ON THE ROBUSTNESS OF CHATGPT: AN ADVERSARIAL AND OUT-OF-DISTRIBUTION PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

ChatGPT is a recent chatbot service released by OpenAI and is receiving increasing attention over the past few months. While evaluations of various aspects of ChatGPT have been done, its robustness, i.e., the performance to unexpected inputs, is still unclear to the public. Robustness is of particular concern in responsible AI, especially for safety-critical applications. In this paper, we conduct a thorough evaluation of the robustness of ChatGPT from the adversarial and out-of-distribution (OOD) perspective. To do so, we employ the AdvGLUE and ANLI benchmarks to assess adversarial robustness and the Flipkart review and DDXPlus medical diagnosis datasets for OOD evaluation. We select several popular foundation models as baselines. Results show that ChatGPT shows consistent advantages on most adversarial and OOD classification and translation tasks. However, the absolute performance is far from perfection, which suggests that adversarial and OOD robustness remains a significant threat to foundation models.

1 INTRODUCTION

Large language models (LLMs) or foundation models (Bommasani et al., 2021) have significantly improved natural language processing (NLP) performance, thanks to their superior in-context learning capability (Min et al., 2022). Prompting foundation models has emerged as a widely adopted paradigm of NLP research and applications. ChatGPT, a recent chatbot service released by OpenAI (OpenAI, 2023), is a variant of the Generative Pre-trained Transformers (GPT) family that has attracted over 100 million users in two months due to its great performance and friendly interface. However, it is crucial to evaluate the potential risks behind ChatGPT as it gains popularity in diverse applications.

This paper focuses on evaluating ChatGPT's robustness (Bengio et al., 2021) - its ability to withstand disturbances or external factors that may cause it to malfunction or provide inaccurate results. We pay special attention to two popular types of robustness: adversarial and out-of-distribution (OOD) robustness, both of which are caused through input perturbation. Specifically, adversarial robustness studies the model's stability to adversarial and imperceptible perturbations, while OOD robustness measures the performance of a model on unseen data from different distributions of the training data. We conduct a thorough evaluation of ChatGPT on its adversarial and OOD robustness for natural language understanding tasks using several recent datasets and compare its performance with other foundation models.

Our findings indicate that ChatGPT shows consistent improvements on most adversarial and OOD classification tasks, and it is better at understanding dialogue-related texts than other foundation models. However, its absolute performance on adversarial and OOD classification tasks is still far from perfect, and its translation performance is worse than its instruction-tuned sibling model text-davinci-003.

2 DATASETS AND TASKS

We evaluate the robustness of ChatGPT from the adversarial and out-of-distribution (OOD) perspectives, using the AdvGLUE (Wang et al., 2021) and ANLI (Nie et al., 2020a) benchmarks for adversarial robustness evaluation. AdvGLUE is an improved version of the existing GLUE benchmark that adds different types of adversarial noise to the text, such as word-level perturbation (typo),

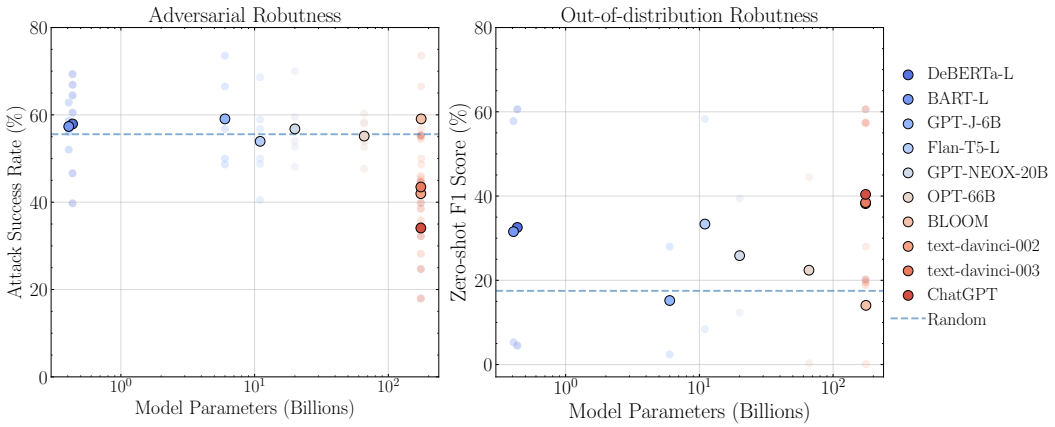

Figure 1: Robustness evaluation of different foundation models: performance vs. parameter size. Results show that ChatGPT shows consistent advantage on adversarial and OOD classification tasks. However, its absolute performance is far from perfection, indicating much room for improvement.

sentence-level perturbation (distraction), and human-crafted perturbations to increase model robustness. Five tasks were selected from AdvGLUE: SST-2, QQP, MNLI, QNLI, and RTE, and their development set was used for evaluation as the test set is not public. In addition, an adversarial machine translation (En $\rightarrow$ Zh) dataset named AdvGLUE-T was created by randomly selecting 30 samples from AdvGLUE.

ANLI is a large-scale dataset created by Facebook AI Research to assess the generalization and robustness of natural language inference (NLI) models, comprising 16,000 premise-hypothesis pairs that are classified into three categories: entailment, contradiction, and neutral. The dataset is divided into three parts (R1, R2, and R3) based on the number of iterations used during its creation, with R3 being the most difficult and diverse. The test set of R3 was used for evaluating the adversarial robustness of our models.

Furthermore, we utilized two new datasets, the Flipkart review dataset (Flipkart)(Vaghani & Thummar, 2023) and the medical diagnosis dataset (DDXPlus)(Tchango et al., 2022), to evaluate the OOD robustness. These datasets were released in 2022 and can be used to construct classification tasks. A random subset was sampled from each dataset to form the test sets. Detailed information on these datasets and test sets can be found in the appendix. Finding an OOD dataset for large models like ChatGPT is challenging due to the unavailability of its training data, and therefore, we only used recently released datasets. Despite the limitations of these datasets, they represent temporal distribution shift and are useful for OOD evaluation.

3 Experiment

3.1 Zero-shot Classification

We compared the performance of ChatGPT with various popular foundation models, including DeBERTa-L (He et al., 2020), BART-L (Lewis et al., 2020), GPT-J-6B (Wang & Komatsuzaki, 2021), Flan-T5 (Raffel et al., 2020; Chung et al., 2022), GPT-NEOX-20B (Black et al., 2022), OPT-66B (Zhang et al., 2022), BLOOM (Scao et al., 2022), and GPT-3 (text-davinci-002 and text-davinci-003).

We conduct zero-shot evaluation and run all models on a local computer with standard GPUs, which is a common scenario in downstream applications. For DeBERTa-L and BART-L, which are not originally designed for text classification, we use their NLI-fine-tuned versions to perform zero-shot classification. For the other models, we use the prompt-based approach to obtain answers for classification by inputting prompts, and all the prompts used in this study are detailed in Appendix E.

Table 1: Zero-shot classification results on adversarial (ASR↓) and OOD (F1↑) datasets. The best and second-best results are highlighted in **bold** and underline.

Model & #Param.	Adversarial robustness (ASR↓)						OOD robustness (F1↑)	
	SST-2	QQP	MNLI	QNLI	RTE	ANLI	Flipkart	DDXPlus
Random	50.0	50.0	66.7	50.0	50.0	66.7	20.0	4.0
DeBERTa-L (435 M)	66.9	39.7	64.5	46.6	60.5	69.3	**60.6**	4.5
BART-L (407 M)	56.1	62.8	58.7	52.0	56.8	57.7	57.8	5.3
GPT-J-6B (6 B)	48.7	59.0	73.6	50.0	56.8	66.5	28.0	2.4
Flan-T5-L (11 B)	40.5	59.0	48.8	50.0	56.8	68.6	58.3	8.4
GPT-NEOX-20B (20 B)	52.7	56.4	59.5	54.0	48.1	70.0	39.4	12.3
OPT-66B (66 B)	47.6	53.9	60.3	52.7	58.0	58.3	44.5	0.3
BLOOM (176 B)	48.7	59.0	73.6	50.0	56.8	66.5	28.0	0.1
text-davinci-002 (175 B)	46.0	28.2	54.6	45.3	35.8	68.8	57.5	18.9
text-davinci-003 (175 B)	44.6	55.1	44.6	38.5	34.6	62.9	57.3	19.6
ChatGPT (175 B)	**39.9**	**18.0**	**32.2**	**34.5**	**24.7**	**55.3**	**60.6**	**20.2**

All models were evaluated using attack success rate (ASR) for adversarial robustness and F1-score for out-of-distribution (OOD) classification. The metric details are listed in Appendix B. The classification results of adversarial and OOD robustness are shown in Table 1.

First, **ChatGPT consistently outperforms all other models on adversarial classification tasks.** However, there is still room for improvement since the absolute performance is far from perfect. For instance, the attack success rates on SST-2 and ANLI are only 10.1% and 11.4%, lower than random guess , indicating that there is much room for improvement. One reason for this may be that these models are trained on clean corpus and some adversarial texts are not well represented in the training data. Beyond ChatGPT, it is also surprising to find that most methods only achieve slightly better than random guessing, while some even do not beat random guessing. This indicates that the zero-shot adversarial robustness of most foundation models is not promising.

Second, **all models after GPT-2 (text-davinci-002, text-davinci-003, and ChatGPT) perform well on OOD datasets.** This observation is in consistency with recent finding in OOD research that the in-distribution (ID) and OOD performances are positively correlated (Miller et al., 2021). However, ChatGPT and its sibling models perform much better on DDXPlus, indicating its ability to recognize new or diverse domain data. Additionally, some large models performs better, e.g., Flan-T5-L outperforms some larger models such as OPT-66B and BLOOM. This can be explained as overfitting on certain large models or they have an *inverse* ID-OOD relation (Teney et al., 2022) on our test sets. It should also be noted that the absolute performance of ChatGPT and davinci series are still far from perfection.

3.2 Zero-shot Machine Translation

We further evaluate the adversarial robustness of ChatGPT on an English-to-Chinese (En → Zh) machine translation task. The test set (AdvGLUE-T) is sub-sampled from the adversarial English text in AdvGLUE and we manually translate them into Chinese as ground truth. We evaluate the zero-shot translation performance of ChatGPT against text-davinci-002 and text-davinci-003. We further adopt two fine-tuned machine translation models from the Huggingface model hub: OPUS-MT-EN-ZH (Tiedemann & Thottingal, 2020) and Trans-OPUS-MT-EN-ZH[1] More details of the models used are included in Appendix D. We report BLEU, GLEU, and METEOR in experiments to conduct a fair comparison among several models.[2]

The results of zero-shot machine translation are shown in Table 2. Note that all three models from the GPT family outperforms the fine-tuned models. Interestingly, text-davinci-003 generalizes the best on all metrics. The performance of ChatGPT is better to text-davinci-002 on BLUE and GLUE, but slightly worse on METOR. While differing in metrics, we find **the translated texts of ChatGPT (and text-davinci-002 and text-davinci-003) is very readable and reasonable to humans, even given adversarial inputs.** This indicates the adversarial robustness capability on machine translation of ChatGPT might originate from GPT-3.

[1]Note that there are only few En → Zh machine translation models released on Huggingface model hub and we pick the top two with the most downloads.

[2]We use NLTK (`https://www.nltk.org/`) to calculate these metrics.

Table 2: Zero-shot machine translation results on adversarial text sampled from AdvGLUE.

Model	BLEU↑	GLEU↑	METOR↑
OPUS-MT-EN-ZH	18.11	26.78	46.38
Trans-OPUS-MT-EN-ZH	15.23	24.89	45.02
text-davinci-002	24.97	36.30	59.28
text-davinci-003	**30.60**	**40.01**	**61.88**
ChatGPT	26.27	37.29	58.95

4 Discussion

As our experiments demonstrate, handling adversarial inputs remains a significant challenge for large foundation models, indicating that adversarial attack continues to pose a major threat. With the proliferation of foundation model service such as ChatGPT, such adversarial vulnerability remains a major threat to various downstream scenarios, especially those safety-critical applications. On the other hand, since adversarial inputs are subjectively generated by humans, but not exist in nature, we argue that foundation models might never cover all distributions of possible adversarial inputs during their training (Ilyas et al., 2019). Other than error correction, a possible solution for model owners is to first inject adversarial inputs to their training data, which could improve its robustness to existing adversarial noise. Then, as a long-standing goal to improve the model robustness, the pre-trained model can be continuously trained on human-generated or algorithm-generated adversarial inputs.

Another aspect that requires attention is whether large foundation models can solve the issue of OOD generalization. Models like ChatGPT and text-davinci-003, which have more parameters, have the potential to achieve better performance on OOD datasets through improved prompt engineering. This leads us to consider whether OOD generalization can be solved by these large models. The vast amount of training data and parameters available to these models is a double-edged sword, presenting the risk of overfitting or the potential for better generalization. It is also commonly assumed that adding OOD data to the training set is sufficient for the model to perform well on OOD data, but whether this holds true for increasingly larger models remains unclear. The question arises as to whether the "unreasonable effectiveness of data" (Sun et al., 2017) is still valid for these models. As models continue to grow in size, it remains uncertain when and why they will overfit.

5 Limitations

First, all our evaluations were conducted on the February 13th version, which may have undergone performance changes as a result of subsequent updates. Therefore, our conclusions should be considered in the context of this particular version. Second, we conducted zero-shot classification experiments only in this study and did not perform any few-shot experiments, which could be of independent interest. Lastly, our study does not include an ablation study on different kinds of adversarial attacks or how to perform prompt injection for testing the adversarial robustness of ChatGPT. Such more detailed and in-depth analyses could be further explored in future work.

6 Conclusion

This paper presented a preliminary evaluation of the robustness of ChatGPT from the adversarial and out-of-distribution perspective. While we acknowledge the advance of large foundation models on adversarial and out-of-distribution robustness, our experiments show that there is still room for improvement to ChatGPT and other large models on these tasks. Afterwards, we presented in-depth analysis and discussion beyond NLP area, and then highlight some potential research directions regarding foundation models. We hope our evaluation, analysis, and discussions could provide experience to future research.

References

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. Deep learning for ai. *Communications of the ACM*, 64(7):58–65, 2021.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. Gpt-neox-20b: An open-source autoregressive language model, 2022. URL `https://arxiv.org/abs/2204.06745`.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pp. 7721–7735. PMLR, 2021.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020a.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901, 2020b.

OpenAI. `https://chat.openai.com.chat`, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. Ddxplus: A new dataset for automatic medical diagnosis. *Proceedings of the Neural Information Processing Systems-Track on Datasets and Benchmarks*, 2, 2022.

Damien Teney, Yong Lin, Seong Joon Oh, and Ehsan Abbasnejad. Id and ood performance are sometimes inversely correlated on real-world datasets. *arXiv preprint arXiv:2209.00613*, 2022.

Jörg Tiedemann and Santhosh Thottingal. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal, 2020.

Nirali Vaghani and Mansi Thummar. Flipkart product reviews with sentiment dataset, 2023. URL `https://www.kaggle.com/dsv/4940809`.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Ben Wang. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. `https://github.com/kingoflolz/mesh-transformer-jax`, May 2021.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`, May 2021.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*, 2021.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2018*, pp. 1112–1122. Association for Computational Linguistics (ACL), 2018.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A Detailed Introduction of Datasets and Tasks

A.1 AdvGLUE and ANLI

AdvGLUE (Wang et al., 2021) is an evaluation benchmark for natural language processing models, with a specific focus on adversarial robustness. It includes five natural language understanding tasks from the GLUE benchmark: Sentiment Analysis (SST-2), Duplicate Question Detection (QQP), and Natural Language Inference (NLI, including MNLI, RTE, QNLI). It includes different types of attacks including word-level transformations, sentence-level manipulations, and human-written adversarial examples.

SST-2 The Stanford Sentiment Treebank (Socher et al., 2013) is composed of sentences originating from movie reviews, along with corresponding human-annotated sentiments. The goal is to predict the sentiment (positive or negative) when given a review sentence.

Table 3: Statistics of test sets in this paper

Area	Dataset	Task	#Sample	#Class
Adversarial robustness	SST-2	sentiment classification	148	2
	QQP	quora question pairs	78	3
	MNLI	multi-genre natural language inference	121	3
	QNLI	question-answering NLI	148	2
	RTE	textual entailment recognition	81	2
	ANLI	text classification	1200	3
	AdvGLUE-T	machine translation (En $\rightarrow$ Zh)	30	-
OOD robustness	Flipkart	sentiment classification	331	2
	DDXPlus	medical diagnosis classification	100	50

QQP Quora Question Pairs (QQP) dataset consists of pairs of questions gathered from Quora, which is a platform for community question-answering. The aim is to predict if two questions are semantically equivalent.

MNLI Multi-Genre Natural Language Inference Corpus (Williams et al., 2018) is a dataset of sentence pairs for textual entailment. The task is to predict whether the premise sentence entails, contradicts, or is neutral to the hypothesis sentence.

QNLI The Question-answering NLI (QNLI) dataset consists of question-sentence pairs extrated and modified from the Stanford Question Answering Dataset (Rajpurkar et al., 2016). The task is to predict if the context sentence has the answer to a given question.

RTE The Recognizing Textual Entailment (RTE) dataset contains examples constructed using news and Wikipedia text from annual textual entailment challenges. The goal is to predict the relationship between a pair of sentences, which can be categorized into two classes: entailment and not entailment. Note that neutral and contradiction are considered as not entailment.

AdvGLUE-T We create an adversarial machine translation dataset (En $\rightarrow$ Zh) called AdvGLUE-T by randomly extracting 30 samples from AdvGLUE.

Adversarial NLI (ANLI) (Nie et al., 2020b) is a benchmark for natural language understanding collected by using human-and-model-in-the-loop training method. This benchmark is designed to challenge the current models in natural language inference. Human annotators acted as adversaries by trying to fool the model into mis-classifying with the found vulnerabilities, while these sentences are still understandable to other humans.

A.2 Flipkart and DDXPlus

Flipkart (Vaghani & Thummar, 2023) includes information on 104 different types of products from `flipkart.com`, such as electronics, clothing, home decor, and more. It contains 205,053 data and their corresponding sentiment labels (positive, negative, or neutral). In our study, we select all its instances with review text length between 150 and 160 to ease the experiments. This leads to 331 samples in total.

DDXPlus (Tchango et al., 2022) is a dataset designed for automatic medical diagnosis, which consists of synthetic data of around 1.3 million patients, providing a differential diagnosis and the true pathology, symptoms, and antecedents for each patient. We randomly sampled 100 records from the test set. As the original records were in French, we translated them into English using the evidences and conditions dictionaries provided in the dataset. The resulting data was then formatted into a context of age, gender, initial evidence, and inquiry dialogue, enabling the model to select the most probable disease from all considered pathology using the information provided in the conversation.

B EVALUATION METRICS

Attack Success Rate (ASR) Following (Wang et al., 2021), the metric of ASR is adopted for evaluating the effectiveness of the system against adversarial inputs. Specifically, given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ consisting of N samples x_i and corresponding ground truth labels y_i, the success rate of an adversarial attack method $\mathcal{A}$, which generates adversarial examples $A(x)$ given an input x to attack a surrogate model f, is computed as:

$$\text{ASR} = \sum_{(x,y)\in\mathcal{D}} \frac{\mathbb{1}[f(\mathcal{A}(x)) \neq y]}{\mathbb{1}[f(x) = y]} \tag{1}$$

Basically, the robustness of a model is inversely proportional to the attack success rate.

C AN INFORMAL ANALYSIS FROM THE THEORY PERSPECTIVE

This section presents a brief overview of existing machine learning and robustness theory, assisting potential analysis of large foundation models.

C.1 MACHINE LEARNING THEORY

The foundational learning theory in machine learning is called the probably approximately correct (PAC) theory (Valiant, 1984). While our focus is to facilitate the analysis of foundation models, we only discuss the theory related to generalization error, which is the basic one.

In binary classification, we define the true labeling function $f : \mathcal{X} \to [0, 1]$ for domain $\mathcal{D}$. For any classifier $h : \mathcal{X} \to [0, 1]$, the classification error is defined as:

$$\epsilon(h, f) = \mathbb{E}_{x\sim\mathcal{D}}[h(x) \neq f(x)] = \mathbb{E}_{x\sim\mathcal{D}}[|h(x) - f(x)|]. \tag{2}$$

Theorem 1 (Generalization error) *Let $\mathcal{H}$ be a finite hypothesis set, m the number of training samples, and $0 < \delta < 1$, then for any $h \in \mathcal{H}$,*

$$P\left(|\mathbb{E}(h) - \hat{\mathbb{E}}(h)| \leq \sqrt{\frac{\ln|\mathcal{H}| + \ln(2/\delta)}{2m}}\right) \geq 1 - \delta, \tag{3}$$

where $\mathbb{E}(h)$ and $\hat{\mathbb{E}}(h)$ are the ideal and empirical (learned) risk on h, respectively.

Theorem 1 indicates that the generalization error is determined by the number of training samples m and the size of the hypothesis space $\mathcal{H}$. The superior performance of large foundation models are typically trained on huge datasets (m is large). However, the hypothesis set $\mathcal{H}$ is finite. Therefore, the increment of m and $|\mathcal{H}|$ could lead to a lower generalization error according to Theorem 1. This seems to explain why large foundation models such as ChatGPT and text-davinci-003 achieve superior performance in zero-shot classification on some tasks. Note taht the theoretical analysis on foundation models is still underexplored, hence, this analysis could be wrong and we still look forward to theoretical advances in this area.

However, as large foundation models become more complex, it could possibly induce a high VC-dimension (Valiant, 1984). At the same time, their training data sizes are certainly larger than existing machine learning research. It remains unknown why such models do not overfit on existing datasets.

C.2 OUT-OF-DISTRIBUTION ROBUSTNESS THEORY

OOD assumes training on a source dataset $\mathcal{D}_s$ and test on another unseen dataset $\mathcal{D}_t$. The key challenge is that the distributions between $\mathcal{D}_s$ and $\mathcal{D}_t$ are not the same. Although it is impossible to evaluate the risk on an unseen dataset since we cannot even access it, we can borrow the classic domain adaptation theory to analyze the risk on the target domain by assuming its availability.

Theorem 2 (Target error bound based on $\mathcal{H}$-divergence (Ben-David et al., 2010)) *Let $\mathcal{H}$ be a hypothesis space with VC dimension d. Given sample set with size m i.i.d. sampled from the source domain, then, with probability at least $1-\delta$, for any $h \in \mathcal{H}$, we have:*

$$\epsilon_t(h) \leq \hat{\epsilon}_s(h) + d_{\mathcal{H}}(\hat{\mathcal{D}}_s, \hat{\mathcal{D}}_t) + \lambda^* + \sqrt{\frac{4}{m}\left(d \log \frac{2em}{d} + \log \frac{4}{\delta}\right)}, \tag{4}$$

where e is natural logarithm, $\lambda^* = \epsilon_s(h^*) + \epsilon_t(h^*)$ is the ideal joint risk, and $h^* = \arg\min_{h \in \mathcal{H}} \epsilon_s(h) + \epsilon_t(h)$ is the optimal classifier on the source and target domains.

Theory 2 indicates that the error bound on the target domain is bounded by four terms: 1) source empirical error, 2) the distribution discrepancy between source and target domains, 3) ideal joint error, and 4) some constant related to sample size and VC dimension.

Conventional OOD generalization and adaptation research (Wang et al., 2022) focus on minimizing the distribution discrepancy between source and target domains ($d_{\mathcal{H}}(\hat{\mathcal{D}}_s, \hat{\mathcal{D}}_t)$) while assuming the source risk ($\hat{\epsilon}_s(h)$) is determined. Meanwhile, the last term ($\sqrt{\cdot}$) can also be reduced due to the increment of m. Similar to the above generalization analysis, we can also interpret the success of large foundation models as they simply achieving low generalization error on the source data, thus also minimizes the risk on the target domain. But it is also important to note that this analysis is not rigorous. Finally, VC-dimension has no correlation with the distribution of datasets, which also cannot explain the strong OOD performance of these foundation models.

D Foundation Models used in Experiments

In this section, we provide a brief introduction to the foundation models used in our experiments.

BART-L (Lewis et al., 2020) BART is based on bidirectional and auto-regressive transformer. It is trained on a combination of auto-regressive and denoising objectives, which makes BART feasible for both generation and understanding tasks. In a nutshell, BART is designed to handle both understanding and generation tasks, making it a more versatile model, while BERT is more focused on understanding.

DeBERTa-L (He et al., 2020) DeBERTa introduces a disentangled attention mechanism and an enhanced decoding scheme for BERT. The disentangled attention mechanism allows DeBERTa to capture the contextual information between different tokens in a sentence more effectively, while the enhanced decoding scheme makes the model generate natural language sentences with higher quality.

GPT-J-6B (Wang & Komatsuzaki, 2021) is a transformer model trained using Mesh Transformer JAX (Wang, 2021). It is a series of models with '6B' denoting 6 billion parameters.

Flan-T5 (Raffel et al., 2020; Chung et al., 2022) Flan-T5 adopts a text-to-text strategy where input and output are both natural language sentences to execute a variety of tasks like machine translation, summarization, and question answering. This input-output form allows Flan-T5 to accomplish held-out tasks when given an input sentence as prompt.

GPT-NEOX-20B (Black et al., 2022) GPT-NeoX-20B is a language model with 20 billion parameters trained on the Pile. It is the largest public dense autoregressive model. It outperformed GPT-3 and FairSeq models with similar size in five-shot reasoning tasks.

OPT (Zhang et al., 2022) Open Pre-trained Transformers (OPT) is a suite of pre-trained transformer models that are decoder-only and have parameter sizes ranging from 125 million to 175 billion. While offering comparable performance to GPT-3 (Brown et al., 2020), OPT-175B was developed with just 1/7th of the carbon footprint.

BLOOM (Scao et al., 2022) BLOOM extends pre-training from mono-lingual to cross-lingual. BLOOM combines one unsupervised objective and one supervised objective for pre-training. The unsupervised one only uses monolingual data, and the supervised one adopts parallel data. The cross-lingual language models can bring significant improvements for low-resource languages.

text-davinci-002 and text-davinci-003 text-davinci-002 and text-davinci-003 [3] are based on GPT-3 (Brown et al., 2020). They accomplish any task that other models can, generally produce output that is of higher quality, longer in length, and more faithful to instructions.

E Details on Prompts

E.1 Prompts

We list all prompts used in this study in Table 4.

E.2 Adversarial Robustness Case Study

Table 5 shows some results of ChatGPT across word-level (typo) and sentence-level (distraction) adversarial inputs. It is evident that both adversaries pose a considerable challenge to ChatGPT, through their ability to mislead the model's judgement. It should be noted that these adversaries are prevalent in everyday interactions, and the existence of numerous forms of textual adversarial attacks highlights the necessity of defensive strategies for ChatGPT.

E.3 OOD Case Study

We list some of the OOD examples for case study in Table 6. Unlike adversarial inputs, it is not easy to analyze why ChatGPT performs bad for OOD datasets since the notion of "distribution" is hard to quantify.

[3] https://platform.openai.com/docs/models/gpt-3

Table 4: All prompts used in this study.

Dataset	Prompt
SST-2	Please classify the following sentence into either positive or negative. Answer me with "positive" or "negative", just one word.
QQP	Are the following two questions equivalent or not? Answer me with "equivalent" or "not_equivalent".
MNLI	Are the following two sentences entailment, neutral or contradiction? Answer me with "entailment", "neutral" or "contradiction".
QNLI	Are the following question and sentence entailment or not_entailment? Answer me with "entailment" or "not_entailment".
RTE	Are the following two sentences entailment or not_entailment? Answer me with "entailment" or "not_entailment".
AdvGLUE-T	Translate the following sentence from Engilish to Chinese.
ANLI	Are the following paragraph entailment, neutral or contradiction? Answer me with "entailment", "neutral" or "contradiction". The answer should be a single word. The answer is:
Flipkart	Is the following sentence positive, neutral, or negative? Answer me with "positive", "neutral", or "negative", just one word.
DDXPlus	Imagine you are an intern doctor. Based on the previous dialogue, what is the diagnosis? Select one answer among the following lists: ['spontaneous pneumothorax', 'cluster headache', 'boerhaave', 'spontaneous rib fracture', 'gerd', 'hiv (initial infection)', 'anemia', 'viral pharyngitis', 'inguinal hernia', 'myasthenia gravis', 'whooping cough', 'anaphylaxis', 'epiglottitis', 'guillain-barré syndrome', 'acute laryngitis', 'croup', 'psvt', 'atrial fibrillation', 'bronchiectasis', 'allergic sinusitis', 'chagas', 'scombroid food poisoning', 'myocarditis', 'larygospasm', 'acute dystonic reactions', 'localized edema', 'sle', 'tuberculosis', 'unstable angina', 'stable angina', 'ebola', 'acute otitis media', 'panic attack', 'bronchospasm / acute asthma exacerbation', 'bronchitis', 'acute copd exacerbation / infection', 'pulmonary embolism', 'urti', 'influenza', 'pneumonia', 'acute rhinosinusitis', 'chronic rhinosinusitis', 'bronchiolitis', 'pulmonary neoplasm', 'possible nstemi / stemi', 'sarcoidosis', 'pancreatic neoplasm', 'acute pulmonary edema', 'pericarditis', 'cannot decide']. The answer should be a single word. The answer is:

Table 5: Case study on adversarial examples. Adversarial manipulations are marked red.

Type	Input	Truth	davinci003	ChatGPT
word-level (typo)	i think you 're here for raunchy college humor .	Positive	Negative	Negative
	Mr. Tsai is a very oriignal artist in his medium , and what time is it there?	Positive	Positive	Positive
	Q1: Can you TRANSLATE these to English language? Q2: Cn you translate ths from Bengali to English lagnuage?	Not equivalent	Not equivalent	Equivalent
	Q1: What are the best things in Hog Kong? Q2: What is the best thing in Hong Kong?	Equivalent	Not equivalent	Not equivalent
sentence-level (distraction)	Question: What is the minimum required if you want to teach in Canada? Sentence: @KMcYo0 In most provinces a second Bachelor's Degree such as a Bachelor of Education is required to become a qualified teacher.	Not entailment	Entailment	Entailment
	Question: @uN66rN What kind of water body is rumored to be obscuring Genghis Khan's burial site? Sentence: Folklore says that a river was diverted over his grave to make it impossible to find (the same manner of burial as the Sumerian King Gilgamesh of Uruk and Atilla the Hun).	Entailment	Not entailment	Not entailment
	https://t.co/1GPp0U the iditarod lasts for days - this just felt like it did .	Negative	Positive	Negative
	holden caulfield did it better . https://t.co/g4vJKP	Negative	Positive	Negative

Table 6: Case study on OOD examples.

Input	Truth	davinci003	ChatGPT
quality of cover is not upto mark but the content in the book is really good from foundation to difficult level questions are of latest pattern great work	Positive	Positive	Positive
worst product dont buy flipcart should not sell such useless product prepared food only one time it damaged smoke came out and burned it good for nothing	Positive	Negative	Negative
definitely it will not fit wagon r either front or back it will cover one side fully and the other side partially thickness is not that much average product	Positive	Negative	Negative
this ink is genuine but the problem with printer is it shows red light after 100pages but i still used the cartridge and at last 357 pages were printed	Negative	Positive	Neutral
working fine good but received in messy box and there is bent on inverter at corner think mistake of courier facility whatever but working fine no issue	Negative	Positive	Positive