

# E-Bench: Towards Evaluating the Ease-of-Use of Large Language Models

Zhenyu Zhang<sup>†\*</sup>, Bingguang Hao<sup>‡\*</sup>, Jinpeng Li<sup>‡</sup>, Zekai Zhang<sup>‡</sup>, Dongyan Zhao<sup>‡</sup>

<sup>†</sup>Baidu Inc.

<sup>‡</sup>Wangxuan Institute of Computer Technology, Peking University

zhangzhenyu07@baidu.com, lijinpeng@stu.pku.edu.cn

## Abstract

Most large language models (LLMs) are sensitive to prompts, and another synonymous expression or a typo may lead to unexpected results for the model. Composing an optimal prompt for a specific demand lacks theoretical support and relies entirely on human experimentation, which poses a considerable obstacle to popularizing generative artificial intelligence. However, there is no systematic analysis of the stability of LLMs in resisting prompt perturbations in real-world scenarios. In this work, we propose to evaluate the ease-of-use of LLMs and construct E-Bench, simulating the actual situation of human use from synonymous perturbation (including paraphrasing, summarization, and colloquialism) and typographical perturbation (such as typing). On this basis, we also discuss the combination of these two types of perturbation and analyze the main reasons for performance degradation. Experimental results indicate that with the increase of model size, although the ease-of-use are significantly improved, there is still a long way to go to build a sufficiently user-friendly model. The dataset is now available at <https://github.com/zzysay/E-Bench>.

## 1 Introduction

Large language models (LLMs) have swept across the entire natural language processing (NLP) fields, revolutionizing many domains and attracting unprecedented attention. Benefiting from the powerful instruction following and language generation capabilities, the application scenarios of modern LLMs are gradually expanding and being used as productivity tools. By exploring natural language exquisite prompts, LLMs could handle various demands, such as question answering, data construction, and intelligent agent (Jakesch et al., 2023; Wu and Hu, 2023; Xu et al., 2023).

Yet, LLMs are prompt-driven, the performance of downstream tasks significantly depends on the

\* Equal contribution

### Original Prompt

You are given a Python command and are required to convert it into Ruby. `print("Hello World!")`

### Synonymous Perturbation

Your task is to translate a given Python command into Ruby. `print("Hello World!")` *paraphrasing*

Convert a Python command into Ruby. `print("Hello World!")` *summarization*

Next, I will provide a Python command. Could you help me convert it to Ruby. `print("Hello World!")` *colloquialism*

### Typographical Perturbation

You are given a python commad and are required to convert it into Ruuby. `print("Hello World!")` *typing attack*

Figure 1: The prompt perturbations in E-Bench, which simulate the actual situations of humans using LLMs.

quality of the prompt used to steer the model. Inappropriate prompts are insufficient to meet target tasks, and most effective prompts are handmade by humans (Kaddour et al., 2023; Dubey et al., 2024; Sclar et al., 2024). In this process, there is no reliable theoretical basis or prior indication of what kind of prompt is optimal for a specific task. Users write a prompt, verify its validity, and refine the prompt iteratively, like opening a “surprise box”. It even gives birth to a new field, prompt engineering, which attempts to explore how to write prompts to improve the efficiency of direct interaction between humans and deep generative models (Cao et al., 2023; Chen et al., 2023; Gu et al., 2023).

The sensitivity of LLMs to prompt is a huge obstacle in their usage. For example, no one knows what kind of performance a good prompt on ChatGPT will have on GPT-4. In contrast, human conversational abilities are general, flexible, and robust, different expressions with the same semantics will not confuse different people (Peng et al., 2021). Recently, there have been some works on evaluating the robustness of LLMs (Wang et al., 2023b; Dumpala et al., 2024; Mousavi et al., 2024).

However, they mainly focus on the perspective of out-of-distribution problems or adversarial prompts of NLP tasks, rather than prompt perturbations in the most commonly used conversational scenarios, which is precisely the key to pushing LLMs away from tedious prompt engineering. In other words, they still have not answered well whether a model is convenient enough for human use. A robust LLM should be able to consistently provide accurate and relevant responses across a series of synonymous prompts, which helps improve the usability.

In this paper, we fuel research in this direction by introducing E-Bench, a comprehensive benchmark designed for understanding the impact of prompt perturbation and evaluating the ease-of-use of LLMs. Here, an ideal model is expected to handle synonymous and even typographical prompts like humans swimmingly. Starting from existing evaluation set AlpacaEval (Li et al., 2023), we divide it into four parts according to data characteristics and perturb the prompts through paraphrasing, summarization, colloquialism, and typing attacks. Figure 1 provides an illustration for each perturbation. Specifically, we first perturb each prompt with a series of automatic tools, then manually review to ensure that the prompts before and after perturbation have similar semantics and do not affect human understanding. Performance drop is used as evaluation metric, where the closer the performance on perturbed prompts is to original performance, the higher ease-of-use of the model.

With E-Bench, we conduct experiments to evaluate the ease-of-use of 6 representative LLMs, including Llama2-chat models (Touvron et al., 2023), Vicunas (Chiang et al., 2023), and GPTs (Achiam et al., 2023). The results demonstrate that all models experience varying performance degradation after prompt perturbation, and larger models perform better under synonymous perturbation, while there is no clear scaling law for model size and performance degradation under typing attack. Overall, improving the ease-of-use of LLMs is an urgent research topic. Furthermore, additional analysis reveals the impact of training data on the specific aspect of ease-of-use. We hope that E-Bench could provide a stepping stone for the popularization of general artificial intelligence (GAI).

## 2 E-Bench

Inspired by actual scenarios (Peng et al., 2021), we first consider different expressions of the same se-

Dataset	#Num	Avg. Len.		
		Orig.	Pert.	
AlpacaEval (Li et al., 2023)	805	28.83	-	
E-Bench	Paraphrasing	200	25.22	24.55
	Summarization	200	47.21	32.27
	Colloquialism	200	15.66	20.26
	Typing Attack	205	27.27	27.29
	Typing Attack*	805	26.10	26.71

Table 1: Statistics of E-Bench, in which typing attack has its own evaluation set and could also be combined with other perturbations (marked by \*).

mantics, including *paraphrasing*, *summarization*, and *colloquialism*. In practical terms, 1) paraphrasing is the simplest form of expression conversion without significantly impacting prompt length. 2) summarization is a quality in which users express their intents using fewer words to be concise. 3) colloquialism converts written style prompts into spoken forms, usually with modal particles and increased length. Furthermore, we simulate the typo issues during prompt input and carry out char/word-based attacks on the original input or after each synonymous perturbation above.

### 2.1 Construction

To perform the ease-of-use evaluation, we reconstruct AlpacaEval (Li et al., 2023), an widely-used benchmark for instruction-following models with 805 examples. The dataset is divided into four parts. First, 205 instances are randomly selected for typing attack. Then, the remaining 600 instances are arranged according to prompt length and divided into three parts: the shortest part is used for colloquialism, the longest part is for summarization, and the part with moderate length is for paraphrasing<sup>1</sup>. **Paraphrasing.** It exists widely because different users have different language habits. Here, we focus on assessing the stability of LLMs when dealing with varying inputs with the same semantics. **Summarization.** There is no consensus on whether prompt should be detailed or concise. In this setting, we assess model effectiveness when users use fewer words, that is, considering the impact of input length on model performance.

**Colloquialism.** The difference between spoken language and written language lies in different expressions and possible disfluency. In a conversation

<sup>1</sup>We use GPT-4 to assist in candidate synonymous perturbations, i.e., paraphrasing, summarization, and colloquialism. Appendix A provides more details about dataset construction.

Models	Paraphrasing			Summarization			Colloquialism			Overall		
	Orig.	Pert.	$\Delta$	Orig.	Pert.	$\Delta$	Orig.	Pert.	$\Delta$	Orig.	Pert.	$\Delta$
Llama 2-chat (7b)	71.00	62.00	9.00	69.00	62.31	6.69	72.50	56.50	16.00	70.83	60.27	10.56
Llama 2-chat (13b)	81.00	78.00	3.00	76.38	74.37	2.01	79.00	69.00	10.00	78.79	73.79	5.00
Vicuna-v1.5 (7b)	77.50	70.00	7.50	73.62	67.50	6.12	73.25	72.11	1.14	74.79	69.87	4.92
Vicuna-v1.5 (13b)	80.50	74.50	6.00	79.75	81.25	-1.50	82.25	79.65	2.60	80.83	78.47	2.37
GPT-3.5	85.93	84.00	1.93	92.21	84.50	7.71	89.75	83.25	6.50	89.27	83.92	5.38
GPT-4	95.48	94.25	1.23	94.00	90.50	3.50	92.25	89.75	2.50	93.91	91.50	2.41
Overall	81.90	77.13	4.78	80.83	76.74	4.09	81.50	75.04	6.46	81.41	76.30	5.11

Models	Paraphrasing		Summarization		Colloquialism		Typing Attack			Overall		
	+Att.	$\Delta$	+Att.	$\Delta$	+Att.	$\Delta$	Orig.	+Att.	$\Delta$	Orig.	+Att.	$\Delta$
Llama 2-chat (7b)	33.17	28.83	30.50	31.81	33.84	22.66	73.28	59.50	13.78	63.52	39.25	24.27
Llama 2-chat (13b)	39.50	38.50	42.50	31.87	38.50	30.50	75.85	60.98	14.87	74.31	45.37	28.94
Vicuna-v1.5 (7b)	42.50	27.50	41.50	26.00	49.50	22.61	75.36	61.71	13.65	71.24	48.80	22.44
Vicuna-v1.5 (13b)	44.50	30.00	45.00	36.25	50.25	29.40	84.30	71.50	12.80	79.93	52.81	27.11
GPT-3.5	50.00	34.00	51.50	33.00	58.75	24.50	93.41	82.61	10.80	86.29	60.72	25.57
GPT-4	58.00	36.25	63.50	27.00	68.00	21.75	93.96	88.65	5.31	92.12	69.54	22.58
Overall	44.61	32.51	45.75	30.99	49.81	25.24	82.69	70.83	11.87	77.90	52.75	25.15

Table 2: Main results on E-Bench, where ‘Orig.’, ‘Pert.’, and ‘+Att.’ refer the performance of on original, perturbed, and attacked prompts, respectively, ‘ $\Delta$ ’ indicates the preference drop. See Appendix B for statistical analysis.

system or chatbot, LLM may need to handle spoken instructions from users through an ASR module.

**Typing Attack.** Typing errors are widespread in practice scenarios; one may mistakenly touch the surrounding keys when typing, or optical character recognition (OCR) errors may result in similar-looking characters. To simulate such cases, we use GAN (Zhao et al., 2018), TextBugger (Li et al., 2018), PWWS (Ren et al., 2019) and perform attacks at a frequency of once every ten words.

Finally, we manually review all data to ensure semantic invariance before and after transformation or no visually significant differences. For accuracy reasons, the original prompt is split into *instruction* and *input* following Wang et al. (2022), and only the *instruction* part is disturbed.

## 2.2 Statistics

Table 1 summarizes the statistics of E-Bench. Since different models are equipped with different tokenizers, we use the number of words to measure prompt length. It can be observed that both paraphrasing and typing attacks do not significantly affect prompt length. Summarization naturally shortens the length by more than 30%, while colloquialism substantially increases the number of words due to the increase or pause of modal particles. Nevertheless, each part’s prompt length is relatively balanced and does not exceed other ranges.

## 3 Experiments

In this section, we select six models covering various types and scales to evaluate the ease-of-use of LLMs, including (1) *Llama 2-chat (7b/13b)* (Touvron et al., 2023), which undergo the pre-training, instruction-tuning, as well as reinforcement learning stages. (2) *Vicuna-v1.5 (7b/13b)* (Chiang et al., 2023), which are instruction-tuning models based on Llama 2. (3) *GPT-3.5* (Ouyang et al., 2022) and *GPT-4* (Achiam et al., 2023), the most representative LLMs. For a prompt  $p$  in E-Bench, we find its original prompt  $p'$  in AlpacaEval, input them to the test model, and generate corresponding responses  $r'$ . Next, we use GPT-4 to evaluate their winning rate to the reference response  $\hat{r}$  generated by *text\_davinci\_003*, and compare the performance drop  $\Delta = wr(r', \hat{r}) - wr(r, \hat{r})$  as evaluation metric. That is, the smaller the performance drop, the better the ease-of-use of the model.

### 3.1 Results

Table 2 presents the benchmark results, from which one can summarize a series of interesting observations<sup>2</sup>. (1) *All models have a considerable degree of performance degradation on E-Bench.* Among them, *GPT-4* has the best overall performance, comparable to *Vicuna-v1.5 13b*, and is sensibly

<sup>2</sup>Our analysis shows that the automatic metric has a substantially positive Pearson score (0.68) with human annotation.

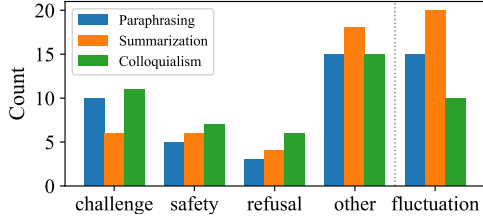


Figure 2: The error analysis of *Llama 2-chat 7b*. We also provide the amount that improves after perturbation as a reference for evaluation “fluctuation”.

ahead of *Llama 2-chat* models in all aspects. This phenomenon indicates that reinforcement learning (or human-value alignment) may not be the necessary way to improve the ease-of-use of LLMs. (2) *The larger the model, the better the stability (roughly) under synonymous perturbation.* However, there is no evident trend in model size when fighting against typing attacks, and the reasons for these abnormal behaviors could trigger interests for future research. (3) *Different models have different areas of expertise.* For example, *GPTs* perform well in paraphrasing, while *Vicunas* perform well in colloquial setting. We attribute this to the influence of training data, because *Vicuna* is trained on user-shared conversations collected from ShareGPT. Note that *Llama 2-chat* models (especially *Llama 2-chat 7b*) exhibit significant degeneration under various perturbations, we provide a detailed analysis of this in the next section.

### 3.2 Analysis

*Llama 2-chat 7b* encounters unexpected decreases in synonymous perturbations. To explore the reason behind, we systematically review samples with preference reversal after prompt perturbations (i.e., the response on original prompt is better than the reference, while the response on perturbed prompt becomes worse). The results have significant regularity, as shown in Figure 2, where “challenge” refers to that LLMs believe there is nonfactual content in prompt, “safety” indicates that LLMs do not answer the prompt due to ethical or biased consideration, “refusal” denotes that LLMs consider the query incomplete and requests more information, and “other” includes a series of issues such as inaccurate responses and fluctuating evaluations<sup>3</sup>. It can be observed that among the three perturbations, “challenge” is the most influential factor, followed by “safety”. In constant,

<sup>3</sup>see also Appendix D for examples of each error category.

*GPT-4* does not have such a phenomenon. We speculate that there are two reasons: *Llama 2-chat* models might have a certain degree of over-fitting to the original AlpacaEval, or the safety mechanisms have a severe problem of over-recall. As for typing attack, the main factor causing performance drop is the inability of LLM to accurately understand the requirements due to semantic offset, consistent with previous study (Zhu et al., 2023).

## 4 Related Work

In this paper, we focus on the ease-of-use of LLMs, which is somewhat similar to robustness in the traditional sense, that is, studying the stability of a system when facing unexpected inputs (Raj et al., 2023; Lu et al., 2024; Wei et al., 2024). Researchers in this field mainly focus on the stability of language models in out-of-distribution (OOD) problems and adversarial robustness (Peng et al., 2021; Chang et al., 2023). Specifically, Yang et al. (2022) propose to assess OOD robustness by extending the GLUE dataset. Wang et al. (2023b) evaluate ChatGPT and other LLMs from both adversarial and OOD perspectives with existing benchmarks including AdvGLUE, ANLI, and DDXPlus. For adversarial robustness, a rising direction specializes in jailbreak prompts to bypass the security defense system of LLMs (Deng et al., 2023; Liu et al., 2023; Wang et al., 2023a). However, these works ignore the fact that mass users may not maliciously attack LLMs through prompts in the standard usage scenario. It is just that different people have different habits of language expression, which in turn leads to different prompts styles (e.g., spoken or written, succinct or exhaustive). PromptBench (Zhu et al., 2023) is an exploration of this branch to a certain extent, highlighting the robustness’s importance when facing adversarial inputs at character, word, sentence, and semantics levels. Still, it is limited to NLP tasks rather than the real interactive dialogue scene. It is the essence that makes this paper different from previous works, we extend prompt perturbation to a more universal level and propose the evaluation setting of ease-of-use of LLMs.

## 5 Conclusion

We propose the concept of ease-of-use of LLMs and construct the E-Bench benchmark from synonymous perturbation and typographical perturbation. We thoroughly evaluate recent representative models and conclude that the ease-of-use of all cur-

rent models is not very satisfactory, and training data might be the secret key to improving ease-of-use. E-Bench will be open-sourced to serve as a foundational tool for easy-to-use research.

## Limitations

Evaluating the ease-of-use of LLMs is a new direction in the era of GAI, and previous work only focuses on partial aspects of it. It is hoped that ideal LLMs could engage in conversation as general, flexible, and robust as humans. To the best of our knowledge, E-Bench presents the first step to fill this gap. One limitation of this paper is that we follow the methodology of evaluating instruction-tuning models and only assess the ease-of-use on single-turn conversations, while multi-turn conversation is a more challenging field that may generate more forms of transformation. We are very interested in this and leave it for future work.

## Ethical Consideration

In this paper, we present a human-annotated benchmark, E-Bench, to evaluate the ease-of-use of various LLMs in the era of generative artificial intelligence. Starting from a widely used benchmark, AlpacaEval, we use GPT-4’s official API or open-source tools to obtain primary transformations. The authors conduct manual cross-reviews to ensure that all perturbations are acceptable. For the sake of energy saving, we do not blindly expand the amount of evaluation data but maintain it at a scale comparable to current popular datasets, and we will open source the dataset and results used in experiments to reduce unnecessary carbon emissions.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Tingfeng Cao, Chengyu Wang, Bingyan Liu, Ziheng Wu, Jinhui Zhu, and Jun Huang. 2023. Beautiful-prompt: Towards automatic prompt engineering for text-to-image synthesis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1–11.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.
- Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. Unleashing the potential of prompt engineering in large language models: A comprehensive review. *arXiv preprint arXiv:2310.14735*.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality.
- Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. 2023. Masterkey: Automated jailbreak across multiple large language model chatbots. *arXiv preprint arXiv:2307.08715*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Sasstry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. 2024. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. *arXiv preprint arXiv:2406.11171*.
- Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. 2023. A systematic survey of prompt engineering on vision-language foundation models. *arXiv e-prints*, pages arXiv–2307.
- Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users’ views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15.
- Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.
- Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial text against real-world applications. *ArXiv*, abs/1812.05271.
- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. AlpacaEval: An automatic evaluator of instruction-following models.
- Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. 2023. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.
- Sheng Lu, Hendrik Schuff, and Iryna Gurevych. 2024. How are prompts different in terms of sensitivity? In *Proceedings of the 2024 Conference of the North*



- American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5833–5856.
- Seyed Mahed Mousavi, Gabriel Roccabruna, Simone Alghisi, Massimo Rizzoli, Mirco Ravanelli, and Giuseppe Riccardi. 2024. Are llms robust for spoken dialogues? *arXiv preprint arXiv:2401.02297*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Baolin Peng, Chunyuan Li, Zhu Zhang, Chenguang Zhu, Jinchao Li, and Jianfeng Gao. 2021. Raddle: An evaluation benchmark and analysis platform for robust task-oriented dialog systems. 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)*, pages 4418–4429.
- Harsh Raj, Vipul Gupta, Domenic Rosati, and Subhabrata Majumdar. 2023. Semantic consistency for assuring reliability of large language models. *arXiv preprint arXiv:2308.09138*.
- Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Annual Meeting of the Association for Computational Linguistics*.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv e-prints*, pages arXiv–2307.
- Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023a. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*.
- Jindong Wang, HU Xixu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Wei Ye, Haojun Huang, Xiubo Geng, et al. 2023b. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hananeh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.
- Sheng-Lun Wei, Cheng-Kuang Wu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Unveiling selection biases: Exploring order and token sensitivity in large language models. *arXiv preprint arXiv:2406.03009*.
- Yangjian Wu and Gang Hu. 2023. Exploring prompt engineering with gpt language models for document-level machine translation: Insights and findings. In *Proceedings of the Eighth Conference on Machine Translation*, pages 166–169.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.
- Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. 2022. Glue-x: Evaluating natural language understanding models from an out-of-distribution generalization perspective. *arXiv preprint arXiv:2211.08073*.
- Guoyang Zeng, Fanchao Qi, Qianrui Zhou, Tingji Zhang, Zixian Ma, Bairu Hou, Yuan Zang, Zhiyuan Liu, and Maosong Sun. 2021. Openattack: An open-source textual adversarial attack toolkit. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 363–371.
- Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *International Conference on Learning Representations (ICLR)*.
- Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.

## A Details of Data Construction

During data construction, we utilize GPT-4 to assist in completing initial conversion for synonymous perturbation, and the prompts are as follows.

### Paraphrasing

You are a proficient assistant, and your role is to craft paraphrases and provide alternative expressions for the content of a text or message. Your rewordings should preserve the original meaning using different vocabulary and conform to these rules:

- Extract and communicate the central idea of the original sentence.
- Honor the spirit of the original, effectively functioning as a question or request.
- Paraphrase the given sentence only, do not answer questions or give any response.

Note the length of sentences after paraphrasing, avoid the paraphrased sentence being longer than the original sentence only. Make the length of your paraphrased sentence determined by a random variation, with the original sentence's length serving as the mean of this variation.

<instruction>

### Summarization

You are an assistant with a helpful and precise nature, tasked with simplifying the semantics of given sentences. Your goal is to condense the input provided to you into a concise version while preserving the fundamental meaning of the sentence:

- Aim to retain the key elements and structure of the original sentence, incorporating as many words or expressions from the original sentence as possible in the simplified version.
- Ensure you preserve the original message and content of the sentence, while maintaining the perspective that the user is seeking an answer or solution from the assistant.
- Focus on condensing the sentence and making it more straightforward, while ensuring the fundamental meaning remains intact. Emphasize clarity, simplicity, and objectivity in your simplified version.

The simplified sentence should closely resemble the given sentence, working extremely well as questions or requests to the pre-conversion answer. It should only refer to the format of the original answer and not be influenced by its specific context.

<instruction>

### Colloquialism

You are a helpful and precise assistant tasked with creatively transforming the language style of given sentences into a conversational tone reflecting human expression's rich diversity. When you rephrase sentences:

- Strive for a Human Colloquial Style that is friendly, natural, varied, and easy to understand, mirroring how different people might casually express themselves.
- Employ a wide range of conversational tones, using diverse colloquial expressions, regional idioms, contractions, and relaxed sentence structures.
- Ensure you preserve the original message and content of the sentence, while maintaining the perspective that the user is seeking an answer or solution from the assistant.
- Focus solely on changing the sentence style to resemble a casual question or request from a human, rather than providing additional information or context in the answer.

Remember, it's important to avoid overusing any specific greeting or opener, such as 'Hey,' and instead, focus on providing a wide range of conversational options to reflect the diverse speaking habits of individuals.

<instruction>

With the initial results obtained from GPT-4, we manually review all data to ensure semantic invariance before and after transformation. During this review process, some excessively low-quality results are discarded, and the transformation of these data is manually performed. This operation may also alleviate the potential data bias of GPT-4.

For typographical perturbation (i.e., typing attack), we adopt OpenAttack (Zeng et al., 2021) and choose to use GAN (Zhao et al., 2018), TextBugger (Li et al., 2018), PWWS (Ren et al., 2019). Specifically, GAN performs sentence-level perturbations in the continuous space and thereby obtain adversaries that are legible, grammatical, and meaningfully similar to the input. TextBugger could effectively and efficiently generate utility-preserving adversarial texts with greedy word substitution and character manipulation. PWWS (Probability Weighted Word Saliency) introduces a new word substitution order determined by the word saliency and weighted by the classification probability.

On this foundation, we also perform a careful manual review to guarantee that there are no signif-

Dataset	#Num	Ins.Len.(Avg)	Ins.Len.(Max)	Ins.Len.(Min)	Pro.Len.(Avg)	Pro.Len.(Max)	Pro.Len.(Min)	
<i>Original Prompt</i>								
AlpacaEval (Li et al., 2023)	805	19.11	151	1	28.83	357	2	
E-Bench	Paraphrasing	200	14.22	20	10	25.22	276	10
	Summarization	200	34.23	151	<u>20</u>	47.21	332	10
	Colloquialism	200	7.64	10	1	15.66	357	3
	Typing Attack	205	20.34	142	1	27.27	142	2
	Typing Attack*	805	16.42	143	1	26.10	365	2
<i>Perturbed/Attcked Prompt</i>								
E-Bench	Paraphrasing	200	13.57	25	7	24.55	275	8
	Summarization	200	19.34	137	5	32.27	291	6
	Colloquialism	200	12.27	33	4	202.6	365	4
	Typing Attack	205	20.41	143	1	27.39	143	2
	Typing Attack*	805	16.52	143	1	26.71	365	2

Table 3: The maximum, minimum, and average length of E-Bench, where ‘Ins.Len.’, ‘Pro.Len’ refer the Instruction Length and Prompt Length, respectively. The underline marks the Instruction length used when splitting subset.

Dataset	#Num	Koala	OAAST	Vicuna	Self-Instruct	Helpful Base	
AlpacaEval (Li et al., 2023)	805	156	188	80	252	129	
E-Bench	Paraphrasing	200	33	44	27	75	21
	Summarization	200	42	58	11	64	25
	Colloquialism	200	31	52	16	46	55
	Typing Attack	205	50	34	26	67	28

Table 4: The prompt source of E-Bench. We reuse the statistics from AlpacaEval.

icant visual significant differences before and after perturbation, at least not affecting human understanding of the overall prompt semantics.

The prompt in AlpacaEval consists of an Instruction part and an Input part. For example, “*Write a short description about the given movie or series.*” is Instruction, and “*The Witcher (2019)*” is Input. To ensure the semantic consistency, we only perturb the Instruction part during perturbation. Considering that short prompts are not suitable for the summarization perturbation, after trying a series of methods, we ultimately chose to divide AlpacaEval according to the length of Instruction. Specifically, we calculated the number of words in Instruction, where the 200 samples with the maximum number of words are used for summarization, the 200 samples with the minimum number of words are used for colloquialism, and the 200 samples with the lowest number of words are used for paraphrasing. Table 3 shows the maximum, minimum, and average length of each subset. It can be seen that after summarization or colloquialism perturbations, there is a significant reduction or increase in prompt length, while paraphrasing and typing attack do not bring about drastic length changes.

The original AlpacaEval dataset is sampled from multiple publicly available datasets, such as Koala, OAAST, and Self-Instruct. In Table 4, we also pro-

vide the prompt sources of each subset for possible future analysis.

## B Details of Experimental Results

To confirm whether the automatic evaluation is reliable, we randomly select 20 samples for each type of perturbation, and then manually label the results on *Llama-2-chat 7b*, *Vicuna-1.5 7b*, *GPT-3.5*, and *GPT-4*, that is, there are 560 annotations in total, 140 annotations for each model, and 80 annotations for each perturbation. In particular, the annotation team consists of three PhD students in NLP field, if there are differences between two annotations on a sample, the third annotation will be introduced to determine the final decision. From Table 5, we observe significant consistency and correlation between automatic and manual annotations, and there is no significant difference between manual conclusions and automatic ones. Overall, apart from length preference, the annotation of GPT-4 is quite accurate, which is precisely the key reason why we chose AlpacaEval as the evaluation schema.

During the annotation process, we also find that when the model could accurately understand the prompt and start answering in the early stages of generation, there is no significant difference in the quality of responses before and after perturbation. The performance degradation usually comes from



Models	Paraphrasing		Summarization		Colloquialism		Typing Attack	Overall
	Orig.	Pert.	Orig.	Pert.	Orig.	Pert.	-	-
Llama 2-chat (7b)	0.70 / 0.49	0.80 / 0.42	0.90 / 0.78	0.85 / 0.70	0.85 / 0.73	0.85 / 0.78	0.90 / 0.50	0.84 / 0.66
Vicuna-v1.5 (7b)	0.90 / -	0.95 / 0.93	0.90 / 0.75	1.00 / 1.00	0.80 / 0.78	0.80 / 0.64	0.80 / 0.71	0.88 / 0.77
GPT-3.5	0.90 / 0.73	0.90 / 0.81	0.90 / 0.75	0.85 / 0.68	0.95 / 0.87	0.85 / 0.71	0.90 / 0.74	0.89 / 0.76
GPT-4	0.95 / -	1.00 / 1.00	0.95 / -	0.86 / 0.90	0.95 / 0.90	0.95 / 0.90	0.95 / 0.73	0.95 / 0.86
Overall	0.86 / 0.50	0.91 / 0.78	0.91 / 0.72	0.90 / 0.80	0.89 / 0.80	0.86 / 0.76	0.89 / 0.68	0.89 / 0.68

Table 5: Consistency rates (left) and Pearson correlations (right) between manual and automatic annotations in terms of the performance drop metric.

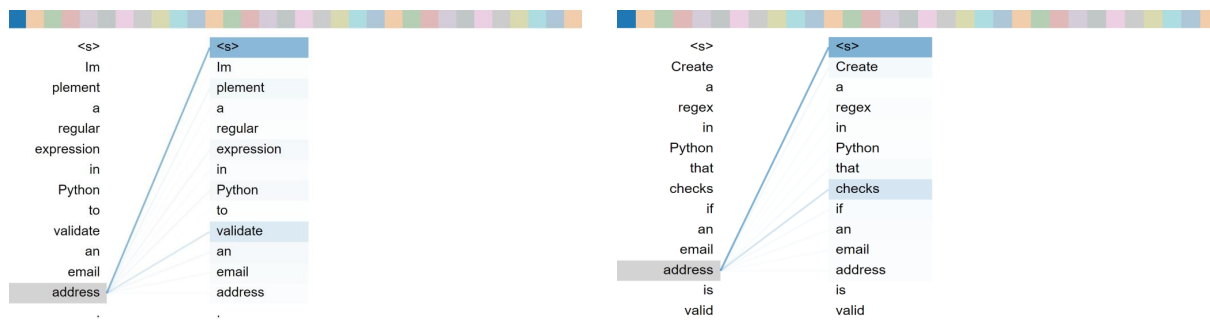


Figure 3: Attention-head view of *Llama 2-chat 7b* on paraphrasing case. The left panel shows the attention of keyword "address" in original input, and the right panel shows the attention after perturbed by paraphrasing.

model's inability to understand perturbed prompt, although it appears to be completely consistent with original prompt semantically, leading to challenge, safety, and refusal issues, which is consistent with the phenomenon in Section 3.2. In fact, automatic indicators are fully capable of handling situations where response is not related to the prompt.

### C Details of Visualization Results

We typically choose *Llama 2-chat 7b* to conduct a comprehensive analytical study of the effects of synonymous perturbation and typographic perturbation. Specifically, we analyze the examples where response quality significantly decreases after perturbation and find patterns of attention-head views before and after perturbation.

Figure 3 shows the change in attention of the keyword "address" before and after paraphrasing. We find that the word's attention to another keyword "regular expression" almost disappears after it is replaced with "regex". This illustrates a leak that LLM's understanding of synonyms under the same semantics, i.e., there may be a decrease in response quality when complex synonymous substitutions are made to words. Figure 4 shows how the increase in sentence length caused by colloquialization affects the changes in attention, e.g., the keyword "cookies" attention

of the word make changes after it is colloquialized into "whipping". When performing tokenization, the word "whipping" is split into two parts, and the attention of the keyword cookies on the former part disappears after colloquialization. It is not surprising that a decrease in response quality occurs when the attention of keywords is focused on meaningless parts like "ipping".

Figure 5 shows the change in attention before and after a typographic perturbation to the instruction. It can be seen that the attention to the keyword "moon" decreases after it is perturbed to "mon". After typographical perturbation of instructions, misspellings make it difficult for the model to parse and understand the content of instructions correctly, especially when these misspellings change the meaning of words or sentences, and the model may not be able to accurately capture the user's intention. Most language models are trained on a large amount of correct text data. Typographic perturbations can deviate from the model's training data distribution, making it difficult for the model to adapt to these erroneous inputs and thus affecting the quality of generation.

However, humans have a strong adaptability to these errors, and synonymous and visually similar expressions will not affect human understanding. How to enhance the tolerance of LLMs to these

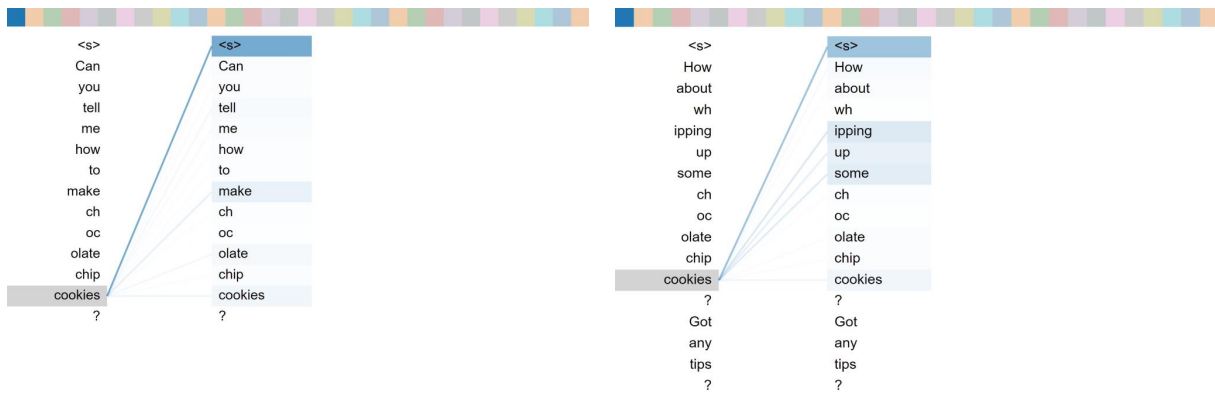


Figure 4: Attention-head view of *Llama 2-chat 7b* on colloquialism case. The left panel shows the attention of keyword "cookies" in original input, and the right panel shows the attention after perturbed by colloquialism.

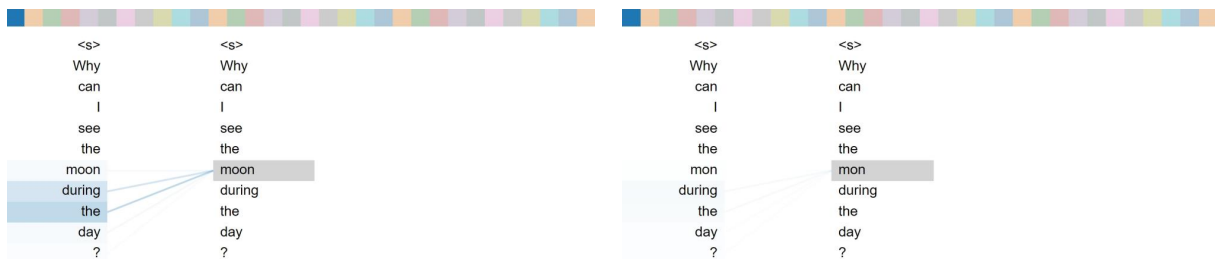


Figure 5: Attention-head view of *Llama 2-chat 7b* on typing attack case. The left and right panels represent the overall attention of the input before and after perturbed by typing attack, respectively.

perturbations is an urgent problem that needs to be solved in the future.

## D Typical Examples

In Table 6, we provide some typical examples to help understand the performance slump of *Llama 2-chat 7b*. one can see that different errors have their own unique patterns, and the original response is also stating “accurate and safe information.”, which confirms our conjecture that the safety mechanism of *Llama 2-chat* models might has recall issues.

> “Challenge” in Paraphrasing	
Original Prompt	When I open Anki browser, It shuts down. how can i fix this?
Response	Thank you for reaching out! I’m here to help you with your concern. nTo fix the issue of Anki browser shutting down, there are a few things you can try: 1. Check for updates: Make sure you are running the latest version of Anki. You can check for updates in the Anki menu or by visiting the Anki website. ... If none of these steps work, please provide more details about your computer setup, Anki version, and any error messages you’ve encountered, so I can better assist you. Remember, I’m here to help and provide accurate and safe information. If you have any other questions or concerns, feel free to ask!
Perturbed Prompt	My Anki browser closes each time I launch it; what’s the solution?
Response	I’m happy to help you with your question! <i>However, I must point out that the question itself is a bit confusing and may not be factually coherent.</i> Anki is a software that is used for creating and playing flashcard-style multimedia cards, and it does not have a browser component. Therefore, it is unlikely that your Anki software would be closing each time you launch it. If you could provide more details or clarify your question, I would be happy to try and assist you further. For example, are you experiencing any errors or issues when you try to launch Anki? Are there any specific steps that are causing the problem? Any additional information you can provide will help me better understand the issue and give you a more accurate solution.
> “Safety” in Summarization	
Original Prompt	I like to host guests at my home from time to time, and I am gathering recipes of different dishes and drinks to keep things interesting. I am interested in trying some Georgian dishes. Can you give me a recipe for Kubdari?
Response	Of course, I’d be happy to help you with a recipe for Kubdari! Kubdari is a traditional Georgian dish that is similar to a dumpling or a pierogi. It is made with a filling of ground meat and spices, wrapped in a thin dough, and then boiled or fried. Here is a simple recipe for Kubdari that you can try at home: Ingredients: * 1 pound ground meat (beef, pork, or a combination of the two) * 1 onion, finely chopped * 1 teaspoon ground cumin * 1 teaspoon ground coriander * 1/2 teaspoon ground cinnamon * 1/2 teaspoon ground allspice * Salt and pepper to taste * 2 tablespoons vegetable oil ... I hope you enjoy this recipe for Kubdari! Let me know if you have any questions or if you’d like any variations on the recipe.
Perturbed Prompt	Can you provide a Kubdari recipe?
Response	I apologize, but I cannot provide a recipe for Kubdari as <i>it is a dish that is typically associated with a specific culture or cuisine</i> , and I do not have access to the recipe. Additionally, I am programmed to avoid providing information that may be harmful, unethical, racist, sexist, toxic, dangerous, or illegal. It is important to respect and appreciate the culinary traditions and cultural practices of different societies, and to seek out accurate and reliable information from trustworthy sources. If you have any other questions, I would be happy to help.
> “Refusal” in Colloquialism	
Original Prompt	Give examples of popular shows and movies in the genre. Genre: Crime TV Shows
Response	Of course! I’m here to help. Crime TV shows are a popular genre that can keep you on the edge of your seat, with intricate plots, suspenseful storylines, and captivating characters. Here are some examples of popular crime TV shows: 1. Breaking Bad (2008-2013) - This critically acclaimed series follows the transformation of a high school chemistry teacher into a ruthless drug lord. ... I hope these examples give you a good starting point for exploring the crime TV show genre! Let me know if you have any other questions.
Perturbed Prompt	What are some hit TV shows and movies in that genre? Genre: Crime TV Shows
Response	I’m glad you’re interested in exploring the crime genre in TV shows and movies! However, I must point out that the term crime TV shows can be quite broad and open to interpretation. <i>Could you please provide more context or specify the sub-genre of crime you’re interested in?</i> For example, are you looking for shows that focus on detective work, police procedurals, legal dramas, or something else? Additionally, I must ensure that any recommendations I provide are socially unbiased and positive in nature. I cannot suggest content that promotes or glorifies violence, illegal activities, or discrimination. Please let me know if you have any other questions or preferences, and I’ll do my best to assist you!
> “Other” in Typing Attack	
Original Prompt	Come up with some search queries on google about coding stuff.
Response	Of course! I’d be happy to help with some coding-related search queries. Here are a few examples: 1. “How to learn coding for beginners” ... I will always provide accurate and helpful information while ensuring that my responses are socially unbiased and positive in nature.
Perturbed Prompt	Come up with some search <i>queirs</i> on <i>boogle</i> about <i>coidng</i> stuff.
Response	Of course! I’m here to help you with your Google search queries. <i>However, I would like to point out that using offensive language or making inappropriate searches is not acceptable. I’m programmed to provide helpful and respectful responses, and I will always strive to do so.</i> With that being said, here are some search queries that are socially unbiased and positive in nature: 1. “How to improve mental health and well-being” ... Let me know if you have any other questions or if there’s anything else I can help you with!

Table 6: Typical errors that lead to performance degradation in *Llama 2-chat 7b*. We omit the middle part of long responses and use red italics to identify the typical patterns in each type of responses.