Benchmarking Prompt Sensitivity in Large Language Models

Abstract. Large language Models (LLMs) are highly sensitive to variations in prompt formulation, which can significantly impact their ability to generate accurate responses. In this paper, we introduce a new task, Prompt Sensitivity Prediction, and a dataset `PromptSET` designed to investigate the effects of slight prompt variations on LLM performance. Using TriviaQA and HotpotQA datasets as the foundation of our work, we generate prompt variations and evaluate their effectiveness across multiple LLMs. We benchmark the *prompt sensitivity prediction* task employing state-of-the-art methods from related tasks, including LLM-based self-evaluation, text classification, and query performance prediction techniques. Our findings reveal that existing methods struggle to effectively address prompt sensitivity prediction, underscoring the need to understand how information needs should be phrased for accurate LLM responses.

1 Introduction

Large language models (LLMs) can generate human-like responses to a wide array of prompts, from answering specific queries to planning for accumulating information to answer complex questions [13]. Despite their usefulness, a notable challenge in working LLMs is their sensitivity to prompt formulation [16]. Small variations in the phrasing, structure, or even punctuation of prompts can often lead to substantially different outputs [24,22,19]. To illustrate this issue, consider the sample prompts shown in Table 1. In this table, we present samples from the TriviaQA [12] and HotPotQA [27] question-answering datasets where the LLM responds correctly and accurately to the original prompts. However, with only slight modifications in wording, we observe that the LLM (in this case LLaMA3.1) fails to provides the correct response.

This challenge, which we refer to as *prompt sensitivity*, highlights the challenges users face when crafting their prompts [7,29]. For this reason, *prompt engineering*, the art of designing effective prompts, has become an active area of research [15,4]. Researchers have already examined the effects of various prompt modifications, including minor structural and formatting changes [24], adversarial prompting [28], and generating prompts with different levels of specificity [20]. We hypothesize that prompt sensitivity could arise because of several reasons. For instance, an LLM may successfully respond to prompts closely aligned with examples seen during training, but struggle with slight modifications that it has not encountered in the past. Another factor could be the model's reliance on specific syntactic or semantic patterns to interpret prompts accurately, which may be impacted due to slight changes in the prompt.

To this end and in this paper, we introduce a novel task and its accompanying dataset specifically curated for *prompt sensitivity prediction*. By curating a collection of prompts and their variations, we aim to predict whether a given LLM would be able to effectively respond to an input prompt or whether it would fail to provide a satisfactory response. Our proposed dataset serves as a benchmark for studying prompt sensitivity, thus setting

Table 1: Samples of sensitive prompts from HotpotQA and TriviaQA datasets.

Dataset	Original Prompt	Alternative Prompt	Original Answer	Alternative Answer	Correct Answer
HotpotQA	What American actor and comedian known for playing the role of Newman in Seinfeld, also stars in the series The Exes on TV Land?	What is the name of the American actor who played Newman in Seinfeld and appears in TV Land's comedy series The Exes	Wayne Knight	Jerry Seinfeld co-star	Wayne Knight
TriviaQA	At which city do the Blue and White Niles meet?	At which geographical location do the Blue and White Niles meet	Sudan's confluence	Khartoum	Khartoum

the stage for forthcoming studies in prompt engineering and the evaluation of LLM responses to prompt variations.

A common prompt engineering strategy is to ask the LLM itself to reformulate the prompt in a way that a more desirable output would be generated for the revised prompt by the LLM. While LLMs can autonomously generate different prompt variations, they cannot assess which variations are most effective, pointing to the fact that LLMs themselves are oblivious to the representation of optimal prompt variations. To establish a benchmark for this challenge, we formally introduce the *Prompt Sensitivity Prediction* task, which is concerned with assessing the effectiveness of a user prompt and its variations. We systematically curate our dataset based on the TriviaQA and HotpotQA [12,27], which consist of prompts that have deterministic and concise answers. To benchmark this task, we draw parallels with established tasks in text classification (TC) [8,5] and query performance prediction (QPP) [10,2,9], as they share resemblance with prompt sensitivity prediction. Our experiments show that such baselines fail to perform effectively for this task, underscoring the need for novel approaches tailored in particular for *prompt sensitivity prediction*. In summary, the contributions of our work in this paper include: **(1)** We define the prompt sensitivity prediction task, outlining the requirements and challenges involved in identifying effective prompts; **(2)** We introduce and publicly release a comprehensive dataset for Prompt Sensitivity Evaluation Task (`PromptSET`), focusing on slight prompt modifications that unveil LLM sensitivity to prompt variations[1]; and, **(3)** We benchmark the prompt sensitivity prediction task using state-of-the-art methods, including text classification, query performance prediction, and LLM-based baselines, to highlight the complexity of the proposed task.

2 Methodology

The Task Definition. Our proposed task of *Prompt Sensitivity Prediction* aims to predict whether a given prompt can be effectively fulfilled by the LLM whose response to the prompt would satisfy the users' information need. More specifically, given a prompt P with a specific information need I_P, we consider a set of similar prompts, denoted $\mathcal{P} = \{P' | Sim\langle P, P' \rangle > \tau \text{ and} I_P == I_{P'}\}$, where each variation P' shares the same information need I_P and maintains a similarity with P above a predefined threshold τ. These prompts $\{P'\}$ are designed to be only slightly modified versions of P, ensuring they still reflect the same information need of the user. The goal of this task is to predict, for a given prompt P_i, whether the LLM will generate a response that accurately respond to the underlying information need I_{Pi}.

The Dataset for the Task. To create the gold standard dataset for the prompt sensitivity task, we adopt a systematic process to generate prompt variations and evaluate their effectiveness as follows:

1. **Selecting Prompts**: We start by choosing a set of initial prompts, denoted as $\mathcal{P}$, where each prompt $p \in \mathcal{P}$ is seeking a distinct information need I_p.
2. **Generating Variations**: For each prompt p in the set, we use an LLM $\mathcal{L}$ to generate N variations $p' = \mathcal{L}(p \mid I_p = I_{p'})$. Here, $\mathcal{L}(p)$ denotes the process of generating

[1] `https://anonymous.4open.science/status/prompt-sensitivity-E6D8`

variations of prompt p, where each variation p' retains high semantic similarity with p, i.e., $(Sim\langle P, P'\rangle > \tau)$ and preserves the original information need I_p.

3. **Filtering Variations**: We process and filter out any generated variations p' that do not meet specific criteria for similarity and alignment with the original prompt p including LLM hallucinated content.
4. **LLM Response Generation**: For each prompt p and its variations $\{p' \in \mathcal{P}'\}$, we ask the LLM to respond to the prompt, denoted $a_p \in \mathcal{A}_\mathcal{P}$ for the original prompt and $a_{p'} \in \mathcal{A}'_\mathcal{P}$ for each variation.

The combination of $\mathcal{P} \cup \mathcal{P}'$ as well as their LLM generated answers $\mathcal{A}_\mathcal{P} \cup \mathcal{A}'_\mathcal{P}$ form the `PromptSET` dataset for this task.

Source Data. To build our dataset, we require a set of prompts that have human annotated answers available as well as having reliable evaluation with deterministic results. Therefore, we selected two widely-used question-answering datasets, TriviaQA [12] and HotpotQA datasets [27] that meet these requirements. TriviaQA is a reading comprehension dataset containing over 650K question-answer-evidence triples. The questions are on average 14 words. Each question has a collection of accepted answers including a list of aliases and normalized version of the answers; the majority of which are specific and short [12]. Furthermore, HotpotQA is a large-scale question-answering dataset consists of 113k training question-answer pairs. Unlike TriviaQA, HotpotQA includes complex multi-hop and comparison questions that require reasoning across multiple documents to answer accurately [27].

In our work, we randomly sampled 12K questions from the train set of each of these datasets paired with their provided answers. After removing questions that were less than 4 words or more than 40 words, we were able to obtain 11,469 unique questions and their respective answers from these datasets. We split these questions using a 70-30 ratio for training and testing, i.e., 8,028 and 3,441 questions, respectively. We consider each of these questions to be prompts that would be submitted to an LLM for a response.

Generating Prompt Variations. For each of the 11,469 unique prompts in our dataset, we generate several prompt variations. The objective of this step to generate slight variations of the original prompt, each of which may or may not be satisfiable by the LLM. To generate prompt variations, we utilize two widely-used LLMs that have demonstrated strong performance across many downstream tasks, namely the pre-trained LLaMA 3.1 with 8B parameters [18] and Mistral-nemo [11]. We chose open-source models to ensure reproducibility and facilitate further research in *prompt sensitivity*. We designed the instructions for the LLMs to be as clear and straightforward as possible, similar to those intended for human use. The primary goal was to generate variations that retain the same semantics of the original prompt, instructing the model to produce a rephrased prompt that does not answer the question directly but maintains the same information need and semantic content [2]. However, while instructed explicitly, LLMs occasionally deviate from the instructions due to hallucination. To address this, we filter out prompts that did not have at least nine valid variations generated and excluded prompts with fewer than four terms to maintain quality and consistency across the dataset. At the end, our dataset consists of 11,469 prompts, each with 9 different variations, resulting in

[2] Due to space constraints, we have provided the full set of instructions in our GitHub repository.

Table 2: Results of baselines on PromptSET.

	Category	Method	PromptSET-TriviaQA				PromptSET- HotPotQA			
			Accuracy	F1	Recall	Precision	Accuracy	F1	Recall	Precision
Mistral Answers	LLM-Based	Mistral	0.5045	0.5858	0.7743	0.4711	0.3735	0.2005	0.6912	0.1173
		LLaMA	0.4656	0.6239	0.9798	0.4577	0.1696	0.2050	0.9419	0.1150
	Text Classification	BERT	0.660	0.659	0.620	0.654	0.526	0.360	0.017	0.813
	Specificity-based QPP	CC	0.506	0.453	0.452	0.454	0.549	0.209	0.524	0.130
		DC	0.484	0.448	0.463	0.434	0.565	0.199	0.475	0.126
		IEF	0.505	0.462	0.469	0.455	0.535	0.204	0.526	0.127
		PageRank	0.481	0.444	0.458	0.431	0.533	0.153	0.370	0.096
	Supervised QPP	BERTPE	0.648	0.627	0.644	0.611	0.710	0.318	0.594	0.217
LLaMA Answers	LLM-Based	Mistral	0.5160	0.6045	0.7704	0.4974	0.3731	0.1978	0.6930	0.1153
		LLaMA	0.4940	0.6507	0.9818	0.4866	0.1674	0.2013	0.9408	0.1127
	Text Classification	BERT	0.664	0.664	0.651	0.650	0.532	0.377	0.034	0.808
	Specificity-based QPP	CC	0.500	0.463	0.449	0.478	0.545	0.199	0.507	0.123
		DC	0.484	0.464	0.465	0.463	0.562	0.190	0.462	0.120
		IEF	0.510	0.482	0.475	0.489	0.535	0.202	0.529	0.125
		PageRank	0.482	0.461	0.461	0.461	0.534	0.151	0.371	0.094
	Supervised QPP	BERTPE	0.659	0.651	0.646	0.656	0.710	0.314	0.596	0.213

114,690K variations. We ran each of these prompts (the original prompt and its nine variations) against the LLMs and generated an answer for each. We then compared the produced answer against the expected answer in the TriviaQA and HotpotQA datasets. Each prompt or prompt variations were labeled as being answerable by the LLMs depending on the answer they generated and whether it aligned with the expected answer. On this basis, the objective of the *Prompt Sensitivity Prediction* task is to predict whether an LLM can correctly answer an input prompt.

Establishing Baselines. To benchmark this task, we identify three types of tasks from the literature that may be applicable to prompt sensitivity prediction:

(1) We employ LLMs directly by asking them to self-assess their ability to predict whether they can accurately answer a given prompt or not [26]. This approach, which we refer to as the LLM self-evaluation baseline, relies on the model's internal confidence in its responses. The prompts and instructions used for this baseline can be found in our GitHub repository for reproducibility.

(2) We further treat prompt sensitivity prediction as a text classification task. We train a text classifier on our dataset's training set and evaluate it on the test set to predict whether the LLM's response to a prompt will meet users' information need. We used the text classifiers implemented in [23].

(3) Prompt sensitivity prediction is also conceptually related to the task of query performance prediction [21] whose goal is to estimate the quality of retrieved documents in response to a user query. For our task, we adapt QPP methods to predict whether a prompt will yield a correct response from an LLM or not. QPP methods typically fall into pre-retrieval and post-retrieval categories. However, since generative settings do not produce traditional "retrieved lists", only pre-retrieval QPP methods are applicable in our context. Furthermore, collection-dependent QPP methods [10] are also not applicable due to their dependence on a document corpus which is not available when using an LLM for generating a response. Thus, we adopted BERT-PE [14], a SOTA pre-retrieval QPP model that uses contextualized embeddings to learn query performance directly. Additionally, we considered the neural embedding specificity-based QPP metrics to assess prompt sensitivity [3,2,1], namely Closeness Centrality (CC), Degree Centrality (DC), PageRank , and Inverse Edge Frequency (IEF), all measured on the ego network of the query terms in the embedding space. We note that since the output of QPP methods is a scalar value, inspired by previous studies [6,5,17], we convert them to binary by classifying values above and below the mean of the data.

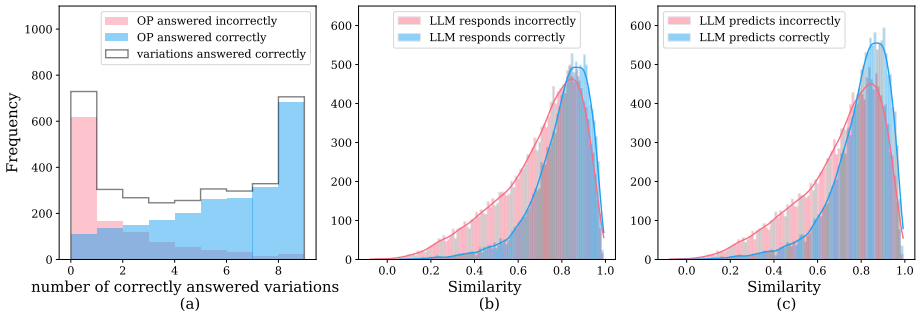

Fig. 1: (a) A histogram of correctly answered variations when the original prompt yields a correct response (in blue) or an incorrect response (in red), along with the total count of correctly answered variations (black step). Figures (b) and (c) display histograms for correct (blue) and incorrect (red) responses and predictions, respectively.

3 Experiments and Findings

Baseline Performance. We analyze the performance of the of baselines on the `PromptSET` test set. Table 2 presents the results for predicting whether each prompt variation could be correctly answered using the LLaMA and Mistral. We report results in terms of accuracy, F1, recall, and precision. As shown in the table, the specificity-based QPP methods (i.e., CC, DC, IEF, and PageRank) perform the lowest among the baselines. Since QPP methods were not specifically designed for prompt sensitivity prediction, their performance is relatively weak on both datasets. We hypothesize that the specificity levels across the original prompts and their variations are too similar, making it challenging for specificity metrics to effectively distinguish between different levels of prompt specificity. On the other hand, BERT-PE demonstrates higher effectiveness in determining whether a prompt can be answered correctly. BERT-PE, which is supervised QPP method, shows competitive performance to text classification-based methods on `PromptSET`-TriviaQA and also outperforms other baselines significantly on `PromptSET`-HotpotQA. This suggests that *supervised* QPP methods might be well-suited for the prompt sensitivity prediction task. We finally note the performance of LLM self-evaluation baseline. This baseline shows reasonable performance on TriviaQA but lacks consistency on HotpotQA. This is inverse to the performance of BERT-PE, indicating that these methods do not show stable performance across different prompt subsets.

Impact of original prompt correctness on variation answerability[3]**.** Here, we aim to investigate whether an LLM's ability to answer the original prompt influences its ability to answer other variations. Specifically, if the LLM can answer the original prompt, does this increase the likelihood of correctly answering the variations as well? To this end, Figure 1(a), marked with a black line, presents histograms of prompts showing the frequency of correctly answered variations out of a total of 10. In this figure, the x-axis represents the count of correctly answered variations for each prompt, grouping prompts by the number of successful variations. The distribution appears roughly balanced rather than long-tailed, indicating that `PromptSET` includes prompts with diverse answerability across reformulated variations, from those with only one answerable variation to those where all variations are answerable. In addition, in Figure 1(a), we further break down the results based on whether the original prompt was answered correctly or not. We observe that when the original prompt is answered correctly (shown in the blue histogram), there is a higher number of variations that also yield correct answers, reflected by an ascending

[3] Due to space constraints, we report the full findings in our GitHub repo.

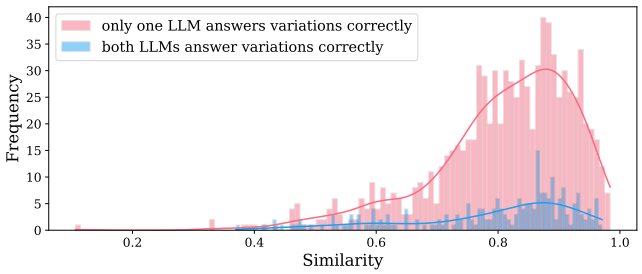

Fig. 2: Distribution of answerability of variations of the questions in PromptSET test sets which both LLMs failed to answer correctly.

pattern in the histogram. Conversely, when the original prompt is answered incorrectly, most of the prompts have only one out of the 10 variations answered correctly, displaying a descending pattern in the red histogram.

Impact of similarity to the original prompt. Here, we aim to investigate the impact of variation similarity to the original prompt on the predictability of LLM responses. Specifically, we ask whether a variation that is more similar to the original prompt has a higher likelihood of generating a correct answer, while less similar variations may lead to lower predictability. To test this hypothesis, we conducted experiments, as shown in Figure 1 (b) and (c).We present the distribution of correctly and incorrectly *answered prompts* and in Figure 1(b), and *predicted responses* in Figure 1(c), based on similarity to the original prompt. Similarity is measured using the cosine similarity of the embedded representations of prompt-variation pairs, calculated with MiniLM, a model known for its strong performance in various NLP and IR tasks [25]. We observe that when a variation closely resembles the original prompt, it is more likely to generate both correct responses and accurate predictions of answerability. This suggests that the model may have encountered this data points before, indicating a strong bias toward its training data and reduced generalizability to less familiar or novel prompt formulations.

Impact of choice of LLM on variation answerability. We further explore whether prompt reformulation can enhance the effectiveness of an LLM. To investigate this, we first filter out questions from the PromptSET test set for which both LLMs, namely LLaMA and Mistral, failed to answer the original prompt correctly. Next, we examine the variations of these questions to see if an alternative prompt allows either LLM to provide a correct answer. The results are shown in Figure 2. For each sample in this figure, both LLMs failed to answer the original prompt correctly. However, in the red cases, at least one of the two LLMs succeeded in answering a variation correctly, while in the blue cases, both LLMs provided correct answers to the variation. This highlights the potential of prompt reformulation as a strategy. We conclude that PromptSET can serve as a valuable resource for prompt reformulation, helping transform an unanswerable prompt into an answerable one through LLM-driven reformulation.

4 Concluding Remarks

This paper investigates the sensitivity of LLMs to prompt variations by introducing the Prompt Sensitivity Prediction task and the PromptSET dataset, based on TriviaQA and HotpotQA. We generate variations of different questions and examine the sensitivity of various LLMs to these variations, all of which share the same underlying information need. Our benchmarking results reveal that existing methods do not fully capture the complexities of prompt sensitivity. These findings underscore the need for further research into prompt variation sensitivity, particularly in developing methods to help users generate more reliable prompts.

References

1. Arabzadeh, N., Zarrinkalam, F., Jovanovic, J., Al-Obeidat, F., Bagheri, E.: Neural embedding-based specificity metrics for pre-retrieval query performance prediction. Information Processing & Management **57**(4), 102248 (2020)
2. Arabzadeh, N., Zarrinkalam, F., Jovanovic, J., Bagheri, E.: Neural Embedding-Based Metrics for Pre-retrieval Query Performance Prediction, p. 78–85. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-45442-5_10, http://dx.doi.org/10.1007/978-3-030-45442-5_10
3. Arabzadeh, N., Zarrinkalam, F., Jovanovic, J., Bagheri, E.: Geometric estimation of specificity within embedding spaces. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. p. 2109–2112. CIKM '19, ACM (Nov 2019). https://doi.org/10.1145/3357384.3358152, http://dx.doi.org/10.1145/3357384.3358152
4. Bhargava, A., Witkowski, C., Shah, M., Thomson, M.W.: What's the magic word? a control theory of llm prompting. ArXiv **abs/2310.04444** (2023). https://doi.org/10.48550/arXiv.2310.04444
5. Collins-Thompson, K., Bennett, P.N.: Predicting Query Performance via Classification, p. 140–152. Springer Berlin Heidelberg (2010). https://doi.org/10.1007/978-3-642-12275-0_15, http://dx.doi.org/10.1007/978-3-642-12275-0_15
6. Faggioli, G., Ferro, N., Muntean, C.I., Perego, R., Tonellotto, N.: A geometric framework for query performance prediction in conversational search. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1355–1365. SIGIR '23, ACM (Jul 2023). https://doi.org/10.1145/3539618.3591625, http://dx.doi.org/10.1145/3539618.3591625
7. Feng, Z., Zhou, H., Zhu, Z., Qian, J., Mao, K.: Unveiling and manipulating prompt influence in large language models. ArXiv **abs/2405.11891** (2024), https://api.semanticscholar.org/CorpusID:269922034
8. Gasparetto, A., Marcuzzo, M., Zangari, A., Albarelli, A.: A survey on text classification algorithms: From text to predictions. Information **13**(2), 83 (2022)
9. Hambarde, K.A., Proença, H.: Information retrieval: Recent advances and beyond. IEEE Access **11**, 76581–76604 (2023)
10. Hauff, C., Hiemstra, D., de Jong, F.: A survey of pre-retrieval query performance predictors. In: CIKM (2008)
11. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023), https://arxiv.org/abs/2310.06825
12. Joshi, M., Choi, E., Weld, D., Zettlemoyer, L.: TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In: Barzilay, R., Kan, M.Y. (eds.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1601–1611. Association for Computational Linguistics, Vancouver, Canada (Jul 2017). https://doi.org/10.18653/v1/P17-1147, https://aclanthology.org/P17-1147
13. Kamalloo, E., Dziri, N., Clarke, C., Rafiei, D.: Evaluating open-domain question answering in the era of large language models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics (2023). https://doi.org/10.18653/v1/2023.acl-long.307, http://dx.doi.org/10.18653/v1/2023.acl-long.307

14. Khodabakhsh, M., Zarrinkalam, F., Arabzadeh, N.: BertPE: A BERT-Based Pre-retrieval Estimator for Query Performance Prediction, p. 354–363. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-56063-7_27, http://dx.doi.org/10.1007/978-3-031-56063-7_27
15. Lo, L.S.: The art and science of prompt engineering: A new literacy in the information age. Internet Reference Services Quarterly **27**, 203 – 210 (2023). https://doi.org/10.1080/10875301.2023.2227621
16. Loya, M., Sinha, D., Futrell, R.: Exploring the sensitivity of llms' decision-making capabilities: Insights from prompt variations and hyperparameters. In: Findings of the Association for Computational Linguistics: EMNLP 2023. p. 3711–3716. Association for Computational Linguistics (2023). https://doi.org/10.18653/v1/2023.findings-emnlp.241, http://dx.doi.org/10.18653/v1/2023.findings-emnlp.241
17. Meng, C., Arabzadeh, N., Aliannejadi, M., de Rijke, M.: Query performance prediction: From ad-hoc to conversational search. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2583–2593. SIGIR '23, ACM (Jul 2023). https://doi.org/10.1145/3539618.3591919, http://dx.doi.org/10.1145/3539618.3591919
18. Meta, L.T.A..: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
19. Mu, Y., Wu, B., Thorne, W., Robinson, A., Aletras, N., Scarton, C., Bontcheva, K., Song, X.: Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. ArXiv **abs/2305.14310** (2023). https://doi.org/10.48550/arXiv.2305.14310
20. Murr, L., Grainger, M., Gao, D.: Testing llms on code generation with varying levels of prompt specificity. ArXiv **abs/2311.07599** (2023). https://doi.org/10.48550/arXiv.2311.07599
21. Poesina, E., Costache, A.V., Chifu, A.G., Mothe, J., Ionescu, R.T.: Pqpp: A joint benchmark for text-to-image prompt and query performance prediction (2024), https://arxiv.org/abs/2406.04746
22. Raj, H., Gupta, V., Rosati, D., Majumdar, S.: Semantic consistency for assuring reliability of large language models. ArXiv **abs/2308.09138** (2023). https://doi.org/10.48550/arXiv.2308.09138
23. Rajapakse, T.C., Yates, A., de Rijke, M.: Simple transformers: Open-source for all. p. 7 pages (2024)
24. Sclar, M., Choi, Y., Tsvetkov, Y., Suhr, A.: Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. arXiv preprint arXiv:2310.11324 (2023)
25. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
26. Yan, Z.: Evaluating the effectiveness of llm-evaluators (aka llm-as-judge). eugeneyan.com (Aug 2024), https://eugeneyan.com/writing/llm-evaluators/
27. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., Manning, C.D.: HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2369–2380. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1259, https://aclanthology.org/D18-1259
28. Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., Yang, L., Ye, W., Gong, N., Zhang, Y., Xie, X.: Promptbench: Towards evaluating the robustness of large language mod-

els on adversarial prompts. ArXiv **abs/2306.04528** (2023). https://doi.org/10.48550/arXiv.2306.04528
29. Zhuo, J., Zhang, S., Fang, X., Duan, H., Lin, D., Chen, K.: Prosa: Assessing and understanding the prompt sensitivity of llms (2024), https://arxiv.org/abs/2410.12405