

AuditLLM: A Tool for Auditing Large Language Models Using Multiprobe Approach

Maryam Amirizaniani
amarayam@uw.edu
University of Washington
Seattle, WA, USA

Elias Martin
eamart34@uw.edu
University of Washington - Bothell
Bothell, WA, USA

Tanya Roosta^{*}
tanyaroosta@gmail.com
UC Berkeley, Amazon
Saratoga, CA, USA

Aman Chadha^{*}
hi@aman.ai
Stanford University, Amazon GenAI
Palo Alto, CA, USA

Chirag Shah
chirags@uw.edu
University of Washington
Seattle, WA, USA

Abstract

As Large Language Models (LLMs) are integrated into various sectors, ensuring their reliability and safety is crucial. This necessitates rigorous probing and auditing to maintain their effectiveness and trustworthiness in practical applications. Subjecting LLMs to varied iterations of a single query can unveil potential inconsistencies in their knowledge base or functional capacity. However, a tool for performing such audits with a easy to execute workflow, and low technical threshold is lacking. In this demo, we introduce “AuditLLM,” a novel tool designed to audit the performance of various LLMs in a methodical way. AuditLLM’s primary function is to audit a given LLM by deploying multiple probes derived from a single question, thus detecting any inconsistencies in the model’s comprehension or performance. A robust, reliable, and consistent LLM is expected to generate semantically similar responses to variably phrased versions of the same question. Building on this premise, AuditLLM generates easily interpretable results that reflect the LLM’s consistency based on a single input question provided by the user. A certain level of inconsistency has been shown to be an indicator of potential bias, hallucinations, and other issues. One could then use the output of AuditLLM to further investigate issues with the aforementioned LLM. To facilitate demonstration and practical uses, AuditLLM offers two key modes: (1) Live mode which allows instant auditing of LLMs by analyzing responses to real-time queries; and (2) Batch mode which facilitates comprehensive LLM auditing by processing multiple queries at once for in-depth analysis. This tool is beneficial for both researchers and general users, as it enhances our understanding of LLMs’ capabilities in generating responses, using a standardized auditing platform.

CCS Concepts

- Large Language Model → Auditing LLMs Tools.

Keywords

Large Language Model, Auditing LLMs Tools

1 Introduction

In the rapidly evolving landscape of artificial intelligence, Large Language Models (LLMs) have emerged as cornerstones of contemporary AI applications, driving innovations in natural language processing (NLP), automated content creation, and decision support systems [29]. Their ability to understand, generate, and interact using human-like language has paved the way for unprecedented advancements across various domains. For example, LLMs have become increasingly integrated into critical sectors, such as healthcare [11, 34, 39], education [7, 16, 19], legal [18] and financial services [13, 41]. Given the mission critical nature of these applications, it is imperative to ensure the LLMs are reliable and robust. Recognizing this, the academic and tech communities have underscored the importance of rigorous auditing mechanisms to probe LLM performance. Previous research has laid a solid foundation for understanding the potential of LLM auditing, focusing on identifying biases [21, 30, 33], hallucinations [5, 27, 38], inconsistency [31, 40], reliability [28, 43] and failures in model responses [25].

Building on the existing foundational work [25], we introduce a novel auditing tool, AuditLLM, which provides a general-purpose solution for auditing LLMs. A simple but effective approach for probing LLMs involves deploying multiple versions of the same query to identify any inconsistencies in the model’s output. To implement this audit technique effectively, a mechanism for crafting these questions or probes is essential. Although humans are capable of formulating various ways to pose the same question for probing purposes, this approach is not scalable. Therefore, we adopt the technique described in [1] to streamline probe generation using a different LLM.

To facilitate this, AuditLLM offers two primary features: **(1) Live Mode:** This mode allows for real-time auditing, where a user’s single inputted question is transformed into five distinct yet related prompts. The LLM’s responses are then generated and assessed for semantic similarity, offering immediate insights into the model’s potential inconsistencies; and **(2) Batch Mode:** Batch mode caters to a broader analysis by accepting multiple questions at once, evaluating the LLM’s performance on a larger scale, and providing a semantic consistency score for each generated response. These innovative features of AuditLLM bolster the robustness, transparency, and ethical deployment of LLMs, thereby advancing the state-of-the-art in Artificial Intelligence.

^{*}Work does not relate to position at Amazon.

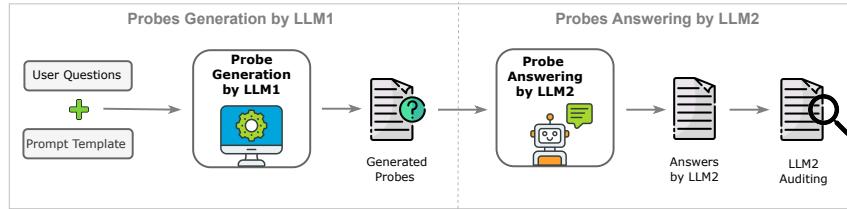


Figure 1: The AuditLLM system employs two LLMs: LLM1 generates probes and LLM2 provides responses to these probes.

The **motivation** for this research is to audit the inconsistencies in LLMs as reflected in question-and-answer tasks, so as to develop a sense for potential bias or hallucinations. The **novelty** of AuditLLM lies in its specialized approach to probing LLMs through the lens of inconsistency detection in the above-mentioned modes. Our **contributions** can be outlined as follows:

- (1) The live mode feature of AuditLLM allows for the immediate auditing of LLMs in real-time in an interactive way.
- (2) The batch mode of AuditLLM supports an extensive evaluation of LLMs by enabling the submission of numerous questions simultaneously, conducting a detailed and comprehensive audit.

In the remainder of this paper, we describe each mode, illustrating their capabilities to detect and highlight inconsistencies in the LLM generated responses.

2 Related Work

The growing application of LLMs in information retrieval (IR) highlights the need for a critical evaluation of their associated risks, a concern extensively recognized within the academic community [2, 3]. Auditing LLMs, aimed at assessing specific attributes to mitigate these risks, has emerged as a crucial research domain. Studies, such as [37], have undertaken thorough evaluations comparing models like GPT-4 to its predecessors in terms of toxicity, stereotype bias, and adversarial robustness. Nonetheless, existing auditing methods often lack generalization capabilities and require substantial human involvement. For instance, Petroni et al. [23] and Jiang et al. [15] employed fill-in-the-blank prompts to audit LLMs’ knowledge of particular entities, but this method struggles with generalization when applied to generative tasks. This limitation stems from the dependency on the specific knowledge encoded within each model, restricting broader applicability across various LLMs. Additionally, initiatives like RealToxicityPrompts[10] and ToxiGen[12] have developed benchmarks with toxic and non-toxic prompts to audit LLM output safety; however, these approaches also require considerable human effort, and the creation of new benchmarks for different domains. Similarly, [4] manually developed a codebook to categorize biases in benchmarks, and [9] involved red team members in writing attack prompts to test LLM robustness, further illustrating the intensive human labor required in these methods.

Systematic auditing of LLMs has been explored by few studies. Mökander et al. [20] introduced a three-tiered auditing pipeline to address the ethical and social challenges of LLMs, covering audits of model designers, pre-release LLM evaluations, and audits

of applications using these models. Although conceptually robust, the framework provides limited practical implementation guidance. Further advancing this field, Ribeiro and Lundberg [26] introduced AdaTest, a collaborative process that combines human expertise and LLM capabilities to craft unit tests for auditing purposes. In this methodology, an LLM generates test suggestions based on existing examples within a given topic, which are then refined by users. Subsequent iterations of testing are organized using a predefined test tree, guiding the LLM in generating new test suggestions. However, this method is somewhat inflexible, depending solely on a single LLM for test generation and lacking a systematic approach to the creation of test questions.

Building on this, AdaTest++ [25] enhances the original model by enabling users to specifically request test suggestions, thereby allowing for more targeted exploration of potential LLM errors and facilitating more purposeful guidance in subsequent testing loops. While AdaTest and AdaTest++ represent significant strides in creating a systematic auditing pipeline, they exhibit notable limitations: (1) the reliance on GPT-3.5 restricts the scope of LLMs auditable within these frameworks; (2) the random generation of test sets lacks a systematic approach and fails to meet specific auditing criteria; and (3) the methodologies are not suited for batch auditing of LLMs. Our proposed framework – AuditLLM – can effectively address these issues as described in the following section.

These efforts reflect advancements in LLM auditing for responsible use, however, none of these approaches have specifically targeted LLM inconsistencies. Our proposed framework, AuditLLM, has been designed specifically to address this gap in auditing LLMs. Given that inconsistencies could be harbingers for issues of bias and hallucination, focusing on auditing an LLM for its inconsistencies allows researchers and developers to be broader in their auditing efforts.

3 AuditLLM Tool Demonstration

AuditLLM¹, serves as a systematic tool to reveal inconsistencies and conduct audits through the utilization of two LLMs: LLM1 and LLM2. As depicted in Figure 1, LLM1 is tasked with the creation of probes based on the given prompt template, whereas LLM2 is responsible for generating responses to these probes. The prompt template for LLM1 is iteratively created using human-in-the-loop technique described by Amirizaniani et al. [1], and shown in Figure 2. The code for AuditLLM is also available on HuggingFace².

¹Implemented with the Gradio Library, this tool is available as a free, open-source project on Huggingface: <https://huggingface.co/spaces/Amirizaniani/AuditLLM>.

²<https://huggingface.co/spaces/Amirizaniani/AuditLLM/blob/main/app.py>

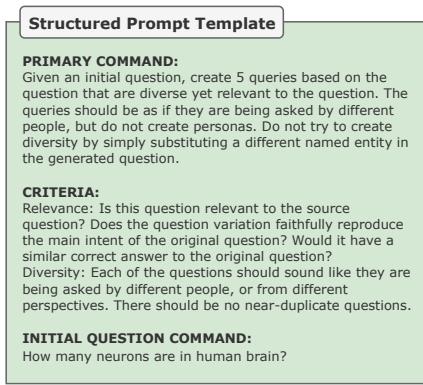


Figure 2: Prompt Template for LLM1 [1].

AuditLLM introduces two distinct modes, live and batch, both designed to use multiple probes generated by LLM1 to detect discrepancies within the responses produced by LLM2. These modes delve into meaning comprehension of generated responses, enabling analysis of the model's consistency by AuditLLM. For a fair comparison, the temperature of LLM2 was set to 0.5 and max_length was configured to 512. The functionalities of each mode are detailed below. A video illustrating the functionality is also hosted on YouTube³.

3.1 Live Mode

The live mode allows for a real-time interrogation of LLMs through the submission of individual queries, providing immediate insights into the model's potential inconsistencies. The live mode comprises the following components, as illustrated in Figure 3:

- **Step 1:** Users are provided with the capability to select from a predefined list of LLMs within the AuditLLM tool for auditing purposes. Currently available models include Llama 2-7B [35], Falcon [22], Zephyr 7B [36], Vicuna [6], and Alpaca [32]. More models are being added through ongoing developments.
- **Step 2:** Users are required to input a query, which serves as the basis for auditing the LLM.
- **Step 3:** Upon receiving user input, AuditLLM initiates the creation of customized probes aimed at evaluating the designated LLM. The prompt template for LLM1 is generated using the method outlined in [1], which utilizes human-in-the-loop validation and is based on criteria of relevance and diversity. We are using Mistral 7B LLM [14] with temperature of 0.0 as our LLM1 to generate five probes that are not only closely related to the user's initial query but also cover a wide range of inquiries, enabling a comprehensive auditing. The prompt template employed is depicted in Figure 2.
- **Step 4:** Users are then afforded the opportunity to select from the generated probes, identifying those that most effectively facilitate the identification of inconsistencies within the audited LLM. This step is critical for narrowing down the focus of the audit to areas of potential concern or interest.
- **Step 5:** In the final step, the audited LLM (LLM2 in Figure 3) generates responses to the user-selected probes from the previous step. The semantic similarity of these responses is quantitatively

determined by leveraging the BERTScore [42] via the sentence-transformer variant of MPNet (all-mpnet-base-v2) and cosine similarity, aiming to highlight sentences with semantic similarity. The culmination of this step is the presentation of a semantic similarity score, providing a metric for comparing responses and thus gauging the LLM consistency in processing and understanding queries.

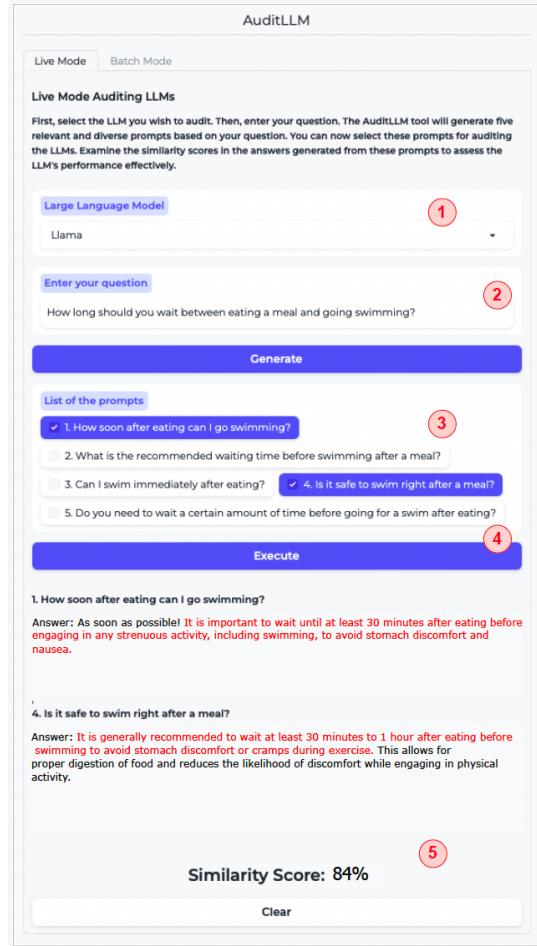


Figure 3: AuditLLM - Live Mode.

Figure 3 showcases an application of this mode in the AuditLLM tool, following the steps previously outlined. AuditLLM highlights the semantic similarity between responses, applying a 60% similarity threshold. Semantically similar sentences are highlighted with their similarity scores across texts, underscoring the consistent quality of the responses to these probes. This analysis of similarities across responses allows us to identify any inconsistencies in the model's outputs. A robust and reliable LLM should produce semantically consistent responses to similar questions, irrespective of their phrasing.

3.2 Batch Mode

The batch mode is designed for broader, more comprehensive assessments, allowing for the simultaneous submission of multiple

³<https://www.youtube.com/watch?v=49UMQOd7Rm0>

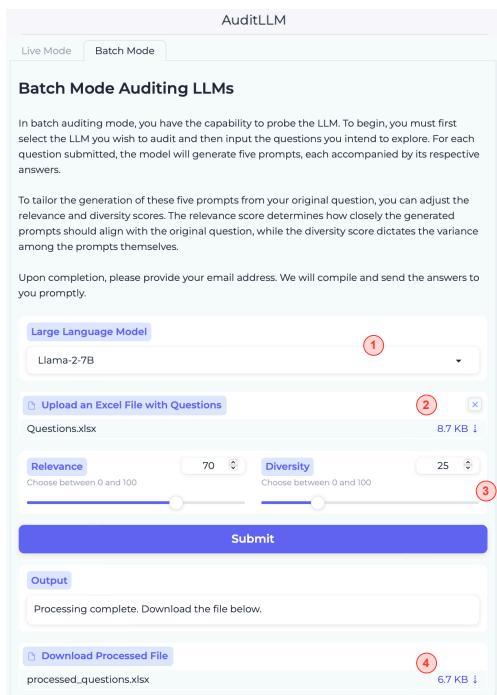


Figure 4: AuditLLM - Batch Mode.

questions for auditing the model’s inconsistencies. The batch mode incorporates the following components, as delineated in Figure 4:

- **Step 1:** The AuditLLM tool offers users the flexibility to choose from a curated selection of the above mentioned LLMs for the purpose of auditing.
- **Step 2:** Users are requested to upload a set of questions in .xlsx format to serve as the basis for auditing the LLM.
- **Step 3:** In this step, we build upon existing research in the ‘Question Generation’ task, developing new probes from user questions with a focus on the criteria of relevance [8] and diversity [24]. This methodology requires that each probe closely aligns with the original user query and exhibits variety among them. Users are given the option to assign scores to each criterion, determining the extent to which the generated probes should exhibit relevance and diversity. These scores are then incorporated into the above mentioned prompt template, guiding the generation process to maintain query relevance and encompass diverse probes. For the probe generation, AuditLLM utilizes a Mistral 7B LLM set at a temperature of 0.0.
- **Step 4:** AuditLLM retrieves responses from the audited LLM via tailored probes, and assesses semantic similarity using the BERTScore, similar to the Section 3.1, via the sentence-transformer to identify closely related sentences. This includes calculating a semantic similarity score to quantitatively evaluate consistency. AuditLLM aggregates these findings, along with relevant details, and makes them available in an Excel format, ready for users to download.

To evaluate the effectiveness of AuditLLM, we conducted a batch mode run using questions from the TruthfulQA dataset [17] as

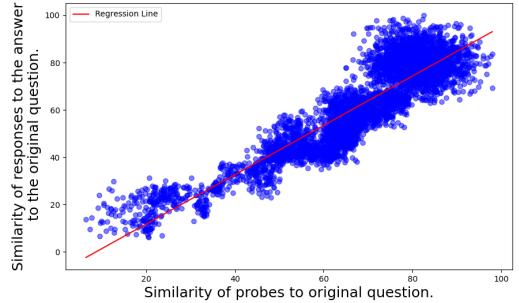


Figure 5: Prompt Similarity vs. Responses Similarity.

inputs. The results are displayed in Figure 5. The x-axis shows the similarity of the five probes to the original question. The y-axis shows the similarity of the generated responses (by LLM2) to the answer for the original question as found in the dataset. We expect that as the similarity among the probes goes up, so should the similarity among the responses. In a well-balanced output, the regression line should have a slope of 45°. A higher slope indicates more similarity or consistency among the responses even when the probes are less similar (more diverse). A lower slope indicates greater divergence in responses for similar probes, indicating higher inconsistencies with the given LLM. One could use the batch mode to run the same set of questions through different LLMs and compare their output distribution or regression lines to compare these LLMs for their consistencies and robustness.

4 Conclusion

AuditLLM provides a methodical auditing tool that examines the consistency of LLM responses to variably phrased queries. Its primary purpose is to evaluate a given language model by utilizing multiple probes generated from a single question. This methodology identifies discrepancies in the model’s understanding or performance, including potential biases, hallucinations, and other related issues [1]. Consequently, this process improves the reliability of these models in practical applications. With features like Live and Batch modes, AuditLLM offers valuable insights into the performance of LLMs through both real-time and extensive probing, making it an essential resource for both researchers and general users seeking to validate and improve the operational integrity of these complex systems.

AuditLLM, while innovative, does have its limitations. Specifically tailored for probing LLMs within the realm of textual data, it utilizes inconsistencies in responses as the primary mechanism for detecting potential issues. Despite incorporating six open-source LLMs recognized for their robust performance on the Huggingface leaderboard⁴, it cannot audit the full spectrum of available LLMs. Additionally, given the substantial size of these models, loading them to generate responses can be time-intensive. Furthermore, variability in responses is observed when identical questions are administered in successive rounds, leading to disparate outcomes if a question is reiterated, as the LLMs yield unique answers. Moreover, AuditLLM’s reliance on the LangChain⁵ library requires that

⁴ https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

⁵ <https://www.langchain.com>

any modifications to LangChain prompt corresponding updates to AuditLLM to ensure continued functionality and compatibility. We are continuing to refine this tool by adding support for more LLMs, plugging in other ways to create multiple probes (e.g., using personas), and providing more methods for assessing inconsistencies.

ACKNOWLEDGEMENTS

This work was supported by the Bill and Melinda Gates Foundation.

References

- [1] Maryam Amirizaniani, Jihan Yao, Adrian Lavergne, Elizabeth Snell Okada, Aman Chadha, Tanya Roosta, and Chirag Shah. 2024. LLMAuditor: A Framework for Auditing Large Language Models Using Human-in-the-Loop. *arXiv preprint arXiv:2402.09346* (2024).
- [2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) (*FAccT '21*). Association for Computing Machinery, New York, NY, USA, 610–623. <https://doi.org/10.1145/3442188.3445922>
- [3] Su Lin Blodgett, Solon Barocas, and Hal Daumé III. 2020. Language (technology) is power: A critical survey of “bias” in nlp. *arXiv preprint arXiv:2005.14050* (2020).
- [4] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *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)*. Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 1004–1015. <https://doi.org/10.18653/v1/2021.acl-long.81>
- [5] Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2023. Hallucination Detection: Robustly Discerning Reliable Answers in Large Language Models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (<conf-loc>, <city>Birmingham</city>, <country>United Kingdom</country>, </conf-loc>) (CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 245–255. <https://doi.org/10.1145/3583780.3614905>
- [6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuai Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023) (2023).
- [7] Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*. 323–325. <https://doi.org/10.1109/ICALT58122.2023.00100>
- [8] Liam Dugan, Eleni Miltakaki, Shriyash Upadhyay, Etan Ginsberg, Hannah Gonzalez, DaHyeon Choi, Chunling Yuan, and Chris Callison-Burch. 2022. A Feasibility Study of Answer-Agnostic Question Generation for Education. In *Findings of the Association for Computational Linguistics: ACL 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 1919–1926. <https://doi.org/10.18653/v1/2022.findings-acl.151>
- [9] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerer, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *arXiv:2209.07858 [cs.CL]*
- [10] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. *arXiv:2009.11462 [cs.CL]*
- [11] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare* 3, 1, Article 2 (oct 2021), 23 pages. <https://doi.org/10.1145/3458754>
- [12] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. *arXiv:2203.09509 [cs.CL]*
- [13] Allen H Huang, Hui Wang, and Yi Yang. 2023. FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research* 40, 2 (2023), 806–841. <https://doi.org/10.1111/1911-3846.12832>
- [14] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).
- [15] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How Can We Know What Language Models Know? *arXiv:1911.12543 [cs.CL]*
- [16] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274. <https://doi.org/10.1016/j.lindif.2023.102274>

- [17] Stephanie Lin, Jacob Hilton, and Owain Evans. 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)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3214–3252. <https://doi.org/10.18653/v1/2022.acl-long.229>
- [18] Yifei Liu, Yiqian Wu, Yating Zhang, Changlong Sun, Weiming Lu, Fei Wu, and Kun Kuang. 2023. ML-LJP: Multi-Law Aware Legal Judgment Prediction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (<conf-loc>, <city>Taipei</city>, <country>Taiwan</country>, <conf-loc>) (SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 1023–1034. <https://doi.org/10.1145/3539618.3591731>
- [19] Jesse G Meyer, Ryan J Urbanowicz, Patrick CN Martin, Karen O'Connor, Ruowang Li, Pei-Chen Peng, Tiffani J Bright, Nicholas Tattonetti, Kyoung Jae Won, Graciela Gonzalez-Hernandez, et al. 2023. ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining* 16, 1 (2023), 20. <https://doi.org/10.1186/s13040-023-00339-9>
- [20] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. Auditing large language models: a three-layered approach. *arXiv preprint arXiv:2302.08500* (2023).
- [21] Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2023. More human than human: Measuring ChatGPT political bias. <https://doi.org/10.1007/s1127-023-01097-2> (2023).
- [22] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116* (2023).
- [23] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language Models as Knowledge Bases? *arXiv:1909.01066 [cs.CL]*
- [24] Vatsal Raina and Mark Gales. 2022. Multiple-Choice Question Generation: Towards an Automated Assessment Framework. *arXiv:2209.11830 [cs.CL]*
- [25] Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershia. 2023. Supporting Human-AI Collaboration in Auditing LLMs with LLMs. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (<conf-loc>, <city>Montréal</city>, <state>QC</state>, <country>Canada</country>, <conf-loc>) (AIES '23)*. Association for Computing Machinery, New York, NY, USA, 913–926. <https://doi.org/10.1145/3600211.3604712>
- [26] Marco Tulio Ribeiro and Scott Lundberg. 2022. Adaptive Testing and Debugging of NLP Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3253–3267. <https://doi.org/10.18653/v1/2022.acl-long.230>
- [27] Mabashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh Menon, Md Parvez, and Zhe Feng. 2023. DelucionQA: Detecting Hallucinations in Domain-specific Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 822–835. <https://doi.org/10.18653/v1/2023.findings-emnlp.59>
- [28] Xinyu Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. Is chatgpt we trust? measuring and characterizing the reliability of chatgpt. *arXiv preprint arXiv:2304.08979* (2023).
- [29] Wenqi Shi, Yuchen Zhuang, Yuanda Zhu, Henry Iwinski, Michael Wattenbarger, and May Dongmei Wang. 2023. Retrieval-Augmented Large Language Models for Adolescent Idiopathic Scoliosis Patients in Shared Decision-Making. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (Houston, TX, USA) (BCB '23)*. Association for Computing Machinery, New York, NY, USA, Article 14, 10 pages. <https://doi.org/10.1145/3584371.3612956>
- [30] Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, Angela Fan, Suzana Ilic, Thomas Wolf, and Matthias Gallé (Eds.). Association for Computational Linguistics, virtual+Dublin, 26–41. <https://doi.org/10.18653/v1/2022.bigsience-1.3>
- [31] Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the Factual Consistency of Large Language Models Through News Summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naapaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 5220–5255. <https://doi.org/10.18653/v1/2023.findings-acl.322>
- [32] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsumori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html> 3, 6 (2023), 7.
- [33] Vishesh Thakur. 2023. Unveiling gender bias in terms of profession across LLMs: Analyzing and addressing sociological implications. *arXiv preprint arXiv:2307.09162* (2023).
- [34] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine* 29, 8 (2023), 1930–1940. <https://doi.org/10.1038/s41591-023-02448-8>
- [35] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasminne Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian CantonFerrer, Moya Chen, Guillem Cucurull, David Esioubi, Jude Fernandes, Jeremy Fu, Wenjin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardasz, Viktor Kerkez, Madiam Khabsa, Isabel Kloumann, Artem Korenev, , Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Miyahlov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poultney, Jeremy Reizenstein, Rashi Rungra, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian XiangKuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
- [36] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944* (2023).
- [37] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xiu, Zidi Xiong, Ritik Dutta, Ryhan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sammi Koyejo, Dawn Song, and Bo Li. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. *arXiv preprint arXiv:2306.11698* (2023).
- [38] Shiping Yang, Renliang Sun, and Xiaojun Wan. 2023. A New Benchmark and Reverse Validation Method for Passage-level Hallucination Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 3898–3908. <https://doi.org/10.18653/v1/2023.findings-emnlp.256>
- [39] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. 2022. A large language model for electronic health records. *NPJ Digital Medicine* 5, 1 (2022), 194. <https://doi.org/10.1038/s41746-022-00742-2>
- [40] Wenta Ye, Mingfeng Ou, Tianyi Li, Xuetao Ma, Yifan Yanggong, Sai Wu, Jie Fu, Gang Chen, and Junbo Zhao. 2023. Assessing Hidden Risks of LLMs: An Empirical Study on Robustness, Consistency, and Credibility. *arXiv preprint arXiv:2305.10235* (2023).
- [41] Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiaoyang Liu. 2023. Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models. In *Proceedings of the Fourth ACM International Conference on AI in Finance (<conf-loc>, <city>Brooklyn</city>, <state>NY</state>, <country>USA</country>, <conf-loc>) (ICAF '23)*. Association for Computing Machinery, New York, NY, USA, 349–356. <https://doi.org/10.1145/3604237.3626866>
- [42] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).
- [43] Li Zhong and Zilong Wang. 2023. A study on robustness and reliability of large language model code generation. *arXiv preprint arXiv:2308.10335* (2023).