Available Online: 3 September 2026 DOI: 10.54254/2977-3903/2026.36593

A statistical framework for reliability evaluation of large language models

Yi Zhu

School of Economics, Hangzhou Dianzi University, Hangzhou, China

13067835506@163.com

Abstract. The growing use of Large Language Models (LLMs) in real-world applications has increased the need for reliable evaluation methods. This study proposes a statistical framework for evaluating LLM reliability by jointly considering model accuracy, hallucination rate, and cross-domain performance stability. Based on the TruthfulQA dataset, 300 test questions were selected using a stratified sampling strategy, and responses were generated by DeepSeek-V4-Flash and GPT-4o-mini. GPT-4o was then employed as an automated evaluator to assess the 600 generated responses. Model reliability was evaluated using three dimensions: Accuracy, Hallucination Rate, and Cross-Domain Stability. In particular, the proposed Cross-Domain Stability (CDS) metric quantifies the consistency of model performance across different knowledge categories. A Weighted Geometric Mean was further employed to construct an overall reliability score. The experimental results show that the proposed framework can identify reliability differences between different LLMs. DeepSeek-V4-Flash achieved an overall reliability score of 0.770, compared with 0.678 for GPT-4o-mini. Sensitivity analysis further shows that the model ranking remains unchanged under different weight configurations, indicating that the proposed framework is reasonably robust to variations in indicator weights. This study provides a statistical framework for quantitatively assessing LLM reliability under multiple evaluation criteria.

Keywords: large language models, reliability evaluation, hallucination, cross-domain stability, statistical framework

1. Introduction

The increasing use of LLMs in real-world applications has raised new challenges for reliability evaluation. Models such as GPT and DeepSeek are increasingly applied to knowledge question answering, intelligent assistance, and domain-specific tasks. However, reliability challenges remain, particularly hallucination, where models generate plausible but factually unsupported information. Moreover, performance variations across knowledge domains make it difficult for a single metric to provide a comprehensive reliability assessment.

Existing evaluation studies have considered various aspects of LLM performance, including accuracy, factuality, robustness, and safety. However, integrating these reliability-related indicators into a unified statistical framework remains challenging.

To address these limitations, this study proposes a statistical framework for LLM reliability evaluation. Based on the TruthfulQA dataset, responses generated by different LLMs are evaluated by GPT-4o along three dimensions: Accuracy, Hallucination Rate, and Cross-Domain Stability. A Weighted Geometric Mean is further employed to construct an overall reliability score. The proposed framework provides a quantitative approach for comparing LLM reliability from multiple perspectives, which may support more informed model selection in practical applications and provide a reference for future research on reliability evaluation methods.

2. Literature review

2.1. Evaluation benchmarks for large language models

Existing LLM evaluation benchmarks have been developed to assess model capabilities across a wide range of tasks, domains, and dimensions. Hendrycks et al. proposed Massive Multitask Language Understanding (MMLU), which evaluates language models' knowledge and reasoning abilities across diverse academic and professional subjects, providing a standardized benchmark for general capability assessment [1]. BIG-bench further expanded LLM evaluation by introducing challenging tasks across multiple capability dimensions, allowing researchers to examine broader capabilities and limitations of language models [2].

Beyond task-oriented evaluation, Liang et al. proposed Holistic Evaluation of Language Models (HELM), which assesses LLMs across multiple dimensions, including accuracy, robustness, fairness, and safety [3]. HELM demonstrates the transition from single-metric evaluation toward multidimensional assessment frameworks.

Factual reliability has also become an important aspect of LLM evaluation. Since LLMs may generate fluent but unsupported responses, Lin et al. introduced TruthfulQA to evaluate whether models can avoid false information when answering questions associated with common misconceptions [4]. The benchmark provides an important basis for factual reliability assessment.

These benchmarks have provided standardized approaches for evaluating LLM capabilities across different tasks and domains. However, existing benchmarks often report performance through separate metrics, while the integration of factual reliability and cross-domain consistency into a unified evaluation framework remains less explored.

2.2. Hallucination and factual reliability evaluation

Hallucination has become a critical issue in LLM reliability evaluation, as models may generate fluent but factually unsupported information. Ji et al. systematically reviewed hallucination in natural language generation and noted that generated content may be inconsistent with factual knowledge or input context, highlighting the importance of hallucination detection and mitigation in generative AI [5]. To systematically evaluate hallucination, Li et al. proposed HaluEval, a benchmark containing human-annotated and model-generated hallucination samples for evaluating hallucination detection in LLM outputs [6]. Beyond benchmark-based methods, Min et al. proposed FActScore, which decomposes generated text into atomic facts and evaluates factual support at a finer granularity [7].

For truthfulness evaluation, TruthfulQA assesses whether language models can avoid false information when answering questions associated with common misconceptions. However, it mainly evaluates truthfulness at the question level and summarizes performance through overall accuracy, which may not capture variations across knowledge domains.

Beyond factuality, output consistency is another aspect of LLM reliability. Raj et al. investigated output stability under semantically similar inputs and showed that models may produce different responses when

input formulations vary [8].

3. Statistical framework for reliability evaluation

3.1. Dataset and sampling strategy

3.1.1. TruthfulQA dataset

The TruthfulQA dataset was used as the data source for the experiments. TruthfulQA is a benchmark designed to evaluate the truthfulness of large language models. Unlike traditional question-answering datasets, which primarily focus on task completion, TruthfulQA places greater emphasis on whether models generate responses that appear plausible but lack factual support. It therefore provides a suitable basis for examining the reliability of LLMs in practical applications.

The dataset covers a range of real-world knowledge domains that are susceptible to misconceptions, misinformation, or social biases, including health, law, history, science, sociology, and statistics, among others. This multi-domain structure provides the empirical basis for the Cross-Domain Stability indicator proposed in this study, allowing model reliability to be examined not only in terms of overall performance but also in terms of performance consistency across different knowledge categories.

3.1.2. Category-based stratified sampling

The number of questions varies substantially across categories in the TruthfulQA dataset. If simple random sampling were applied to the dataset as a whole, the proportions of some categories in the sample might deviate from those in the original dataset, potentially resulting in insufficient representation of low-frequency categories. This could affect the analysis of model reliability across different domains. To approximately preserve the category structure of the original dataset while reducing the potential impact of category imbalance on the evaluation results, this study adopts a category-based stratified sampling method. Suppose that the TruthfulQA dataset contains K categories. Let the number of samples in the k-th category be N_k, and let the total number of samples in the dataset be N:

$$N = \sum_{k=1}^{K} N_k \tag{1}$$

Given a target sample size of n, the theoretical number of samples allocated to the k-th category is defined as follows:

$$n_k = n \times \frac{N_k}{N} \tag{2}$$

However, simply rounding the proportional allocation may lead to two problems.

First, because the sample sizes of different categories vary considerably, the proportional allocation may assign fewer than one sample to some low-frequency categories, causing them to be excluded from the final sample.

Second, the subsequent analysis requires the calculation of category-level performance stability. If some categories contain no samples or too few samples, the model performance within those categories cannot be reliably estimated.

Therefore, a minimum sample size constraint was further imposed on the category allocation:

$$n_k \geq 2 \tag{3}$$

This constraint ensures that each category contains at least two samples.

After applying the minimum sample size constraint, the allocation of samples across categories is further adjusted to account for rounding differences while maintaining the predefined total sample size.

The resulting stratified sample preserves the overall category structure of the original TruthfulQA dataset while ensuring that low-frequency categories are adequately represented for subsequent evaluation.

3.2. LLM response generation and evaluation framework

This study adopts a two-stage evaluation procedure consisting of a Response Generation Stage and an Automatic Evaluation Stage [9].

In the first stage, the target LLMs generate responses to the benchmark questions. The generated responses are then automatically evaluated by a different Evaluator LLM, producing structured evaluation results for subsequent reliability metric calculation. This framework reduces the cost associated with relying entirely on human evaluation while enabling the quantitative assessment of factual accuracy and hallucination risk.

3.2.1. Response generation process

In the response generation stage, the input consists of the set of test questions obtained through stratified sampling.

For each question, the target model generates a response using a standardized prompt template. To reduce the influence of model randomness on the experimental results, fixed generation parameters are used throughout the response generation process.

Specifically, the models are instructed to:

Provide concise and factual answers;

Avoid unnecessary elaboration;

Explicitly state uncertainty when uncertain.

This design serves two main purposes. First, it reduces potential evaluation bias caused by differences in response length and wording across models. Second, it keeps the generated responses focused on factual information, thereby facilitating subsequent automatic evaluation.

Following this process, each test question is paired with a model-generated response:

$$R_i = f(Q_i, P, M, \theta) \tag{4}$$

where Q_i denotes the i-th test question, P represents the standardized prompt template, M represents the target language model, and θ denotes the generation settings, including temperature and response length constraints. R_i represents the response generated by the target model for the corresponding question.

The resulting response dataset can be represented as:

$$D = \{(Q_i, R_i)\}_{i=1}^{n} \tag{5}$$

where D denotes the generated response dataset containing n question-response pairs.

3.2.2. Automatic evaluation using LLM-as-a-judge

Due to the high time cost associated with large-scale human evaluation and the potential subjectivity among different evaluators, this study adopts an automatic evaluation approach based on large language models to assess the generated responses.

This approach employs an independent large language model as an evaluator, which automatically assesses the quality of generated responses based on the question content, reference answers, and responses produced by the target models.

For each generated response, the evaluator model receives the following information:

Question

Best Answer

Correct Answers

Incorrect Answers

Response

The evaluation process can be represented as follows:

$$E_i = f(Q_i, R_i, A_i^+, A_i^-) \tag{6}$$

where A_i^+ denotes the set of correct reference answers, A_i^- denotes the set of incorrect reference answers, and E_i represents the evaluation result.

To facilitate subsequent statistical analysis, the evaluation results are converted into structured labels, including Correctness and Hallucination.

Correctness Label:

$$\text{Correctness} = \begin{cases} 1, & \text{if the response is factually correct} \\ 0, & \text{otherwise} \end{cases}$$

Hallucination Label:

$$\text{Hallucination} = \begin{cases} 1, & \text{if the response contains factual hallucinations} \\ 0, & \text{otherwise} \end{cases}$$

Finally, each model response is converted into a structured data record: $(Q_i, R_i, Correctness_i, Hallucination_i)$.

These structured records provide the basis for calculating Accuracy, Hallucination Rate, and Cross-Domain Stability in the subsequent analysis.

3.3. Construction of reliability metrics

This study evaluates LLM reliability from three perspectives: response correctness, factual reliability, and cross-domain performance stability. Accordingly, three core metrics are introduced: Accuracy, Hallucination Rate, and Cross-Domain Stability (CDS).

A composite reliability score is then constructed by integrating these metrics.

3.3.1. Accuracy

Accuracy is used to measure the factual correctness of the responses generated by the model.

Based on the Correctness Label, each response is assigned a binary label: $Correctness_i$:

Accordingly, the overall accuracy of the model is defined as:

$$Accuracy = \frac{\sum_{i=1}^{n} Correctness_i}{n} \tag{7}$$

where n denotes the number of test samples, and $Correctness_i$ indicates whether the response to the i-th question is correct.

A higher Accuracy indicates stronger factual answering ability of the model on the test dataset.

3.3.2. Hallucination rate

Hallucination Rate is used to measure the risk of factual hallucinations generated by the model.

Based on the Hallucination Label generated during the automatic evaluation stage, the model's hallucination rate is defined as:

$$HallucinationRate = \frac{\sum_{i=1}^{n} Hallucination_i}{n} \tag{8}$$

Where $Hallucination_i$ indicates whether the i-th response contains a factual hallucination.

A lower Hallucination Rate indicates a lower likelihood of the model generating factually incorrect information and, therefore, higher model reliability.

Because Hallucination Rate has the opposite direction from reliability, when calculating the composite reliability score, it is transformed as $1 - HallucinationRate$.

This transformation ensures that all metrics are oriented in the same direction, such that a higher value consistently represents higher reliability.

3.3.3. Cross-domain stability

Although Accuracy and Hallucination Rate can reflect the overall performance of a model, they cannot fully capture performance differences across different knowledge domains.

For example, a model may achieve relatively high accuracy on health-related questions but perform poorly in the law or history domains. If only overall Accuracy is considered, such disparities across domains may be overlooked.

Therefore, this study proposes the Cross-Domain Stability (CDS) metric to measure the stability of model performance across different knowledge domains.

First, the accuracy of the model within each Category is calculated as:

$$A_k = \frac{C_k}{N_k} \tag{9}$$

The weight of each Category is defined as:

$$\omega_k = \frac{N_k}{N} \tag{10}$$

The weighted mean accuracy across all Categories is then calculated as:

$$\overline{A} = \sum_{k=1}^{K} \omega_k A_k \tag{11}$$

Where C_k denotes the number of correctly answered questions in the k-th Category;

Next, the weighted standard deviation of accuracy across all Categories is calculated as:

$$\sigma_A = \sqrt{\sum_{k=1}^{K} \omega_k (A_k - \overline{A})^2} \tag{12}$$

A larger standard deviation indicates greater variation in model performance across different knowledge domains.

Therefore, the Cross-Domain Stability (CDS) metric is defined as:

$$CDS = 1 - \sigma_A \tag{13}$$

A higher CDS value indicates that the model achieves more balanced performance across different knowledge domains, suggesting stronger cross-domain reliability.

3.3.4. Reliability score

To integrate multiple dimensions of model reliability, this study further constructs a comprehensive reliability score.

The score is defined as follows:

$$Reliability\ Score = Accuracy^{\omega_1} \times (1 - HallucinationRate)^{\omega_2} \times CDS^{\omega_3} \tag{14}$$

In this study, ($\omega_1, \omega_2, \omega_3$) = (0.4, 0.3, 0.3) is selected as the baseline weight configuration.

This configuration emphasizes the importance of answer correctness while simultaneously considering factual risks and cross-domain stability.

Compared with linear weighted aggregation, the geometric mean reduces the possibility that a strong performance in one dimension compensates for insufficient performance in other dimensions. Therefore, it is

more suitable for reliability evaluation scenarios that require balanced performance across multiple dimensions [10].

Furthermore, to examine whether the comprehensive score depends on a specific weight configuration, a sensitivity analysis is conducted in the experimental section. Different weight combinations are tested to observe changes in model rankings.

4. Experiments and results

4.1. Experimental setup

4.1.1. Experimental dataset

Based on the dataset selection and stratified sampling strategy described in Section 3, this study constructs the experimental test set from TruthfulQA. The final test set contains 300 questions covering 37 knowledge domains, with each question associated with a category label for subsequent category-level analysis.

4.1.2. Model selection and response generation

To evaluate the proposed reliability framework, two efficient large language models, DeepSeek-V4-Flash and GPT-4o-mini, were selected as target models.

Each model generated one response for every question in the test set, resulting in 600 responses in total. To minimize the impact of response length and stylistic differences on evaluation results, a unified generation configuration was applied during the response generation process. The detailed generation settings are summarized in Table 1.

Table 1. Response generation settings

parameter	setting
Temperature	0
Maximum response length	Short factual responses only
Answer format	One- to two-sentence factual responses

This design reduces noise caused by differences in response length and style, allowing the evaluation to better reflect factual reliability rather than generation style.

4.1.3. Answer evaluation procedure

After response generation, GPT-4o was employed as the evaluator model to automatically assess all 600 generated responses. Following the evaluation framework described in Section 3.2.2, the evaluator produced response-level labels for subsequent calculation of Accuracy and Hallucination Rate.

4.2. Overall reliability evaluation

Based on the reliability evaluation framework introduced in Section 3, the performance of the two models is first compared under the proposed evaluation scheme.

The comparison results are presented in Figure 1.

As shown in Table 2, DeepSeek-V4-Flash achieved a higher overall reliability score (0.770) than GPT-4o-mini (0.678). Moreover, DeepSeek-V4-Flash outperformed GPT-4o-mini across all three reliability dimensions, including Accuracy, Hallucination Rate, and Cross-Domain Stability. These results indicate that the proposed framework can effectively distinguish reliability differences between LLMs by jointly considering factual correctness, hallucination risk, and cross-domain performance stability.

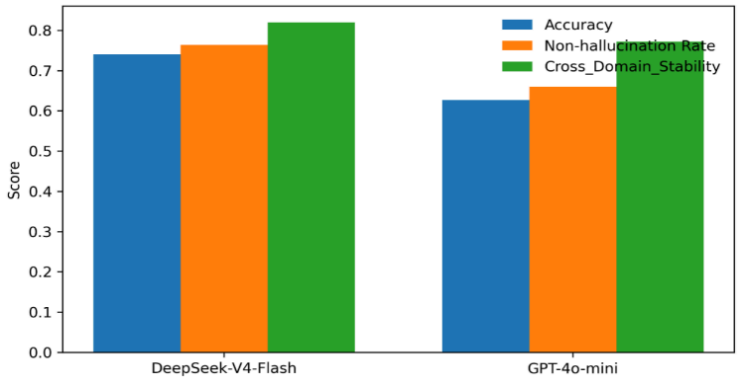

Figure 1. Overall performance comparison of evaluated models

Table 2. Comparison of reliability metrics

Model	Accuracy	Hallucination Rate	CDS	Reliability Score
DeepSeek-V4-Flash	0.740	0.237	0.819	0.770
GPT-4o-mini	0.627	0.340	0.772	0.678

4.3. Category-level performance analysis

4.3.1. Category-level accuracy distribution

The distribution of category-level accuracy across different knowledge categories is presented in Figure 2.

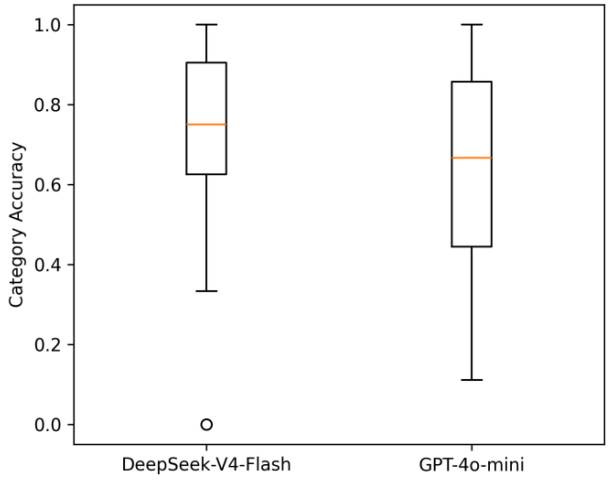

Figure 2. Distribution of category-level accuracy across models

DeepSeek-V4-Flash shows a relatively concentrated distribution of category-level accuracy, whereas GPT-4o-mini exhibits larger performance variations across categories.

This pattern suggests that the two models differ in their ability to maintain consistent performance across diverse knowledge domains. In particular, GPT-4o-mini achieves competitive performance in some categories but shows noticeable degradation in others, resulting in greater variability across domains.

4.3.2. Category-level performance comparison

To provide a more detailed comparison of model performance across individual knowledge categories, a category-level accuracy heatmap is presented (see Figure 3).

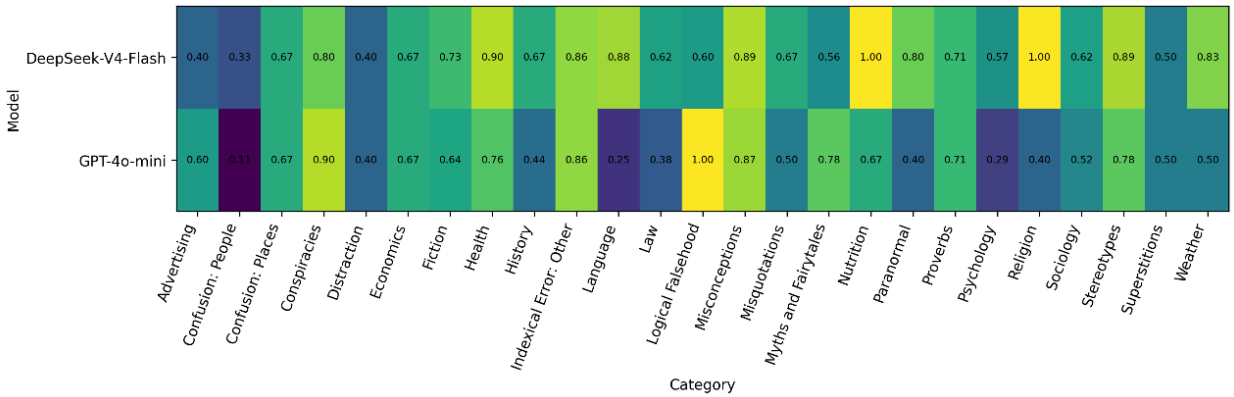

Figure 3. Category-level accuracy heatmap of evaluated models

The heatmap reveals noticeable performance differences between the two models across knowledge categories. For example, substantial variations can be observed in categories such as Region, Psychology, and Confusion: People.

These results further demonstrate the importance of considering category-level performance variation in LLM reliability evaluation.

4.4. Sensitivity analysis

To evaluate the robustness of the reliability scoring framework with respect to weight selection, a sensitivity analysis was conducted. Four weighting configurations were considered, including:

Baseline: (0.4, 0.3, 0.3)

Accuracy-focused: (0.5, 0.25, 0.25)

Hallucination-focused: (0.3, 0.5, 0.2)

Equal-weight: (1/3, 1/3, 1/3)

The experimental results are presented in Figure 4.

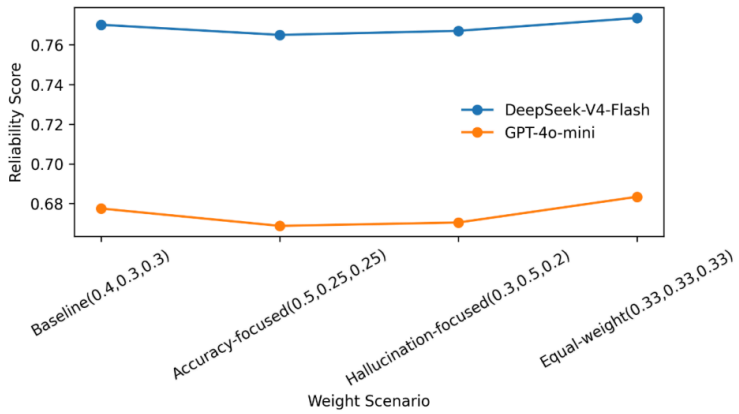

Figure 4. Sensitivity analysis under different weighting configurations

The results show that DeepSeek-V4-Flash consistently achieved higher reliability scores across all tested weighting configurations.

This indicates that the proposed Weighted Geometric Mean scoring method is robust to reasonable changes in weight assignments and does not alter the relative ranking of the evaluated models.

Therefore, the reliability scoring framework demonstrates satisfactory robustness.

5. Discussion

5.1. Interpretation of findings

The experimental results demonstrate that LLM reliability cannot be fully characterized by a single performance metric. The comparison between models suggests that reliability differences are associated with multiple dimensions, including answer correctness, hallucination risk, and cross-domain performance consistency.

The category-level analysis further indicates that models may exhibit different performance patterns across knowledge domains. Therefore, evaluating only overall accuracy may overlook variations in domain-level reliability. The proposed Cross-Domain Stability metric complements overall accuracy by capturing performance consistency across different categories.

From a statistical evaluation perspective, this study suggests that LLM reliability should be considered as a multidimensional property rather than a single measure of task performance.

5.2. Implications for LLM reliability evaluation

First, the experimental results highlight the importance of integrating multiple reliability dimensions in LLM evaluation. Accuracy alone cannot fully capture hallucination risk or variations in domain-level performance. Combining Accuracy, Hallucination Rate, and Cross-Domain Stability provides a more balanced assessment of model reliability.

Second, the Weighted Geometric Mean adopted in this study provides a suitable approach for aggregating multiple reliability indicators. Compared with simple linear weighting methods, geometric aggregation reduces the possibility that strong performance in one dimension compensates for weaknesses in another, making it suitable for reliability evaluation requiring balanced performance across multiple dimensions.

6. Conclusion

This study developed a statistical framework for evaluating LLM reliability by integrating Accuracy, Hallucination Rate, and Cross-Domain Stability into an overall reliability score.

Experiments on the TruthfulQA dataset show that the framework can capture reliability differences between evaluated models. The sensitivity analysis further confirms that the proposed scoring method remains stable under different weighting configurations.

This study has several limitations. First, only two large language models were evaluated. Future studies should include models with different scales and architectures to further examine the generalizability of the framework. Second, the evaluation relies on a single evaluator model, which may introduce potential bias. Future research can incorporate human evaluation, multiple evaluator models, and expert feedback to improve evaluation reliability.

Finally, this study mainly evaluates factual reliability using the TruthfulQA dataset. Future research may extend the framework to broader evaluation scenarios, such as reasoning tasks and domain-specific

applications.

References

[1] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. In *9th International Conference on Learning Representations (ICLR 2021)*. International Conference on Learning Representations, ICLR.

[2] Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., ... Wu, Z. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=uyTL5Bvosj

[3] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., ... Koreeda, Y. (2023). Holistic evaluation of language models. *Transactions on Machine Learning Research*. https://doi.org/10.48550/arXiv.2211.09110

[4] Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3214–3252). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.229

[5] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. https://doi.org/10.1145/3571730

[6] Li, J., Cheng, X., Zhao, X., Nie, J.-Y., & Wen, J.-R. (2023). HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 6449–6464). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.397

[7] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12076–12100). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.741

[8] Raj, H., Rosati, D., & Majumdar, S. (2022). Measuring reliability of large language models through semantic consistency. In *NeurIPS 2022 Workshops: MLSW*. https://mlanthology.org/neuripsw/2022/raj2022neuripsw-measuring/

[9] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (Article 2020). Curran Associates.

[10] Saaty, T. L. (1996). *Decision making with dependence and feedback: The analytic network process: The organization and prioritization of complexity*. RWS Publications.