Fiction vs Friction: Challenges in Evaluating LLMs on Data Visualization Tasks

Shani C Spivak*
spivak.s@northeastern.edu
Northeastern University
Boston, MA, USA

Melanie Tory
m.tory@northeastern.edu
The Roux Institute at Northeastern University
Portland, Maine, USA

ABSTRACT

Large language models (LLMs) are often marketed as all-purpose tools, capable of assisting users with a variety of tasks. This has led to the use of LLMs across domains and tasks, and has exposed some fundamental limitations of LLMs. For data visualization tasks (where an LLM is asked to create a visualization or answer a question with a visualization), we call out challenges associated with query specification and the difficulty of verifying results. Differently phrased queries may have the same analytic goal, while similarly phased queries may lead to dramatically different results. Add to this the plethora of visualization guidelines and design choices, and the complexity of evaluating LLMs on data visualization tasks grows quickly. While correct and credible answers take time to sort out, plausible-looking, but limited, hallucinated, or otherwise incorrect model responses are instant and ubiquitous. We explore the challenges associated with this space, and call for consideration of combinations of techniques to spot check model responses and surface errors.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; *HCI theory, concepts and models*; **HCI design and evaluation methods**; **Visualization**; **Visualization design and evaluation methods**; *Empirical studies in visualization.*

KEYWORDS

Large language models, Machine learning, Evaluation, Data visualization, Error characterization, Benchmarking, Human-centered AI, Explainability

ACM Reference Format:
Shani C Spivak and Melanie Tory. 2018. Fiction vs Friction: Challenges in Evaluating LLMs on Data Visualization Tasks. In *Proceedings of (HEAL @ CHI 2025).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Large language models (LLMs) have been touted as inherently context-independent, all purpose tools, trained on a wide swath of information from different sources and capable of a multitude of tasks [32, 36, 40, 42]. The argument goes that the bulk of human knowledge can lead to models that excel at a range of tasks, well beyond the purview of previous natural language processing (NLP) models. Log in to OpenAI to use ChatGPT today and you may see the following pop-up message: "Meet the o3-mini family. Introducing o3-mini and o3-mini-high — two new reasoning models that **excel at coding, science, and anything else that takes a little more thinking.**" This issue of **fit for purpose**, both in the sense that LLMs are treated as all-purpose models when they aren't, and that LLM evaluations can't be generalized to meet this broad use, represents a major limitation of LLM evaluation. Many evaluations and audit processes applied to large language models (LLMs) assume a specific human-facing task, like data visualization (e.g. nvBench), or domain, like employment (e.g. NYLL 144). The result is that model testing, evaluation, validation, and verification diverge in fit and scope from the general-purpose nature of LLMs.

Here we focus on a specific challenge in the context of LLMs for visualization: LLMs have a propensity to generate seemingly plausible answers which sometimes turn out to be wrong [19, 23, 25]. This represents another dimension of difficulty for evaluation because LLM errors can sometimes seem subtle and difficult to identify. This is especially concerning in LLM-enabled data visualization (LLM4VIS), since people often take data visualizations at face value [10, 37] and may not have the time, resources, or ability to verify results.

McNutt et al. [31] discuss how a visualization may look plausible, but lead to incorrect conclusions. These "visualization mirages" may be purposefully or inadvertently developed, and require critical thinking on behalf of a user to identify. This is true whether or not an LLM is used to generate a data visualization, but the complex and often opaque nature of LLMs further complicates surfacing errors. Next, we look at available evaluation techniques. In balance with the time and effort required to evaluate LLM4VIS, and considering that plausible-looking, but incorrect model responses are instant and ubiquitous, we discuss their applicability and limitations. For us, this problem, which involves both Fiction (plausible-looking but incorrect LLM responses) and Friction (the time and effort required to verify LLM responses), is central to LLM4VIS work, as the tension between the two may dictate the likelihood of users bothering to check a provided response. This leads to a balancing act for LLM4VIS evaluation: surface fiction without causing too much friction in the process. While these issues are not necessarily unique to LLM4VIS, this is a space where users have to balance semantic interpretation and validation with verification of a representation that inherently abstracts away a certain level of detail.

HEAL @ CHI 2025, April 2025, Yokohama, Japan

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

2 TRADITIONAL PERFORMANCE MEASUREMENT TECHNIQUES

Many have studied LLM evaluation and auditing practices through various lenses. In financial, medical, education, and other field-specific applications, performance metrics, benchmarks, audits, and other approaches are used to evaluate LLMs. These help to identify and quantify LLM strengths and weaknesses, though each come with their own limitations. These limits may reflect a focus on a specific task or domain, as well as limitations inherent to the method applied. Here we review various evaluation approaches used in LLM4VIS, along with their limitations.

2.1 LLM Metrics

Different metrics have been used to evaluate LLM outputs, including measurements of accuracy, measurements of ethical thresholds including privacy protection, misinformation reduction, fairness, and transparency, measures of fairness including bias mitigation, and measures of robustness including resistance to manipulation and attacks [21]. Hu et al. [21] divide LLM metrics into three categories: 1) multiple-classification metrics which evaluate an LLM's ability to classify text into multiple groups; 2) Token-similarity metrics which evaluate how well the LLM-generated text aligns with reference text; and 3) Question-answering metrics which evaluate LLMs specifically on question-answering tasks. Multiple classification metrics may include accuracy, precision, recall, and F1 score. Token-similarity metrics may include Perplexity, Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE) 1 or 2, ROUGE-L, BertScore, and Metric for Evaluation of Translation with Explicit Ordering (METEOR). Question-answering metrics may include Strict Accuracy (SaCC), Lenient Accuracy (LaCC), and Mean Reciprocal Rank (MRR).

The highly structured and focused nature of these metrics limit their generalizability, and may also differ from real-world scenarios and use cases. Many focus solely on the text of a response, which limits their utility in evaluating LLM4VIS. These metrics also provide a narrow and potentially misleading picture of performance, and metrics can be gamed [4, 27, 49]. Applying these metrics to data visualizations is additionally complicated by the fact that many paths may lead to the same correct answer and a variety of very different resulting visualizations may be correct for the same query.

2.2 Benchmarks

Many benchmarks have been developed to evaluate LLM performance on specific tasks or sets of tasks, including math [22], reasoning [41, 60], multilingual reasoning [45], judgment [63], coding [64], function-calling [39], and multitask accuracy [18].

Benchmarks which evaluate LLMs on data visualization tasks often focus on specific visualizations, query types, tasks, and domains, limiting their generalizability [9, 14, 30, 46, 61]. While these benchmarks enable improved testing of LLMs on various data visualization tasks, some limit which aspects of an LLM response is evaluated (e.g. the visualization, the code used to generate the visualization, the provided explanation) and others limit what types of charts and queries can be tested (e.g. bar charts, scatterplots, etc.). Ford et al. show that LLM-generated charts do not match the accuracy of non-LLM-generated charts based on VQA performance measures [13]. LLMs have also been tested on visual literacy tasks. These tests, including variations of the Visualization Literacy Assessment Test (VLAT), highlight several model limitations, though authors note limitations to this testing, including the potential impacts of model prior knowledge, prompt engineering, and the limits of the test set (8 tasks, 12 data visualizations) [7, 20, 28, 38, 43].

Noted benchmark limitations, beyond those related to narrow applicability, include issues with reliability [52], consistency [58], and other complex failure states [34], as well as benchmark exploitation, dataset contamination, and evaluation bias [5].

2.3 AI Audits

AI audits are meant to be independent evaluations of a model's processes and outputs to highlight potential concerns to stakeholders [2, 32]. LLM audits may be specific to an application or may focus on surfacing specific harms like discrimination, privacy and security gaps, prevention of misleading or malicious information generation, and other negative human impacts. Because LLMs may be used for a variety of tasks, and no audit can comprehensively evaluate all potential tasks, audits provide a limited window of LLM capabilities and concerns. Audits may help spot check LLM4VIS applications for a defined set of query-visualization pairs, but a limited variety of data visualizations may be tested as part of an audit.

2.4 Metamorphic Testing

Lastly we explore the adaptation of software testing practices to both visualizations and LLMs. McNutt et al. suggest looking to metamorphic testing as a way to surface visualization mirages [31]. Metamorphic testing is a technique often used in software testing that uses properties of a program to generate test cases by varying an input in ways that should have a known effect on the output. McNutt et al. consider the "test oracle problem" (distinguishing between correct and incorrect behavior), by iterating through several input changes and verifying that metamorphic relations should remain invariant across these changes actually do so. In adapting metamorphic testing for LLMs, studies often aim to define metamorphic relations (the input transformations with known effects on the output), which then serve as modular evaluation metrics [16, 24, 29, 54]. While we believe metamorphic testing presents a viable path to more effective evaluations of LLM4VIS, some testing may be slower and prone to error, as not all scenarios (e.g. errors, visualizations mirages) are well-covered.

3 ETHICAL CONSIDERATIONS AND HUMAN-CENTERED TECHNIQUES FOR LLM EVALUATIONS

3.1 Trust and Alignment

Ethical risks associated with technologies are often grouped into two categories: quantifiable"hard impacts" (e.g. biased hiring, privacy failures, environmental impacts, etc.) and less measurable, but nonetheless significant, "soft impacts" (e.g. behavior changes, moral implications, overreliance, etc.) [48, 50, 51]. Tigard et al. [50] note the difficulty of evaluating soft impacts and advocate for embedding ethical considerations into the model development process to better

align within the context of the model and potentially preempt some concerns, but note the challenges inherent in attempting to identify a wide array of potential social, moral, and ethical choices and implications. Here we discuss the role of ethics- and human-centered evaluation techniques and consider their application in evaluating LLM4VIS.

3.1.1 User Trust Measurement and Manipulation. Guo et al. discuss the importance of explanations and interactivity in building user trust, noting specifically that visualizations help with interpretation and promote trust [17]. While reinforcing the importance of performance metrics and visualizations, other work also found that including transparency features increased user trust and their understanding of an AI model [11]. Zheng et al. develop a framework to evaluate whether an LLM's response aligns with its intrinsic knowledge as a way to gauge trustworthiness [62]. Shang et al. studied affective and cognitive trust in LLMs and found that GPT models were able to manipulate trust via cognitive and affective routes in diverse contexts like emotional support, technical aid, and social planning [44]. Any evaluation of LLM4VIS should take this into account and test not just for correctness, since, as we know from McNutt et al. technically correct charts could still be misleading [31], but also for trustworthiness.

3.1.2 LLM Alignment and Misalignment. Z. Wang et al. survey approaches to align LLMs with human expectations and identify four tracks: 1) Various types of reward models; 2) Feedback systems; 3) Reinforcement learning; and 4) Optimization [55]. Huang et al. break down hallucinations induced during alignment as resulting from capability misalignment and belief misalignment [23], and note a gap in research related to capability misalignment within LLMs. This is important for LLM4VIS because potential ambiguity in query language could lead to a response that is not aligned with a query goal. As noted in our abstract, differently phrased queries may have the same analytic goal, but at the same time, similarly phased queries may lead to dramatically different results. H.W. Wang et al. gauged various LLMs' perceptual awareness by having them identify takeaways from variously oriented bar charts and found that even leading LLMs struggle with semantic diversity and factual accuracy. Using an example chart and human takeaways helped to induce some alignment, but model were found to be heavily dependent on context rather than data.

3.2 Explainable AI Techniques and Limitations

Cambria et al. surveyed work at the intersection of explainable AI (XAI) and LLMs and found limited work dedicated to developing explanation methods for LLMs [8]. They call on LLM researchers to proactively incorporate explainability practices in LLM design and implementation; they also call on XAI researchers to explore more approachable methodologies. These challenges extend to LLM4VIS.

3.2.1 How Does Traceability Work for LLMs? There has been some work to develop LLM explanations through attribution, though this is limited to text responses [15]. Others trace LLM-generated code back to user requirements [35]. While there were many limitations to these traceability efforts, the work showed promise in improving LLM responses. Traceability is especially applicable to LLM4VIS because of the complex paths LLMs take to generate visualizations through reasoning and code generation.

3.2.2 How Does Accountability Work for LLMs? Understanding how accountability is parsed when a consequence stems from the use of an LLM requires clear institutional boundaries, and may rely on external evaluations like red-teaming and auditing to act as an enabling and reinforcing support structure [3, 26, 33]. Anderljung et al. organmize six requirements for providing this external scrutiny of fronteir models such as LLMs: Access, Searching attitude, Proportionality to the risks, Independence, Resources, and Expertise, calling for the application of this ASPIRE framework. We consider accountability to be important in LLM4VIS because the potential for negative human impact is high when users depend on LLMs for unverified analysis in high stakes scenarios in fields like healthcare, finance, and education.

3.2.3 Presenting LLM Uncertainty and Confidence. Another approach taken to evaluate LLM responses involves quantifying their predictive uncertainty [1, 57, 59], by estimating epistemic uncertainty stemming from a lack of knowledge (unknown unknowns) and/or aleatoric uncertainty stemming from irreducible randomness (such as when multiple answers or interpretations are possible).

Similarly, various studies have attempted to produce confidence scores for LLM results [6, 47, 53, 56]. Sun et al. evaluate four methods for estimating confidence scores based on softmax, raw token scores, verbalized confidences, and a combination of these methods [47].

All of these approaches are couched in the distinction between open-source and proprietary models, which have different limits on uncertainty and confidence estimation, and may be further limited by application in contexts where ground truth may be harder to quantify. Ehsan et al. also point to unanticipated and unintended negative downstream effects from adding AI explanations [12]. Still, providing this information is especially important in LLM4VIS where model uncertainty or confidence in a response could make a large difference in how users make use of the visualization.

4 DISCUSSION

Having reviewed the wide array of approaches to evaluating LLMs, and the specific complexities of LLM4VIS, We call for deeper consideration of the issue of plausible but incorrect LLM responses when developing LLM evaluations. Ideally, evaluations are centered around specific goals which are grounded in human impact, and in evaluating a model against these goals, we should consider how something is optimized as well as what is being optimized. We believe this to be important because users may be less likely to apply a burdensome process, especially when a plausible looking LLM response is so readily and nearly instantly available.

The domain of application also comes with specific constraints and nuances that affect evaluation. When LLMs are evaluated on tasks in certain fields like medicine, finance, education, not only do additional privacy, security, and ethical considerations apply, human impact becomes more central and potentially harder to fully quantify. It is even more important in these cases that we go beyond plausibility and verify responses.

Because no one evaluation approach can cover more than a limited scope of testing, we consider a combination of traditional and

ethical, and human-centered techniques in evaluating LLM4VIS applications. This could include using alignment techniques coupled with benchmark testing in initial stages of development, presenting uncertainty metrics to users with each model response, and incorporating well-defined accountability and remediation procedures once deployed, along with audits as a way to spot check applications.

We call for testing to evaluate how combinations of techniques can help mitigate the risks associated with LLM4VIS applications, while addressing the plausibility problem.

5 CONCLUSION

With the somewhat inherent trust placed on seemingly objective data visualizations [10, 37], it is increasingly important to develop effective evaluations of LLM4VIS. We believe that no single evaluation technique can address the most significant risks associated with LLM4VIS, which could include negative consequences for individuals and groups when applied in healthcare, financial, and other high-impact sectors. We call instead for the development of combinations of the discussed techniques.

REFERENCES

[1] Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. 2024. How many Opinions does your LLM have? Improving Uncertainty Estimation in NLG. https://openreview.net/forum?id=JIlh7OzipV

[2] Maryam Amirizaniani, Elias Martin, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. AuditLLM: A Tool for Auditing Large Language Models Using Multiprobe Approach. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (Boise, ID, USA) *(CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 5174–5179. https://doi.org/10.1145/3627673.3679222

[3] Markus Anderljung, Everett Thornton Smith, Joe O'Brien, Lisa Soder, Benjamin Bucknall, Emma Bluemke, Jonas Schuett, Robert Trager, Lacey Strahm, and Rumman Chowdhury. 2023. Towards Publicly Accountable Frontier LLMs: Building an External Scrutiny Ecosystem under the ASPIRE Framework. arXiv:2311.14711 [cs.CY] https://arxiv.org/abs/2311.14711

[4] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: the risk of rationalization. arXiv:1901.09749 [cs.LG] https://arxiv.org/abs/1901.09749

[5] Sourav Banerjee, Ayushi Agarwal, and Eishkaran Singh. 2024. The Vulnerability of Language Model Benchmarks: Do They Accurately Reflect True LLM Performance? arXiv:2412.03597 [cs.CL] https://arxiv.org/abs/2412.03597

[6] Evan Becker and Stefano Soatto. 2024. Cycles of Thought: Measuring LLM Confidence through Stable Explanations. arXiv:2406.03441 [cs.CL] https://arxiv.org/abs/2406.03441

[7] Alexander Bendeck and John Stasko. 2025. An Empirical Evaluation of the GPT-4 Multimodal Language Model on Visualization Literacy Tasks. *IEEE Transactions on Visualization and Computer Graphics* 31, 1 (2025), 1105–1115. https://doi.org/10.1109/TVCG.2024.3456155

[8] Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Navid Nobani, and Andrea Seveso. 2024. XAI meets LLMs: A Survey of the Relation between Explainable AI and Large Language Models. arXiv:2407.15248 [cs.CL] https://arxiv.org/abs/2407.15248

[9] Nan Chen, Yuge Zhang, Jiahang Xu, Kan Ren, and Yuqing Yang. 2025. VisEval: A Benchmark for Data Visualization in the Era of Large Language Models. *IEEE Transactions on Visualization and Computer Graphics* 31, 1 (2025), 1301–1311. https://doi.org/10.1109/TVCG.2024.3456320

[10] Michael Correll. 2019. Ethical Dimensions of Visualization Research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300418

[11] Jaimie Drozdal, Justin Weisz, Dakuo Wang, Gaurav Dass, Bingsheng Yao, Changruo Zhao, Michael Muller, Lin Ju, and Hui Su. 2020. Trust in AutoML: exploring information needs for establishing trust in automated machine learning systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 297–307. https://doi.org/10.1145/3377325.3377501

[12] Upol Ehsan and Mark O. Riedl. 2024. Explainability pitfalls: Beyond dark patterns in explainable AI. *Patterns* 5, 6 (2024), 100971. https://doi.org/10.1016/j.patter.2024.100971

[13] James Ford, Xingmeng Zhao, Dan Schumacher, and Anthony Rios. 2024. Charting the Future: Using Chart Question-Answering for Scalable Evaluation of LLM-Driven Data Visualizations. arXiv:2409.18764 [cs.CV] https://arxiv.org/abs/2409.18764

[14] Siwei Fu, Kai Xiong, Xiaodong Ge, Siliang Tang, Wei Chen, and Yingcai Wu. 2020. Quda: Natural Language Queries for Visual Data Analytics. arXiv:2005.03257 [cs.CL] https://arxiv.org/abs/2005.03257

[15] Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and Revising What Language Models Say, Using Language Models. arXiv:2210.08726 [cs.CL] https://arxiv.org/abs/2210.08726

[16] Guoxiang Guo, Aldeida Aleti, Neelofar Neelofar, and Chakkrit Tantithamthavorn. 2024. MORTAR: Metamorphic Multi-turn Testing for LLM-based Dialogue Systems. arXiv:2412.15557 [cs.SE] https://arxiv.org/abs/2412.15557

[17] Lijie Guo, Elizabeth M. Daly, Oznur Alkan, Massimiliano Mattetti, Owen Cornec, and Bart Knijnenburg. 2022. Building Trust in Interactive Machine Learning via User Contributed Interpretable Rules. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 537–548. https://doi.org/10.1145/3490099.3511111

[18] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. arXiv:2009.03300 [cs.CY] https://arxiv.org/abs/2009.03300

[19] Michael Townsen Hicks, James Humphries, and Joe Slater. 2024. ChatGPT is bullshit. *Ethics and Inf. Technol.* 26, 2 (jun 2024), 10 pages. https://doi.org/10.1007/s10676-024-09775-5

[20] Jiayi Hong, Christian Seto, Arlen Fan, and Ross Maciejewski. 2025. Do LLMs Have Visualization Literacy? An Evaluation on Modified Visualizations to Test Generalization in Data Interpretation. arXiv:2501.16277 [cs.PF] https://arxiv.org/abs/2501.16277

[21] Taojun Hu and Xiao-Hua Zhou. 2024. Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions. arXiv:2404.09135 [cs.CL] https://arxiv.org/abs/2404.09135

[22] Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, Yue Wu, Ming Yin, Shange Tang, Yangsibo Huang, Chi Jin, Xinyun Chen, Chiyuan Zhang, and Mengdi Wang. 2025. MATH-Perturb: Benchmarking LLMs' Math Reasoning Abilities against Hard Perturbations. arXiv:2502.06453 [cs.LG] https://arxiv.org/abs/2502.06453

[23] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2, Article 42 (Jan. 2025), 55 pages. https://doi.org/10.1145/3703155

[24] Sangwon Hyun, Mingyu Guo, and M. Ali Babar. 2024. METAL: Metamorphic Testing Framework for Analyzing Large-Language Model Qualities. In *2024 IEEE Conference on Software Testing, Verification and Validation (ICST)*. IEEE, Washington, DC, 117–128. https://doi.org/10.1109/ICST60714.2024.00019

[25] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (March 2023), 38 pages. https://doi.org/10.1145/3571730

[26] Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. 2024. Navigating LLM Ethics: Advancements, Challenges, and Future Directions. arXiv:2406.18841 [cs.CY] https://arxiv.org/abs/2406.18841

[27] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2019. *The (Un)reliability of Saliency Methods*. Springer International Publishing, Cham, 267–280. https://doi.org/10.1007/978-3-030-28954-6_14

[28] Sukwon Lee, Sung-Hee Kim, and Bum Chul Kwon. 2017. VLAT: Development of a Visualization Literacy Assessment Test. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 551–560. https://doi.org/10.1109/TVCG.2016.2598920

[29] Zhahao Li, Jinfu Chen, Haibo Chen, Leyang Xu, and Wuhao Guo. 2024. Detecting Bias in LLMs' Natural Language Inference Using Metamorphic Testing. In *2024 IEEE 24th International Conference on Software Quality, Reliability, and Security Companion (QRS-C)*. IEEE, Washington, DC, 31–37. https://doi.org/10.1109/QRS-C63300.2024.00015

[30] Yuyu Luo, Jiawei Tang, and Guoliang Li. 2021. nvBench: A Large-Scale Synthesized Dataset for Cross-Domain Natural Language to Visualization Task. arXiv:2112.12926 [cs.HC] https://arxiv.org/abs/2112.12926

[31] Andrew McNutt, Gordon Kindlmann, and Michael Correll. 2020. Surfacing Visualization Mirages. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376420

[32] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. Auditing large language models: a three-layered approach. *AI and Ethics* 4 (2023), 1085–1115. https://doi.org/10.1007/s43681-023-00289-2

[33] Kelsie Nabben. 2024. AI as a constituted system: accountability lessons from an LLM experiment. *Data & Policy* 6 (2024), e57. https://doi.org/10.1017/dap.2024.58

[34] Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. 2024. Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models. arXiv:2406.02061 [cs.LG] https://arxiv.org/abs/2406.02061

[35] Marc North, Amir Atapour Abarghouei, and Nelly Bencomo. 2024. Code Gradients: Towards Automated Traceability of LLM-Generated Code. In *2024 IEEE 32nd International Requirements Engineering Conference*. IEEE, Washington, DC, 321–329. https://doi.org/10.1109/RE59067.2024.00038

[36] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[37] Anshul Vikram Pandey, Anjali Manivannan, Oded Nov, Margaret Satterthwaite, and Enrico Bertini. 2014. The Persuasive Power of Data Visualization. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2211–2220. https://doi.org/10.1109/TVCG.2014.2346419

[38] Saugat Pandey and Alvitta Ottley. 2025. Benchmarking Visual Language Models on Standardized Visualization Literacy Tests. arXiv:2503.16632 [cs.HC] https://arxiv.org/abs/2503.16632

[39] Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. 2023. Gorilla: Large Language Model Connected with Massive APIs. arXiv:2305.15334 [cs.CL] https://arxiv.org/abs/2305.15334

[40] Jay Peters. 2025. OpenAI lays out plans for GPT-5. https://www.theverge.com/news/611365/openai-gpt-4-5-roadmap-sam-altman-orion

[41] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. arXiv:2311.12022 [cs.AI] https://arxiv.org/abs/2311.12022

[42] Kylie Robison. 2025. Inside OpenAI's $14 million Super Bowl debut. https://www.theverge.com/openai/608476/openai-super-bowl-chatgpt-commercial?utm_source=chatgpt.com

[43] Ariane Moraes Bueno Rodrigues, Gabriel Diniz Junqueira Barbosa, Hélio Côrtes Vieira Lopes, and Simone Diniz Junqueira Barbosa. 2024. Assessing Data Visualization Literacy: Design Implementation and Analysis of a Comprehensive Test. In *Proceedings of the XXIII Brazilian Symposium on Human Factors in Computing Systems (IHC '24)*. Association for Computing Machinery, New York, NY, USA, Article 52, 13 pages. https://doi.org/10.1145/3702038.3702091

[44] Ruoxi Shang, Gary Hsieh, and Chirag Shah. 2025. Trusting Your AI Agent Emotionally and Cognitively: Development and Validation of a Semantic Differential Scale for AI Trust. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES '24)*. AAAI Press, San Jose, California, USA, 1343–1356.

[45] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language Models are Multilingual Chain-of-Thought Reasoners. arXiv:2210.03057 [cs.CL] https://arxiv.org/abs/2210.03057

[46] Arjun Srinivasan, Nikhila Nyapathy, Bongshin Lee, Steven M. Drucker, and John Stasko. 2021. Collecting and Characterizing Natural Language Utterances for Specifying Data Visualizations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 464, 10 pages. https://doi.org/10.1145/3411764.3445400

[47] Yi-Jyun Sun, Suvodip Dey, Dilek Hakkani-Tur, and Gokhan Tur. 2024. Confidence Estimation For LLM-Based Dialogue State Tracking. , 1083-1090 pages. https://api.semanticscholar.org/CorpusID:272689376

[48] Tsjalling Swierstra and Hedwig te Molder. 2012. *Risk and Soft Impacts*. Springer Netherlands, Dordrecht, 1049–1066. https://doi.org/10.1007/978-94-007-1433-5_42

[49] Rachel Thomas and David Uminsky. 2020. The Problem with Metrics is a Fundamental Problem for AI. arXiv:2002.08512 [cs.CY] https://arxiv.org/abs/2002.08512

[50] Daniel W. Tigard, Maximilian Braun, Svenja Breuer, Amelia Fiske, Stuart McLennan, and Alena Buyx. 2024. Embedded Ethics and the "Soft Impacts" of Technology. *Bulletin of Science, Technology & Society* 44, 3-4 (2024), 73–83. https://doi.org/10.1177/02704676241298162

[51] Simone van der Burg. 2009. Taking the "Soft Impacts" of Technology into Account: Broadening the Discourse in Research Practice. *Social Epistemology* 23, 3-4 (2009), 301–316. https://doi.org/10.1080/02691720903364191 arXiv:https://doi.org/10.1080/02691720903364191

[52] Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. 2025. Do Large Language Model Benchmarks Test Reliability? arXiv:2502.03461 [cs.LG] https://arxiv.org/abs/2502.03461

[53] Yuvraj Virk, Premkumar Devanbu, and Toufique Ahmed. 2024. Enhancing Trust in LLM-Generated Code Summaries with Calibrated Confidence Scores. arXiv:2404.19318 [cs.SE] https://arxiv.org/abs/2404.19318

[54] Guanyu Wang, Yuekang Li, Yi Liu, Gelei Deng, Tianlin Li, Guosheng Xu, Yang Liu, Haoyu Wang, and Kailong Wang. 2024. MeTMaP: Metamorphic Testing for Detecting False Vector Matching Problems in LLM Augmented Generation. In *Proceedings of the 2024 IEEE/ACM First International Conference on AI Foundation Models and Software Engineering* (Lisbon, Portugal) *(FORGE '24)*. Association for Computing Machinery, New York, NY, USA, 12–23. https://doi.org/10.1145/3650105.3652297

[55] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Zixu, Zhu, Xiang-Bo Mao, Sitaram Asur, Na, and Cheng. 2024. A Comprehensive Survey of LLM Alignment Techniques: RLHF, RLAIF, PPO, DPO and More. arXiv:2407.16216 [cs.CL] https://arxiv.org/abs/2407.16216

[56] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. arXiv:2306.13063 [cs.CL] https://arxiv.org/abs/2306.13063

[57] Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. 2024. To Believe or Not to Believe Your LLM. arXiv:2406.02543 [cs.LG] https://arxiv.org/abs/2406.02543

[58] Zhe Yang, Yichang Zhang, Tianyu Liu, Jian Yang, Junyang Lin, Chang Zhou, and Zhifang Sui. 2024. Can Large Language Models Always Solve Easy Problems if They Can Solve Harder Ones? arXiv:2406.12809 [cs.CL] https://arxiv.org/abs/2406.12809

[59] Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking LLMs via Uncertainty Quantification. arXiv:2401.12794 [cs.CL] https://arxiv.org/abs/2401.12794

[60] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? arXiv:1905.07830 [cs.CL] https://arxiv.org/abs/1905.07830

[61] Yuge Zhang, Qiyang Jiang, Xingyu Han, Nan Chen, Yuqing Yang, and Kan Ren. 2024. Benchmarking Data Science Agents. arXiv:2402.17168 [cs.AI] https://arxiv.org/abs/2402.17168

[62] Danna Zheng, Danyang Liu, Mirella Lapata, and Jeff Z. Pan. 2024. TrustScore: Reference-Free Evaluation of LLM Response Trustworthiness. https://openreview.net/forum?id=roex1Ylo02

[63] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL] https://arxiv.org/abs/2306.05685

[64] Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, Simon Brunner, Chen Gong, Thong Hoang, Armel Randy Zebaze, Xiaoheng Hong, Wen-Ding Li, Jean Kaddour, Ming Xu, Zhihan Zhang, Prateek Yadav, Naman Jain, Alex Gu, Zhoujun Cheng, Jiawei Liu, Qian Liu, Zijian Wang, David Lo, Binyuan Hui, Niklas Muennighoff, Daniel Fried, Xiaoning Du, Harm de Vries, and Leandro Von Werra. 2024. BigCodeBench: Benchmarking Code Generation with Diverse Function Calls and Complex Instructions. arXiv:2406.15877 [cs.SE] https://arxiv.org/abs/2406.15877