Consistency in Language Models: Current Landscape, Challenges, and Future Directions

Jekaterina Novikova [1 2] **Carol Anderson** [1] **Borhane Blili-Hamelin** [1] **Domenic Rosati** [3] **Subhabrata Majumdar** [1 4]

Abstract

The hallmark of effective language use lies in consistency: expressing similar meanings in similar contexts and avoiding contradictions. While human communication naturally demonstrates this principle, state-of-the-art language models (LMs) struggle to maintain reliable consistency across task- and domain-specific applications. Here we examine the landscape of consistency research in LMs, analyze current approaches to measure aspects of consistency, and identify critical research gaps. Our findings point to an urgent need for quality benchmarks to measure and interdisciplinary approaches to ensure consistency while preserving utility.

1. Introduction

Consistency—broadly defined as using language similarly in similar settings or avoiding contradictions when using language—is among the most important forms of generalization in the use of language. This ability to maintain consistent outputs is essential for building reliable AI systems that users can trust and depend on. Consistency is both a natural expectation that users have when interacting with language technologies and a prerequisite to deploying them in high-stakes domains (Elazar et al., 2021; Jang et al., 2022; Kim et al., 2025). However, most advanced large language models (LLMs) struggle with consistency and frequently demonstrate inconsistent behavior (Elazar et al., 2021; Raj et al., 2025). Although such examples have been documented in multiple studies, there are no standard approaches to assessing model consistency. As such, there is an ongoing risk of overestimating the performance of state-of-the-art models, as well as of underestimating the risks and potential harms elicited by them.

[1]AI Risk and Vulnerability Alliance [2]Vanguard, Enteprise AI Research [3]Dalhousie University [4]Vijil. Correspondence to: Subhabrata Majumdar <subho@vijil.ai>.

Published at ICML 2025 Workshop on Reliable and Responsible Foundation Models.

Despite early attempts to measure and enhance the consistency of language models (LM) and to understand the roots of inconsistency, research on this topic faces multiple challenges. These include a lack of agreement on terminology and evaluation metrics, and limitations on data and model availability. In this paper, we present a review of current research on consistency in LMs, highlight the most pressing challenges, and provide recommendations for future research. We restrict our attention to *text-only* LMs, which a majority of existing research is based on. See Appendix A for a brief discussion on multimodal consistency.

2. A Review of Consistency Research

Consistency has connections to critical areas in AI research: hallucination (generating made-up information contradicting references), factuality (agreement with real-world knowledge), misinformation (false claims misleading users), and reasoning (logical coherence across statements). We survey literature on consistency in LMs from 2019 to 2025, focusing on peer-reviewed publications and influential preprints that explicitly address consistency metrics, theory, and enhancement.

Terminology The terminology used to describe the consistency of LMs is often confusing, as there is not a single, commonly agreed-upon definition of consistency. Authors either come up with their own definition of the concept that aligns best with the specifics of the work they focus on, or use an overly broad definition, or sometimes just omit defining the term altogether. As a result, existing studies present multiple narrowly focused definitions of consistency that often cover very different aspects of model behavior and sometimes even contradict each other.

Given this interest in model behavior and the implications for potential model applications, in this paper, we limit the otherwise broader concept of consistency to *behavioral consistency*. In psychology, behavioral consistency is closely related to the predictability of behavior, which is equally important for the applications of LMs. Based on how behavioral consistency is approached in the literature, we categorize the different types of consistency into two large groups: logical/formal and nonlogical/informal.

Logical consistency in LLMs was introduced by Jang et al. (2022), as the ability of the model to make decisions without logical contradiction. The rules and principles of formal logic are applied to assess the behavior of a model in a methodical way, allowing for standardized and intuitive measurement. Based on these principles, Jang et al. (2022) classified consistency into negational, symmetric, transitive, and additive types. Negational consistency follows the logical negation property (p is true $\Leftrightarrow \neg p$ is false), i.e. LM's predictions should be opposite for texts with the opposite meanings. Symmetric consistency follows the rule $f(x, y) = f(y, x)$ and implies that the predictions of an LM should be invariant to the input text swap. Transitive consistency can measure deductive reasoning ability and follows the property of transitive inference, represented as $X \rightarrow Y \wedge Y \rightarrow Z$ then $X \rightarrow Z$. This type of consistency was analyzed in natural language inference (NLI) tasks (Li et al., 2019) and question-answering (Q&A) (Asai & Hajishirzi, 2020b; Mitchell et al., 2022).

Semantic consistency, another subpart of the Jang et al. (2022) definition, is one of the most widely used concepts in existing consistency research studies. The idea of semantic consistency is derived from the semantic equivalence property, represented as $f(X) = f(Y)$ if X and Y mean the same. Elazar et al. (2021), and multiple studies later on (Raj et al., 2022; Ohmer et al., 2024), explored this as the ability of a model to make consistent decisions in semantically equivalent contexts.

Nonlogical or informal consistency covers all the other definitions that do not follow the rules of formal logic. For example, Bonagiri et al. (2024) highlight the importance of moral consistency, as the ability to preserve noncontradictory moral values across different situations (Arvanitis & Kalliris, 2020; Marcus, 1980), in LLM alignment. Their approach consists of generating semantically equivalent scenarios and employing consistency checks to see if a target LLM gets the same Semantic Graph Entropy (SaGE) score while responding to these scenarios. Jain et al. (2025) investigated norm inconsistency, defined as the condition in which LLMs apply different norms in similar situations, on applying LLMs in high-risk domains.

Informational and/or factual consistency is another subpart of the Jang et al. (2022) definition frequently used in consistency research. Manakul et al. (2023) used the term *informational consistency*, without explaining or defining it further, to develop a method for fact-checking the responses of black-box models. The term *factual consistency* is often used in the context of automatic summarization (Wang et al., 2020). Factual inconsistency is often referred to as hallucinations and/or faithfulness, i.e., models that generate new information that contradicts the source document (Tam et al., 2023; Maynez et al., 2020). Definitions of factual consistency are often not clearly specified, and instead are replaced with human annotations.

A recent study (Parcalabescu & Frank, 2024) on natural language explanations contrasts faithfulness and *self-consistency*. Self-consistency examines whether similar inputs produce consistent explanations—essentially measuring explanation *stability* across input variations. Faithfulness, meanwhile, evaluates whether the explanation behind a certain model-generated answer *accurately* reflects the model's reasoning process to come up with that answer. While related, they involve different evaluation approaches. Self-consistency requires testing multiple input variations (which may not generalize well across datasets) and does not necessarily involve checking for accuracy of the explanations. On the other hand, faithfulness focuses on the accuracy of individual explanations without such constraints.

Analyzed Tasks A slim majority of studies on LM consistency investigate well-established NLP tasks. Most commonly analyzed tasks include Q&A (Mündler et al., 2024; Raj et al., 2022; Berglund et al., 2024; Li et al., 2023; Wang et al., 2020; Asai & Hajishirzi, 2020a), summarization (West et al., 2024; Cui et al., 2024; Tam et al., 2023; Wang et al., 2020), NLI (Jang et al., 2022; Jang & Lukasiewicz, 2023; Camburu et al., 2020; West et al., 2024; Dziri et al., 2019) and reasoning (Zhang et al., 2024b; Liu et al., 2024b; Chen et al., 2024; Wang et al., 2023) Approximately a third of existing studies do not rely on standard NLP tasks, usually using custom tasks such as generating continuations of sentences from Wikipedia (Mündler et al., 2024). A small number of studies employ use-case specific approaches, for example, measuring stock price prediction accuracy based on textual information such as earnings calls and news articles (Yang et al., 2023).

Dataset Size and Availability The number of testing samples varies substantially across different studies, from a few hundred to tens of thousands. One standard approach to creating a test dataset for measuring consistency is to multiply the prompts in one or more existing benchmarks using perturbation rules or prompt templates (Jang et al., 2022; Fierro & Søgaard, 2022). Another approach is to enhance existing benchmarks with human- or LLM-generated annotations (Liu et al., 2023). To do this, a common method is to create paraphrases of an existing dataset using automatic parahrasing methods (Bonagiri et al., 2024) and/or human annotators (Elazar et al., 2021). The majority of testing datasets are shared publicly, although in some cases the authors only describe the dataset creation process without providing access to the actual dataset.

Evaluated Models More than two-thirds of the studies we examined use transformer-based generative LMs with

decoder-only or encoder-decoder architectures, such as the GPT and OPT series models, BART, and T5 (Jang et al., 2022; Li et al., 2023; Jang & Lukasiewicz, 2023; West et al., 2024; Berglund et al., 2024; Raj et al., 2022; Manakul et al., 2023; Mündler et al., 2024; Zhang et al., 2024b; Liu et al., 2024b; Zhang et al., 2024a; Cheng et al., 2024; Cui et al., 2024; Tam et al., 2023; Wang et al., 2023; Chen et al., 2024; Wang et al., 2020; Nie et al., 2021). The parameter sizes for the models tested range from a few billion to hundreds of billion. Slightly more than half of the papers test proprietary models such as GPT-4 (Li et al., 2023; Jang & Lukasiewicz, 2023; West et al., 2024; Berglund et al., 2024; Mündler et al., 2024; Zhang et al., 2024b; Liu et al., 2024b; Zhang et al., 2024a; Cui et al., 2024; Chen et al., 2024), whose exact sizes have not been publicly disclosed but in some cases are rumored to exceed a trillion parameters. Some studies also consider other types of LMs: about a quarter of papers (Jang et al., 2022; Elazar et al., 2021; Asai & Hajishirzi, 2020a; Yang et al., 2023; Nie et al., 2021; Qin et al., 2021) focus on encoder-only, BERT-style models such as BERT, RoBERTa, and ALBERT.

Evaluation of Consistency Consistency evaluation typically uses two approaches: (1) *input-based sampling*, creating paraphrases or equivalent prompts to test consistent responses to similar inputs, or (2) *output-based sampling*, generating multiple outputs from identical inputs. Output-based sampling with high temperature may artificially inflate inconsistency by forcing models to sample normally-avoided tokens, potentially misrepresenting model behavior.

Metrics to measure different notions of consistency typically depend on pairwise similarity metrics. They compute base metrics such as BERTScore, ROUGE, Entailment, or Contradiction for pairs of outputs given similar inputs and/or context, and aggregate over multiple pairs. In earlier studies, the base metrics were based on token-matching similarities (Elazar et al., 2021). Later papers graduated to notions of semantic similarity that are robust to syntactic variations that can change the wording or structure of a phrase of text while keeping the meaning the same or similar (Raj et al., 2022; Rabinovich et al., 2023; Manakul et al., 2023). Aggregation of a metric across pairs is typically done by simple averaging, with the exception of Mündler et al. (2024), which uses sequential aggregation of contradiction scores to measure factual consistency, and Raj et al. (2025); Kuhn et al. (2023) who use semantic entropy across the entire set of outputs.

Challenges Two important aspects of consistency remain underresearched. First, current work tends to focus excessively on consistency in generations at decoding time. In this process, it ignores encoder-only models and how (in)consistent inputs shape the performance on downstream standard NLP tasks like sentiment prediction. Another underexplored direction is adversarial attacks to degrade consistency. Despite extensive research on adversarial robustness (e.g. the AdvGLUE benchmark (Wang et al., 2022)) and jailbreaks, very few studies explore how inconspicuous or subtle manipulation of prompts can lead to inconsistent LLM responses (Lin et al., 2024a). We do not yet fully understand how much malicious perturbations coupled with slightly different input text can degrade output quality.

The availability of model weights and training datasets—allowing stronger transparency and reproducibility—aid in investigating the root causes of inconsistency. Lin et al. (2024b) showed that analyzing the internal state of the model can improve the transparency of the model and lay the foundation for mitigating hallucinations and inconsistencies. Not only closed-weight models, but also unpublished source code and datasets make it nearly impossible to reproduce and verify claims and findings of some existing publications (Semmelrock et al., 2023).

3. Discussion and Recommendations

As mentioned earlier, we need standardization of terms and definitions for a better understanding of the progress of consistent language model development. Beyond this, we recommend the following focus areas for future research.

Multilingual Consistency Similarly to other topics in NLP research, the overwhelming majority of studies on LM consistency are English-based, significantly limiting our understanding of the topic. To broaden this understanding, more research is needed on both monolingual consistency in non-English languages and on cross-language consistency behaviors.

There is a substantial gap between the amount of training data available for English and that available for all other languages (Üstün et al., 2024). While more than 7,000 languages are spoken around the world today, an astounding 73% of the popular datasets used to train LLMs are primarily or entirely English (Longpre et al., 2023). This severe sampling bias in dataset construction results in disparities in model performance between languages, even in well-studied tasks (Lai et al., 2023). Inherent differences between languages may also significantly influence the consistency of the LMs trained on them. Structural features such as word order or inflectional morphology can vary in their stability across languages (Dediu & Cysouw, 2013). These differences can make it more difficult to train models to produce consistent output for certain languages, even when all languages are equally represented in the training data. More research is necessary to understand the effect of linguistic differences and limitations of multilingual training data on consistency in non-English languages.

Recent work has demonstrated significant challenges in

cross-lingual consistency, i.e., whether a model produces compatible or equivalent outputs when the same query is presented in different languages. Shen et al. (2024) found that LLMs exhibit inconsistent safety behaviors across languages, with safety guardrails being more easily circumvented in non-English languages. Xing et al. (2024) observed that LLMs produce inconsistent factual information when asked about the same knowledge in different languages, suggesting knowledge representation gaps across languages. Qi et al. (2023) examined factual consistency across languages and found that languages more dissimilar to English are less likely to reflect synthetically inserted factual associations through model editing. Jin et al. (2023) evaluated cross-lingual inconsistency specifically in healthcare questions and found discrepancies in medical advice across languages. Zhou & Zhang (2024) explored how political biases manifest inconsistently in bilingual models, revealing that models may express different political positions depending on the input language.

These findings collectively highlight a critical gap in current LLM capabilities: the ability to maintain consistent factual information, safety guardrails, and reasoning in different languages. Cross-lingual consistency represents an important direction for future research, especially as LLMs are deployed globally across linguistic boundaries.

Consistency Evaluation Evaluating consistency has several unique challenges. Most previous studies have used automatic metrics alone to assess consistency in LMs. Although automatic evaluation can ensure objectivity and fast assessment, human evaluation is important to establish an acceptable baseline, especially in highly sensitive or subjective culture-specific applications (e.g. social appropriateness), or when automatic metrics are sufficiently high. Automatic metrics often struggle to capture the different nuances of consistency (factual, logical, semantic), while human evaluation suffers from subjectivity and cognitive biases. The contextual nature of consistency requires evaluation across multiple responses, different phrasings, and various contexts, making comprehensive assessment computationally expensive and logistically challenging. Further complicating matters, consistency evaluation interacts with other dimensions such as factuality, helpfulness, and safety—a model may be internally consistent but factually incorrect, or it may sacrifice consistency to maintain safety.

While several consistency benchmarks have recently emerged, there remains a need for more comprehensive evaluation frameworks that measure all different aspects of consistency in LMs across diverse tasks. Recent benchmarks have made important contributions but typically focus on specific consistency types or limited task domains (Jang et al., 2022; Bonagiri et al., 2024; Cui et al., 2024; Liu et al., 2024b; Paleka et al., 2024; Liu et al., 2024a; Gilhuly & Shahzad, 2025). Future work should focus on developing more holistic benchmarks that address the breadth of consistency challenges outlined above.

Impact Inconsistent output can cause users of language-based systems to receive conflicting or incorrect information. This is problematic in scenarios where factual accuracy is crucial (Tam et al., 2023; Wang et al., 2024)—such as in medical, legal, or financial contexts—especially when such information is used for decision making. In critical systems, such as autonomous vehicles or medical diagnosis support, inconsistent responses can lead to critical safety risks. In less critical applications, inconsistent responses lead to poor user experience, cause frustration, and reduce overall utility (Lazar et al., 2023; van Bergen et al., 2024; Zhang et al., 2024a). Inconsistency can also reflect and magnify the underlying societal biases and stereotypes in the training data, leading to potentially discriminatory outcomes for certain user groups, amplifying unfair use and causing representational harm (Blodgett et al., 2020).

Inconsistency may have some advantages in specific situations. Lower degrees of consistency can lead to diverse and creative outputs, which can be valuable in tasks requiring originality or brainstorming. Inconsistency might reflect the ability of a model to adapt to different contexts or user needs, potentially providing more personalized responses. Inconsistent outputs of a model prompt users to engage more critically with generated content, avoid overreliance, and seek additional verification. This can be beneficial in educational applications, provided the level of possible inconsistency is carefully calibrated.

Improving Consistency There are surprisingly few approaches that actually increase the consistency of LMs. Current proposals to do so fall into two narrow categories. The first approach employs fine-tuning to improve consistency between multiple generations from a LM when supplied with the same or similar inputs. Elazar et al. (2021) used a custom loss function, Raj et al. (2025) used knowledge distillation from more consistent teacher models, and Raj et al. (2025); Zhao et al. (2024) used synthetic datasets of groups of consistent input-outputs. The second approach attempts to improve self-consistency, i.e. consistency between a model's reasoning process and the final answer (Deng et al., 2023; Wang et al., 2023; Wei et al., 2022).

Albeit promising, the above methods primarily address symptoms rather than the fundamental causes of inconsistency. There remains a critical need for research investigating the structural basis of consistency in LMs' representational spaces, consistency-oriented pre-training, and architectures designed to maintain consistency across diverse contexts. Such foundational approaches may eliminate the trade-offs between consistency and other valuable properties like creativity and diversity.

4. Call to Action

We call on the research community to address several key challenges: (1) developing standardized definitions and taxonomies of consistency types; (2) creating comprehensive, multilingual, and cross-lingual benchmarks for consistency evaluation; (3) establishing robust evaluation protocols that combine automatic metrics with human evaluation; (4) investigating the relationship between consistency and other important properties such as factuality, safety, and helpfulness; and (5) developing efficient methods to enhance consistency without sacrificing other beneficial model capabilities. To this end, we emphasize the need for interdisciplinary collaboration, bringing together perspectives from linguistics, psychology, philosophy, and ethics to better understand the multifaceted nature of consistency in human and machine language use. By addressing these challenges collectively, we can move toward LMs that exhibit more reliable, trustworthy, and human-aligned behavior across diverse contexts and applications.

References

Arvanitis, A. and Kalliris, K. Consistency and moral integrity: A self-determination theory perspective. *Journal of Moral Education*, 49(3):316–329, 2020.

Asai, A. and Hajishirzi, H. Logic-guided data augmentation and regularization for consistent question answering. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5642–5650, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.499. URL https://aclanthology.org/2020.acl-main.499.

Asai, A. and Hajishirzi, H. Logic-guided data augmentation and regularization for consistent question answering. *arXiv preprint arXiv:2004.10157*, 2020b.

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., and Evans, O. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=GPKTIktA0k.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, 2020.

Bonagiri, V. K., Vennam, S., Govil, P., Kumaraguru, P., and Gaur, M. SaGE: Evaluating moral consistency in large language models. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 14272–14284, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1243.

Camburu, O.-M., Shillingford, B., Minervini, P., Lukasiewicz, T., and Blunsom, P. Make up your mind! adversarial generation of inconsistent natural language explanations. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4157–4165, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.382. URL https://aclanthology.org/2020.acl-main.382.

Chen, A., Phang, J., Parrish, A., Padmakumar, V., Zhao, C., Bowman, S. R., and Cho, K. Two failures of self-consistency in the multi-step reasoning of LLMs. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=5nBqY1y96B.

Cheng, F., Zouhar, V., Arora, S., Sachan, M., Strobelt, H., and El-Assady, M. Relic: Investigating large language model responses using self-consistency. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. Association for Computing Machinery, 2024. doi: 10.1145/3613904.3641904.

Cui, W., Zhang, J., Li, Z., Damien, L., Das, K., Malin, B., and Kumar, S. Dcr-consistency: Divide-conquer-reasoning for consistency evaluation and improvement of large language models, 2024. URL https://arxiv.org/abs/2401.02132.

Dediu, D. and Cysouw, M. Some structural aspects of language are more stable than others: A comparison of seven methods. *PloS one*, 8(1):e55009, 2013.

Deng, Y., Prasad, K., Fernandez, R., Smolensky, P., Chaudhary, V., and Shieber, S. Implicit chain of thought reasoning via knowledge distillation, 2023. URL https://arxiv.org/abs/2311.01460.

Dziri, N., Kamalloo, E., Mathewson, K., and Zaiane, O. Evaluating coherence in dialogue systems using entailment. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3806–3812, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1381. URL https://aclanthology.org/N19-1381.

Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., and Goldberg, Y. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410. URL https://aclanthology.org/2021.tacl-1.60.

Fierro, C. and Søgaard, A. Factual consistency of multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3046–3052, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-acl.240.

Gilhuly, C. and Shahzad, H. Consistency evaluation of news article summaries generated by large (and small) language models. *arXiv preprint arXiv:2502.20647*, 2025.

Jain, S., Calacci, D., and Wilson, A. *As an AI Language Model, "Yes I Would Recommend Calling the Police": Norm Inconsistency in LLM Decision-Making*, pp. 624–633. AAAI Press, 2025.

Jang, M. and Lukasiewicz, T. Consistency analysis of ChatGPT. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15970–15985, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.991. URL https://aclanthology.org/2023.emnlp-main.991.

Jang, M., Kwon, D. S., and Lukasiewicz, T. Becel: Benchmark for consistency evaluation of language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3680–3696, 2022.

Jin, Y., Chandra, M., Verma, G., Hu, Y., Choudhury, M. D., and Kumar, S. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. *Proceedings of the ACM on Web Conference 2024*, 2023. URL https://api.semanticscholar.org/CorpusID:264405758.

Kim, S. S. Y., Vaughan, J. W., Liao, Q. V., Lombrozo, T., and Russakovsky, O. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *CHI Conference on Human Factors in Computing Systems*, pp. 1–26, Yokohama, Japan, 2025. ACM. ISBN 979-8-4007-1394-1/25/04. doi: 10.1145/3706598.3714020.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023. URL https://arxiv.org/abs/2302.09664.

Lai, V., Ngo, N., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., and Nguyen, T. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 13171–13189, 2023.

Lazar, J., Feng, J. H., Lazar, A., and Wentz, B. Frustration: Still a common user experience. *ACM Transactions on Computer-Human Interaction*, 30(3):1–22, 2023. doi: 10.1145/3582432.

Li, T., Gupta, V., Mehta, M., and Srikumar, V. A logic-driven framework for consistency of neural models. *arXiv preprint arXiv:1909.00126*, 2019.

Li, X. L., Shrivastava, V., Li, S., Hashimoto, T., and Liang, P. Benchmarking and improving generator-validator consistency of language models. *ArXiv*, abs/2310.01846, 2023. URL https://api.semanticscholar.org/CorpusID:263609159.

Lin, W., Gerchanovsky, A., Akgul, O., Bauer, L., Fredrikson, M., and Wang, Z. Llm whisperer: An inconspicuous attack to bias llm responses. *arXiv preprint arXiv:2406.04755*, 2024a.

Lin, Z., Guan, S., Zhang, W., Zhang, H., Li, Y., and Zhang, H. Towards trustworthy llms: a review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review*, 57(9):1–50, 2024b.

Liu, Y., Li, Y., Zhang, J., Fan, P., Zhou, Y., and Liang, K. Aligning with logic: Measuring, evaluating and improving logical preference consistency in large language models. *arXiv preprint arXiv:2410.02205*, 2024a. Updated 2025.

Liu, Z., Lee, I., Du, Y., Sanyal, S., and Zhao, J. Score: A framework for self-contradictory reasoning evaluation. *arXiv preprint arXiv:2311.09603*, 2023.

Liu, Z., Lee, I., Du, Y., Sanyal, S., and Zhao, J. Self-contradictory reasoning evaluation and detection, 2024b. URL https://arxiv.org/abs/2311.09603.

Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., et al. The data provenance initiative: A large scale audit of dataset licensing & attribution in ai. *arXiv preprint arXiv:2310.16787*, 2023.

Manakul, P., Liusie, A., and Gales, M. J. F. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, 2023. URL https://arxiv.org/abs/2303.08896.

Marcus, R. B. Moral dilemmas and consistency. *Journal of Philosophy*, 77(3):121–136, 1980. doi: 10.2307/2025665.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. On faithfulness and factuality in abstractive summarization. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173/.

Mitchell, E., Noh, J. J., Li, S., Armstrong, W. S., Agarwal, A., Liu, P., Finn, C., and Manning, C. D. Enhancing self-consistency and performance of pre-trained language models through natural language inference. *arXiv preprint arXiv:2211.11875*, 2022.

Mündler, N., He, J., Jenko, S., and Vechev, M. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation, 2024. URL https://arxiv.org/abs/2305.15852.

Nie, Y., Williamson, M., Bansal, M., Kiela, D., and Weston, J. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1699–1713, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.134. URL https://aclanthology.org/2021.acl-long.134.

Ohmer, X., Bruni, E., and Hupkes, D. From form (s) to meaning: Probing the semantic depths of language models using multisense consistency. *Computational Linguistics*, pp. 1–51, 2024.

Paleka, D., Hadjikyriacou, A., Daneshjou, R., Gleave, A., and Steinhardt, J. Consistency checks for language model forecasters. *arXiv preprint arXiv:2412.18544*, 2024. Updated January 2025.

Parcalabescu, L. and Frank, A. On measuring faithfulness or self-consistency of natural language explanations. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6048–6089, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.329.

Qi, J., Fern'andez, R., and Bisazza, A. Cross-lingual consistency of factual knowledge in multilingual language models. In *Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://api.semanticscholar.org/CorpusID:264145744.

Qin, L., Xie, T., Huang, S., Chen, Q., Xu, X., and Che, W. Don't be contradicted with anything! CI-ToD: Towards benchmarking consistency for task-oriented dialogue system. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2357–2367, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.182. URL https://aclanthology.org/2021.emnlp-main.182.

Rabinovich, E., Ackerman, S., Raz, O., Farchi, E., and Anaby Tavor, A. Predicting question-answering performance of large language models through semantic consistency. In Gehrmann, S., Wang, A., Sedoc, J., Clark, E., Dhole, K., Chandu, K. R., Santus, E., and Sedghamiz, H. (eds.), *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pp. 138–154, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.gem-1.12.

Raj, H., Rosati, D., and Majumdar, S. Measuring reliability of large language models through semantic consistency, 2022.

Raj, H., Gupta, V., Rosati, D., and Majumdar, S. Improving consistency in large language models through chain of guidance. *Transactions on Machine Learning Research*, 2025. URL https://arxiv.org/abs/2502.15924.

Semmelrock, H., Kopeinik, S., Theiler, D., Ross-Hellauer, T., and Kowald, D. Reproducibility in machine learning-driven research. *arXiv preprint arXiv:2307.10320*, 2023.

Shen, L., Tan, W., Chen, S., Chen, Y., Zhang, J., Xu, H., Zheng, B., Koehn, P., and Khashabi, D. The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2668–2680, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.156. URL https://aclanthology.org/2024.findings-acl.156.

Tam, D., Mascarenhas, A., Zhang, S., Kwan, S., Bansal, M., and Raffel, C. Evaluating the factual consistency of large language models through news summarization. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.),

Findings of the Association for Computational Linguistics: ACL 2023, pp. 5220–5255, Toronto, Canada, July 2023. Association for Computational Linguistics.

Tan, Z., Yang, X., Ye, Z., Wang, Q., Yan, Y., Nguyen, A., and Huang, K. Ssd: Towards better text-image consistency metric in text-to-image generation. *arXiv preprint arXiv:2210.15235*, 2022.

Üstün, A., Aryabumi, V., Yong, Z.-X., Ko, W.-Y., D'souza, D., Onilude, G., Bhandari, N., Singh, S., Ooi, H.-L., Kayid, A., et al. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*, 2024.

van Bergen, R., van der Schalk, B., Kökciyan, N., Otterbacher, J., Haider, J., and Terzimehić, N. "as an ai language model, i cannot": Investigating llm denials of user requests. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–16. ACM, 2024. doi: 10.1145/3613904.3642135.

Wang, A., Cho, K., and Lewis, M. Asking and answering questions to evaluate the factual consistency of summaries. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5008–5020, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.450. URL https://aclanthology.org/2020.acl-main.450.

Wang, B., Xu, C., Wang, S., Gan, Z., Cheng, Y., Gao, J., Awadallah, A. H., and Li, B. Adversarial glue: A multi-task benchmark for robustness evaluation of language models, 2022. URL https://arxiv.org/abs/2111.02840.

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Wang, Y., Wang, M., Manzoor, M. A., Liu, F., Georgiev, G., Das, R. J., and Nakov, P. Factuality of large language models in the year 2024, 2024. URL https://arxiv.org/abs/2402.02420.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

West, P., Lu, X., Dziri, N., Brahman, F., Li, L., Hwang, J. D., Jiang, L., Fisher, J., Ravichander, A., Chandu, K., Newman, B., Koh, P. W., Ettinger, A., and Choi, Y. The generative AI paradox: "what it can create, it may not understand". In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CF8H8MS5P8.

Xing, X., He, Z., Xu, H., Wang, X., Wang, R., and Hong, Y. Evaluating knowledge-based cross-lingual inconsistency in large language models. *ArXiv*, abs/2407.01358, 2024. URL https://api.semanticscholar.org/CorpusID:270870062.

Yang, L., Ma, Y., and Zhang, Y. Measuring consistency in text-based financial forecasting models. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13751–13765, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.769. URL https://aclanthology.org/2023.acl-long.769.

Zhang, M., Jin, L., Song, L., Mi, H., and Yu, D. Inconsistent dialogue responses and how to recover from them. In Graham, Y. and Purver, M. (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 220–230, St. Julian's, Malta, March 2024a. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.16.

Zhang, W., Shen, Y., Wu, L., Peng, Q., Wang, J., Zhuang, Y., and Lu, W. Self-contrast: Better reflection through inconsistent solving perspectives. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3602–3622, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.197.

Zhao, Y., Yan, L., Sun, W., Xing, G., Wang, S., Meng, C., Cheng, Z., Ren, Z., and Yin, D. Improving the robustness of large language models via consistency alignment, 2024. URL https://arxiv.org/abs/2403.14221.

Zhou, D. and Zhang, Y. Political biases and inconsistencies in bilingual gpt models—the cases of the u.s. and china. *Scientific Reports*, 14, 10 2024. doi: 10.1038/s41598-024-76395-w.

A. Appendix: Multimodal Consistency

Until 2022, every consistency study was analyzing robustness of LMs to various text perturbations or to semantically equivalent texts only. Starting in 2022, some interest in

non-textual modalities started to appear that comes primarily from text-to-image model analysis. For example, Tan et al. (2022) explores the challenge of generating consistent and high-quality images from given texts in the task of visual-language understanding and highlights the need to design a better text-image consistency metric, a problem that remains under-explored in the community. In their study, Tan et al. (2022) present a novel CLIP-based metric named Semantic Similarity Distance (SSD) that leads to significantly better text-image consistency while maintaining decent image quality. The attempts to quantify the consistency in text-to-image models are continued by Berglund et al. (2024), which proposes a novel semantic consistency score for image generation that has strong agreement with human annotators. Recently, there was an attempt to evaluate the understanding capability of generative models in both language and vision domains (West et al., 2024). West et al. (2024) conducted interrogative evaluation of image understanding models via visual question answering in an open-ended setting. They investigated whether the models produce consistent output when interrogated about the content of their generated image, and figured out that although models can outperform humans in generation, they regularly show evidence of inconsistency between their generation and understanding performance.