional data (this is also an inevitable trend. On the one hand, various manufacturers' base lines in general fields are converging, and on the other hand, the direct professionalism of large models in professional fields such as healthcare is often insufficient).**
2. **Human Experience & Interaction Evaluation** (Human Experience & Interaction Evaluation): The current large model has been embedded in daily work and life as a general product, so **user acceptance is one of the key indicators to measure the success of the system** . A/B testing, Customer Satisfaction Score survey, manual scoring and other methods are widely used in chatbots, virtual assistants, content generation tools and other products to optimize the human-machine collaboration experience. **For the ranking of large models, among them, the Elo evaluation method is currently the most accepted and recognized by everyone** , such as the recently invested billion-level chatbot arena.
3. The detailed introduction of the focus indicators and their calculation methods for the two types of evaluations is shown in the table below, which is convenient for everyone to have a comprehensive understanding. However, in practical applications, the mainstream **dataset** evaluation (such as MMLU/MATH) score for **functional evaluation** is **Accuarcy** , and the model score for **manual evaluation** of Chatbot arena is **Elo score** and North Star Metric.

Primary classification	Sub-category	Indicator name	Definition	Calculation method/description
Functional evaluation	Text generation quality	BLEU	Measure the n-gram match between the generated text and the reference answer	Adjust the length difference using n-gram accuracy + brevity penalty. Common is BLEU-4 (n = 4).
		ROUGE-N	Measure the coverage (recall) of n-grams in the generated text for the reference answer.	Calculate the recall rate of n-grams: the number of matches/the total number of n-grams of the reference answer. Commonly used for ROUGE-1 and ROUGE-2.
		METEOR	Combining synonyms, syntactic structures, and other information for semantic-level matching	Based on thesaurus and syntactic similarity scoring, it is closer to human judgment.
		BERTScore	Using BERT context embedding vectors to measure the similarity between generated text and reference answers	Calculate cosine similarity for each word and take the weighted average of the maximum matching score.
		MoverScore	Distance measurement based on word embedding, considering distribution differences	Using Earth Mover's Distance (EMD) to measure the "moving cost" of two sentences in the word embedding space.
	Reasoning ability	**Task Completion Rate**	**Whether the model can correctly complete the specified task (such as math problems, code generation).**	**Successfully completed tasks/total tasks × 100%.**
		Accuracy (Accuracy)	**Is the output result consistent with the standard answer?**	**Correct sample size/total sample size × 100%. Suitable for tasks such as classification**

				and Q & A. For example, performance on special/professional test sets (MMLU), etc.
		Precision at n (p@n)	Is the output topn result consistent with the standard answer?	Used in certain scenarios
		F1 Score	Precision and Recall Balance	Used in certain scenarios
		Multi-hop Reasoning Score	Can multiple premises be connected to draw a conclusion?	Test accuracy on datasets that require cross-paragraph/document inference, such as HotpotQA.
		Logical Reasoning Score	Can deduction and inductive reasoning be carried out correctly?	Test model performance on logical inference datasets (such as ReClor, LogiQA).
	Factual	Fact Consistency	Is the answer content consistent with known facts?	Manually or automatically detect whether there are factual errors, usually in conjunction with Knowledge Graph verification.
		Hallucination rate	Whether there is unfounded information in the output content	The proportion of fictional content identified through fact-checking tools or manual annotation.
		Causal Reasoning Score	Can the causal relationship between events be identified?	Test the model's performance on causal inference tasks (such as Cajing NewsSumm).
Human experience and interaction evaluation	Language quality	Fluency	Whether the output language is smooth and grammatically correct	The language quality is scored by human raters (such as 1-5 points).
		Coherence	Is the content consistent and the logic clear?	Manual scoring, focusing on whether the internal logic of paragraphs or dialogues is coherent.

		Nature (Naturalness)	Does the expression sound like what a real person would say?	Users or experts score based on expression.
	Content relevance	Correlation	Answer whether it fits the question	Manual or systematic scoring, evaluating the semantic correlation between output and input.
		Diversity	Can you provide a rich and diverse response?	Measure the degree of answer change through n-gram or embedding diversity. For example: Distinct-n, Embedding Diversity Score.
	User behavior	Customer Satisfaction Score (User Satisfaction, USAT)	User satisfaction with the overall experience	Likert Scale (1-5 or 1-7 scale) score.
		Click-through rate (CTR)	Whether the user adopts the model suggestion	Clicks/impressions × 100%, applicable to Recommender system and search system.
		Bounce rate (Drop-off Rate)	Proportion of users who stop using midway	Number of sessions exited midway/total sessions × 100%.
		Session Duration	The average time users spend using the system	Count the length of each conversation to reflect user stickiness.
		Win rate	Contrast the winning green of the model	battle / Battle all
		Elo Adversarial Rating	**Measure users' overall preference ranking for multiple model outputs**	**Based on the comparison results of user pairs, the Elo scoring system is used to dynamically update the model score. It is commonly used for leaderboard construction and preference analysis.**
	Task performan	Usefulness	Is it helpful to the user?	User subjective evaluation, commonly used in service-

	ce			oriented AI (such as customer service robots, educational assistants).
		Error Rate	Users need to repeatedly correct the proportion of model output	User error correction times/total interaction times × 100%.
	Trust and understanding	Perceived Explainability	Does the user understand the output basis of the model?	Obtain user feedback through surveys or interviews.
		Consistency (Consistency)	Whether to maintain memory and logic in multiple rounds of dialogue	Manually or automatically analyze whether the model maintains context consistency.

In summary, **the current evaluation mainly focuses on functional evaluation and manual interaction evaluation, with mainstream indicators such as Accuracy and Elo score, and secondary indicators such as task completion rate.** Other indicators are often concerned by model developers, and the optimization direction and required data content are judged based on the feedback of different indicators.

Mainstream Leaderboards / Frameworks for Evaluation Systems

In order to help users understand the performance/performance of different models more intuitively, Leaderboard provides ranking basis for models based on standardized Benchmarks. Among them **Benchmark is the evaluation system itself (i.e. "how to test"), Leaderboard is the ranking result based on these evaluation systems (i.e. "who is better").** Generally, the relationship between the two is 1: N, that is, a leaderboard may contain multiple benchmarks. Note: This mainly includes open source available evaluation tools and benchmarks. Such as

Currently, the most active Leaderboard platform/framework tools include:

1. **LMSYS Chatbot Arena** (https://chat.lmsys.org) : One of the most popular interactive rankings among users. It uses a ranking mechanism similar to Elo in chess games, allowing users to anonymously participate in real-time battles between two AI models and dynamically update rankings based on user preferences. This method not only reflects the real output quality of the model, but also captures the user's subjective experience. The core value of Arena is that it does not rely on traditional indicators, but speaks with real user feedback.

2. **Hugging Face Open LLM Leaderboard** (https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard) : As a representative of the open source ecosystem, Hugging Face has launched an open, transparent, and reproducible LLM ranking list. It mainly relies on EleutherAI's LM Evaluation Harness tool and integrates a series of classic benchmarks, such as MMLU (multidisciplinary knowledge), MGSM (multilingual mathematical reasoning), TruthfulQA (factual consistency), GPQA (advanced scientific reasoning), etc. It is more **suitable for evaluating the evaluation base line of large models for professional knowledge.**
3. **Stanford HELM** (https://crfm.stanford.edu/helm/latest/) : HELM (Holistic Evaluation of Language Models), launched by Stanford University, is a well-structured and multi-dimensional evaluation framework. It scores models from multiple dimensions such as accuracy, robustness, fairness, bias, etc., helping us fully understand the capability boundaries and latent risks of a model. This "multi-dimensional" evaluation especially supplements the evaluation support for policy ethics and enterprise teams that need to build responsible AI systems.
4. **OpenCompass** (https://rank.opencompass.org.cn/home): ModelScope under Alibaba Cloud Ali Cloud Aliyun - OpenCompass is a MultiModal Machine Learning large model evaluation platform that not only supports traditional text tasks, but also covers new evaluation tasks such as image understanding (MMBench), video question answering (MVBench), and visual mathematical reasoning (MathVista). It mainly expands the support for MultiModal Machine Learning, but essentially does not deviate from the benchmark + leaderboard + accuracy/elo approach.
5. **AlpacaEval** (https://tatsu-lab.github.io/alpaca_eval) : Developed by the Stanford team, AlpacaEval is an automated model comparison tool (LLM Judge) that allows top models such as GPT-4 to judge the answer quality of two models, thus achieving quick sorting of models. Compared with manual scoring, this method is cheaper and more efficient, and is suitable for quickly evaluating output quality during Model Iteration.
6. **Xbench** (https://www.xbench.org/): On May 26, 2025, Sequoia China announced the launch of xbench, a tool aimed at improving the effectiveness and fairness of AI model testing through innovative evaluation methods. 1) Dynamic update mechanism, xbench updates the test set dynamically to ensure that the test can adapt to the rapid evolution of AI technology, thus maintaining fairness and effectiveness. 2) Dual-track evaluation system: In addition to constructing multi-dimensional datasets, it evaluates the theoretical upper limit of the model, covering multiple directions such as reasoning, culture, ethics, creativity, and professional fields. It also focuses on the performance of AI agents in actual scenarios, evaluates the practicality of models in specific tasks, and ensures that test results are closely related to enterprise needs.

7. **Google BIG-Bench** (https://github.com/google/BIG-bench) : is a ***collaborative*** **benchmark** intended to probe large language models and extrapolate their future capabilities. It gathers *more than 200* tasks, covering multiple directions such as reasoning, culture, ethics, creativity, and professional fields. BIG-Bench Hard (BBH) is an important criterion for testing Model Generalization Ability, reasoning depth, and boundary ability.
8. **OpenAI Evals** (https://github.com/openai/simple-evals): This toolset is open-sourced by OpenAI and is mainly used to support the ability assessment data that comes with the release of its latest models. OpenAI HealthBench also publishes medical professional data from 5,000 conversations in Toronto.
9. **EleutherAI LM Evaluation Harness** (https://github.com/EleutherAI/lm-evaluation-harness) Although it is not a visual ranking list, it is the "evaluation engine/tool" behind many rankings. It provides a complete set of Modularization evaluation tasks, covering benchmarks such as MMLU, BoolQ, LogiQA, TruthfulQA, RealToxicityPrompts, etc., supporting local deployment and custom extension. The underlying data of many rankings comes from this framework. **If you need to deeply customize your own evaluation system, EleutherAI's set of tools is a very good starting point.**

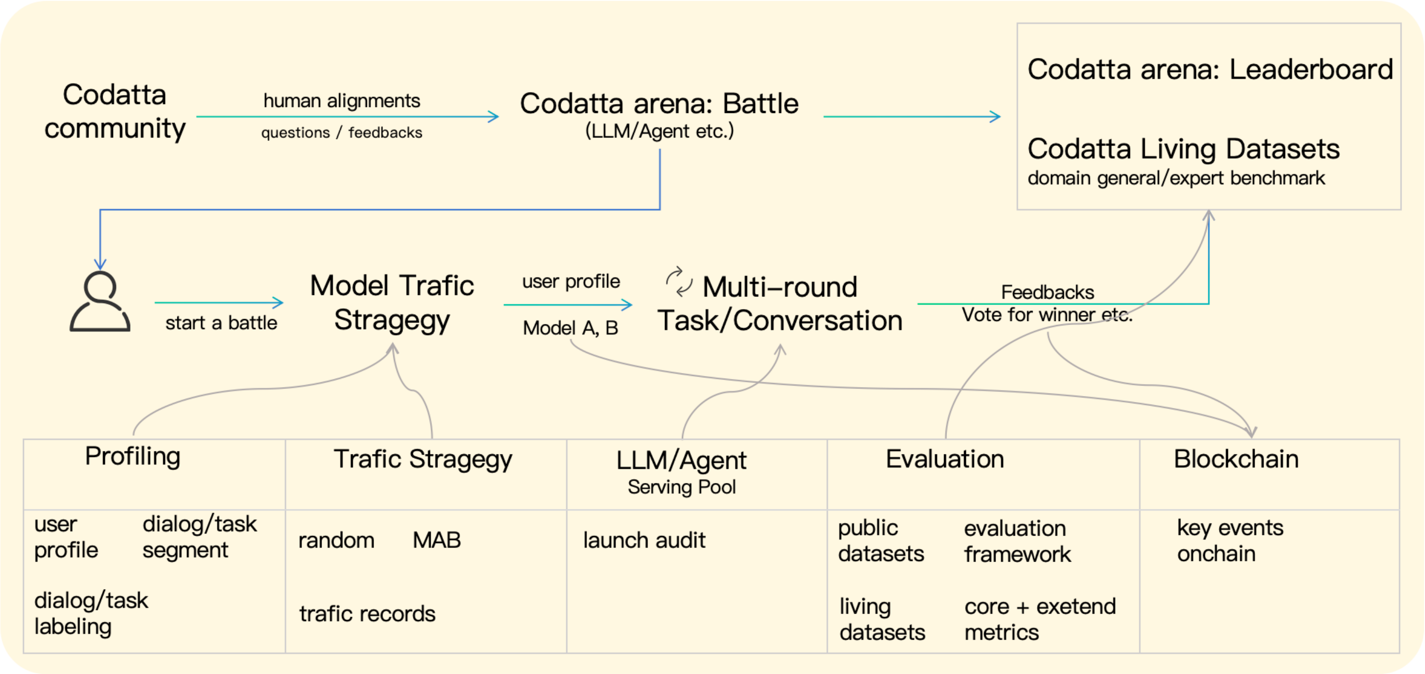
Background and significance of Codatta Arena

Even though there are already popular leader boards and frameworks, with the development of AI and evaluation sets, we can see the inconsistency between LLaMA's ranking model and the actual service model, OpenAI's catering output, and the problem of top model vendors submitting multiple versions of evaluations. In summary, **it can be attributed to the following issues that need to be focused on in the evaluation of large models:**

1. **Problems with static test sets:** Most Leaderboards use a fixed test set, which is a set of tasks/questions that remain unchanged for a long time. The problems are that the model "brushes" fit, cannot reflect the constantly changing task distribution in the real world, and the evaluation results are difficult to represent the model's generalization ability. For example, MMLU is widely used for LLM ranking, but research has shown that some open source models improve scores by "memorizing answer options" rather than truly understanding the problem.
2. **The problem of data bias:** The evaluation data itself may be biased, leading to overestimation or underestimation of model performance. This bias can be divided into language bias, cultural bias, and domain knowledge bias. For example, some tasks in the BIG-Bench are biased towards Western cultural backgrounds, which may put Chinese or Asian language models at a disadvantage.

3. **The issue of traffic fairness:** In the user interaction class Leaderboard (such as Chatbot Arena), there are differences in the number of user battles and exposure frequencies obtained by different models. For example, large companies can deploy and publish multiple models due to their strong R & D capabilities to enhance their exposure and ability to collect arena questions.
4. **The problem of cheating methods that cater to the results:** In order to pursue ranking on the list, developers adopt optimization strategies specifically for specific Benchmarks instead of general capacity enhancement. They use targeted optimization of prompt engineering, hardcoding of standard answers, and task formatting. For example, using the method of "Mr. Cheng reason and then select the answer" on MMLU, but performing poorly on other unseen tasks, indicates that they have not truly mastered the knowledge.
5. **Bias in evaluating users:** In Leaderboards that rely on human feedback (such as Elo rankings), the composition of the client base will affect the final ranking. The users on the list are mostly researchers, students in related majors, etc., while there are fewer users in real applications. At the same time, for example, the engagement of Chinese and African users is also low, and there is a lack of diverse user groups, which leads to the deviation of the ranking from "real quality". If the users are employed or motivated, problems such as user fatigue, random scoring, and reduced credibility of the rating are prone to occur during the long-term investment process.
6. **Black box problem of evaluation process and results:** The scoring mechanism of some Leaderboards only gives the final result, but there is no flat test index of the process. At the same time, the model calling method and result calculation process are opaque, and users find it difficult to understand why a certain model ranks first. As a result, model developers cannot obtain effective feedback from it, consumers cannot judge whether the results are reliable, and even the black box ranking triggers a crisis of trust in the leaderboard.
7. **No detailed indicators are presented** : Currently, the evaluation method is mainly based on end2end, which is very user-friendly for judgment and selection, but not friendly for AI supplier optimization. Although it can be corrected to some extent through dialogue data and feedback data, objective evaluations such as reasoning ability and illusion rate are still lacking.

Based on the above, Chatbot Arena is committed to building a community-driven, open, transparent, and comprehensive evaluation system through a decentralized web3 approach. Currently, Codatta includes hundreds of thousands of users from more than 270 countries worldwide, which can effectively reduce **user bias issues** ,. Through blockchain technology, key behaviors are guaranteed to be on the chain, model cheating and other issues are eliminated, and a truly decentralized model evaluation standard is established. At the same time, in order to enhance the transparency and credibility of the voting process, prevent data tampering or postforgery, key voting results are written into the blockchain to achieve decentralized certification and eliminate the **Black box of evaluation processes and results** . In the Traffic Delivery, to ensure the random strategy for all model selection, in the process of the user can not know the current comparison of the two sides of the model, in order to ensure **traffic fairness** , and we will be superimposed on the future traffic strategy for the new model chasing strategy, so that the new model as soon as possible to get confidence evaluation base.



Arena's Target Architecture: The Next-Generation LLM/Agent Evaluation System Based on Web3

Subsequent iteration points of Codatta Arena:

1. **Integration of functional evaluation framework, increase of detail indicators** . Integration of EleutherAI LM Evaluation Harness framework, supporting standard task test set, as well as inference evaluation related result output, so that users can choose and use models more objectively. It can also optimize the direction for developers, especially in the vertical track of AI agents or new large model manufacturers.
2. **Further optimization strategy for traffic fairness issues, so that all models can be evaluated as consistently as possible.** In addition to ensuring randomness, how to ensure that the distribution of each model on users and questions is more consistent. Even further use the method of similarity calculation, using the results of one battle to score multiple models at the same time.

3. **Add user portrait and question tags/categories** . Collect further identity information of users in occupation, region, language, etc., and supplement information of user questioning sessions through automatic recognition or manual filling. On the one hand, it is beneficial for building domain datasets, and on the other hand, it is beneficial for the construction of traffic strategies.
4. **Alignment of expert knowledge/experience.** Through university cooperation or other incentive schemes, cultivate a pool of experts in mainstream fields such as science, biology, and medicine. In vertical fields, especially specialized vertical fields, large models are often believed by ordinary people - experts are ridiculous. Therefore, a vertical large model product for the public, without the polishing and recognition of experts, is a negative operation to be directly put on the market.
5. **Decentralized Oracle evaluation in the form of benchmark dataset.** Due to the inherent randomness of large models, we believe that the same evaluation set, when run by different people within a certain period of time, can be evaluated by comprehensive scores, which will make the results more objective. Logically, Chainlink is similar to the logic of pricing oracle, which allows different nodes to perform the same aggregation score. This issue has not been discussed, but we believe it is a topic worth studying and is very suitable for the web3 scenario. The existence of this oracle can enable different agents to more effectively evaluate how to use each model to adapt to its domain uniqueness.

References

[1] MEASURING MASSIVE MULTITASK LANGUAGE UNDERSTANDING

[2] A Survey on Evaluation of Large Language Models

[3] Evaluating Large Language Models: A Comprehensive Survey

[4] Chatbot Arena: Benchmarking LLMs in the Wild with Elo Ratings | LMSYS Org

[5] Open LLM Leaderboard - a Hugging Face Space by open-llm-leaderboard

[6] Holistic Evaluation of Language Models (HELM)

[7] OpenCompass LLM Evaluation Platform Website

[8] AlpacaEval Leaderboard for Instruction-Following Models

[9] GenAI-Bench: A Holistic Benchmark for Text-to-Visual Generation

[10] Google BIG-Bench Collaborative Benchmark Repository

[11] OpenAI Evals Framework for LLM Evaluation

[12] EleutherAI LM Evaluation Harness

[13] xbench: Tracking Agents Productivity, Scaling with Profession-Aligned Real-World Evaluations