On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective

Jindong Wang[1,*], **Xixu Hu**[1,2‡†], **Wenxin Hou**[3†], **Hao Chen**[4], **Runkai Zheng**[1,5‡], **Yidong Wang**[6], **Linyi Yang**[7], **Wei Ye**[6], **Haojun Huang**[3], **Xiubo Geng**[3], **Binxing Jiao**[3], **Yue Zhang**[7], **Xing Xie**[1]

[1]Microsoft Research, [2]City University of Hong Kong, [3]Microsoft STCA, [4]Carnegie Mellon University, [5]Chinese University of Hong Kong (Shenzhen), [6]Peking University, [7]Westlake University

https://github.com/microsoft/robustlearn

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. Moreover, ChatGPT shows astounding performance in understanding dialogue-related texts and we find that it tends to provide informal suggestions for medical tasks instead of definitive answers. Finally, we present in-depth discussions of possible research directions.

1 Introduction

Large language models (LLMs), or foundation models (Bommasani et al., 2021), have achieved significant performance on various natural language process (NLP) tasks. Given 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 is a recent chatbot service released by OpenAI (OpenAI, 2023), which is a variant of the Generative Pre-trained Transformers (GPT) family. Thanks to its friendly interface and great performance, ChatGPT has attracted over 100 million users in two months.

It is of imminent importance to evaluate the potential risks behind ChatGPT given its increasing worldwide popularity in diverse applications. While previous efforts have evaluated various aspects of ChatGPT in law (Choi et al., 2023), ethics (Shen et al., 2023), education (Khalil and Er, 2023), and reasoning (Bang et al., 2023), we focus on its *robustness* (Bengio et al., 2021), which, to our best knowledge, has not been thoroughly evaluated yet. Robustness refers to the ability to withstand disturbances or external factors that may cause it to malfunction or provide inaccurate results. It is

*Contact: jindong.wang@microsoft.com.
†Equal contribution.
‡Work done during internship at Microsoft Research Asia.

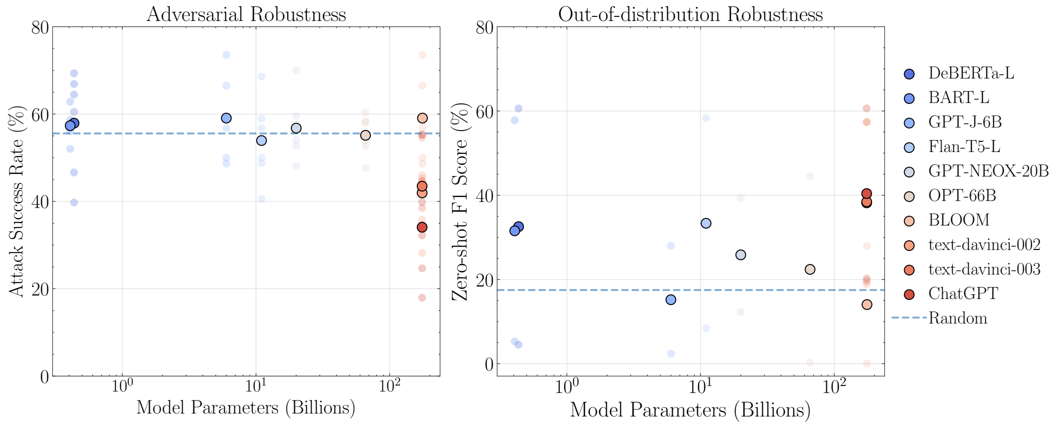

Figure 1: Robustness evaluation of different foundation models: performance vs. parameter size. Results show that ChatGPT shows consistent advantage on adversarial and OOD classification tasks. However, its absolute performance is far from perfection, indicating much room for improvement.

important to practical applications especially the safety-critical scenarios. For instance, if we apply ChatGPT or other foundation models to fake news detection, a malicious user might add noise or certain perturbations to the content to bypass the detection system. Without robustness, the reliability of the system collapses.

Robustness threats exist in a wide range of scenarios: out-of-distribution (OOD) samples (Wang et al., 2022), adversarial inputs (Goodfellow et al., 2014), long-tailed samples (Zhang et al., 2021), noisy inputs (Natarajan et al., 2013), and many others. In this paper, we pay special attention to two popular types of robustness: the adversarial and OOD robustness, both of which are caused through input perturbation. Specifically, adversarial robustness studies the model's stability to adversarial and imperceptible perturbations, e.g., adding trained noise to an image or changing some keywords of a text. On the other hand, OOD robustness measures the performance of a model on unseen data from different distributions of the training data, e.g., classifying sketches using a model trained for art painting or analyzing a hotel review using a model trained for appliance review. More background of these robustness is elaborated in Section 2.2.

Zero-shot robustness evaluation. While robustness research often requires training and optimization (e.g., fine-tuning, linear probing, domain adaptation and generalization, Section 2.2), in this work, we focus on *zero-shot* robustness evaluation. Given a foundation model, we perform inference directly on the test dataset for evaluation. We argue that it becomes more expensive and unaffordable to train, or even load existing (and future, larger) foundation models. For instance, the largest Flan-T5 model has 11 billion parameters (Chung et al., 2022), which is already beyond the capability of most researchers and practitioners. Thus, their zero-shot performance becomes important to downstream tasks. On the other hand, foundation models are typically trained on huge volumes of datasets with huge amount of parameters, which seems to challenge conventional machine learning theories (Appendix C):

Are large foundation models all we need for robustness?

In this work, we conduct a thorough evaluation of ChatGPT on its adversarial and OOD robustness for natural language understanding tasks. It is challenging to select appropriate datasets for evaluating ChatGPT since it is known to be trained on huge text datasets as of 2021. Eventually, we leverage several recent datasets for our evaluation: AdvGLUE (Wang et al., 2021) and ANLI (Nie et al., 2020a) for adversarial robustness and two new datasets for OOD robustness: Flipkart review (Vaghani and Thummar, 2023) and DDXPlus medical diagnosis datasets (Tchango et al., 2022). Furthermore, we randomly selected 30 samples from AdvGLUE to form an adversarial translation dataset to evaluate the translation performance. These datasets represent various levels of robustness, thus provide a fair evaluation. The detailed information of these datasets are introduced in Section 3. We then select several popular foundation models from Huggingface model hub and OpenAI service[4] to compare with ChatGPT. In summary, we have 9 tasks and overall $2,089$ test examples.[5]

Our findings. We perform zero-shot inference on all tasks using these models and Fig. 1 summarizes our main results. The major findings of the study include:

1. What ChatGPT does well:
    - ChatGPT shows consistent improvements on most adversarial and OOD classification tasks.
    - ChatGPT is good at translation tasks. Even in the presence of adversarial inputs, it can consistently generate readable and reasonable responses.
    - ChatGPT is better at understanding dialogue-related texts than other foundation models. This could be attributed to its enhanced ability as a chatbot service, leading to good performance on DDXPlus dataset.
2. What ChatGPT does not do well:
    - The absolute performance of ChatGPT on adversarial and OOD classification tasks is still far from perfection even if it outperforms most of the counterparts.
    - The translation performance of ChatGPT is worse than its instruction-tuned sibling model text-davinci-003.
    - ChatGPT does not provide definitive answers for medical-related questions, but instead offers informed suggestions and analysis. Thus, it can serve as a friendly assistant.
3. Other general findings about foundation models:
    - Task-specific fine-tuning helps language models perform better on related tasks, e.g., NLI-fine-tuned RoBERTa-L has similar performance to Flan-T5-L.
    - Instruction tuning benefits large language models, e.g., Flan-T5-L achieves comparable performance to text-davinci-002 and text-davinci-002 with significantly less parameters.

Beyond evaluations, we share more reflections in the discussion and limitation sections, providing experience and suggestions to future research. Finally, we open-source our code and results at `https://github.com/microsoft/robustlearn` to facilitate future explorations.

2 Background

2.1 Foundation Models, ChatGPT, and Existing Evaluation

Foundation models have become a popular research and application paradigm for natural language process tasks. Since foundation models are trained on large volumes of data, they show significant performance improvement on different downstream tasks such as sentiment analysis, question answering, automatic diagnosis, logical reasoning, and sequence tagging. ChatGPT is a generative foundation model that belongs to the GPT-3.5 series in OpenAI's GPT family, coming after GPT (Radford et al., 2018), GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), and InstructGPT (Ouyang et al., 2022). In contrast to its predecessors, ChatGPT makes it easy for every one to use just through a browser with enhanced multi-turn dialogue capabilities. Although the technical details of ChatGPT is still not released, it is known to be trained using reinforcement learning from human feedback (RLHF) (Christiano et al., 2017) with instruction tuning. Other than natural language processing, there are also emerging efforts in building foundation models for computer vision (Dehghani et al., 2023), music generation (Agostinelli et al., 2023), biology (Luo et al., 2022; Lee et al., 2020), and speech recognition (Radford et al., 2022).

Previous efforts evaluate ChatGPT in different aspects (van Dis et al., 2023). Bang et al. (2023) proposes a multi-task, multi-modal, and multilingual evaluation of ChatGPT on different tasks. They showed that ChatGPT performs reasonably well on most tasks, while it does not bring great performance on low-resource tasks. Similar empirical evaluations are also made by Gozalo-Brizuela and Garrido-Merchan (2023); Azaria (2022). Specifically, Qin et al. (2023) also did several evaluations

[4]Huggingface: `https://huggingface.co/models`. OpenAI service: `https://openai.com/api`.

[5]Although the sample size may seem small, it must be noted that due to the current unavailability of an API service for ChatGPT, conducting experiments on a larger scale is challenging.

and they found that ChatGPT does not do well on fine-grained downstream tasks such as sequence tagging. In addition to research from artificial intelligence, researchers from other areas also showed interest in ChatGPT. Hacker et al. (2023); Shen et al. (2023) expressed concerns that ChatGPT and other large models should be regulated since they are double-edged swords. The evaluations on ethics are done in (Zhuo et al., 2023). There are reflections and discussions from law (Choi et al., 2023), education (Khalil and Er, 2023; M Alshater, 2022; Susnjak, 2022; Guo et al., 2023), human-computer interaction (Tabone and de Winter, 2023), medicine (Jeblick et al., 2022), and writing (Biswas, 2023). To the best of our knowledge, a thorough robustness evaluation is currently under-explored.

2.2 Robustness

In the following, we present the formulation of robustness with the classification task (other tasks can be formulated similarly). We are given a K-class classification dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, where $\mathbf{x} \in \mathbb{R}^d$ and $y \in [K]$ are its d-dimensional input and output, respectively. We use $\ell[\cdot, \cdot]$ to denote the loss function.

Adversarial robustness An adversarial input (Goodfellow et al., 2014) $\mathbf{x}'$ is generated by adding a ϵ-bounded, imperceptible perturbation δ to the original input $\mathbf{x}$. The optimal classifier can be learned by optimizing the following objective (Madry et al., 2017):

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(\mathbf{x}, y) \in \mathcal{D}} \max_{|\delta| \leq \epsilon} \ell[f(\mathbf{x} + \delta), y].$$

Out-of-distribution robustness On the other hand, OOD robustness (generalization) (Wang et al., 2022; Shen et al., 2021) aims to learn an optimal classifier on an unseen distribution by training on existing data. One popular formulation for OOD robustness is to minimize the average risk on all distributions e, which is sampled over the set of all possible distributions (could be large than $\mathcal{D}$):

$$\min_{f \in \mathcal{H}} \mathbb{E}_{e \sim \mathcal{Q}} \mathbb{E}_{(\mathbf{x}, y) \in \mathcal{D}^e} \ell[f(\mathbf{x}), y].$$

Yang et al. (2022) presented GLUE-X, a benchmark based on GLUE and then conducted a thorough evaluation of the OOD robustness of language models by training on in-distribution (ID) sets and then testing on OOD sets. Ours, however, performs zero-shot evaluation. The OOD robustness of ChatGPT cannot be evaluated by GLUE and GLUE-X benchmarks since it may include the entire GLUE datasets in its training data.

3 Datasets and Tasks

3.1 Adversarial Datasets

We adopt AdvGLUE (Wang et al., 2021) and adversarial natural language inference (ANLI) (Nie et al., 2020a) benchmarks for evaluating adversarial robustness. AdvGLUE is a modified version of the existing GLUE benchmark (Wang et al., 2019) by adding different kinds of adversarial noise to the text: word-level perturbation (typo), sentence-level perturbation (distraction), and human-crafted perturbations. We adopt 5 tasks from AdvGLUE: SST-2, QQP, MNLI, QNLI, and RTE. Since the test set of AdvGLUE is not public, we adopt its development set instead for evaluation. Although AdvGLUE is a classification benchmark, we additionally construct an adversarial machine translation (En $\rightarrow$ Zh) dataset, termed AdvGLUE-T, by randomly selecting 30 samples from AdvGLUE.

ANLI is a large-scale dataset designed to assess the generalization and robustness of natural language inference (NLI) models, which was created by Facebook AI Research. It comprises 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. Therefore, we select the test set of R3 for evaluating the adversarial robustness of our models. Detailed information of AdvGLUE and ANLI can be found in Appendix A.1.

Table 1: 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

3.2 Out-of-distribution Datasets

We adopt two new datasets[6] for OOD robustness evaluation: Flipkart (Vaghani and Thummar, 2023) and DDXPlus (Tchango et al., 2022). Flipkart is a product review dataset and DDXPlus is a new medical diagnosis dataset, both of which are released in 2022. These two datasets can be used to construct classification tasks. We randomly sample a subset of each dataset to form the test sets. Detailed introduction and construction of each test set can be found in Appendix A.2. Table 1 shows the statistics of each dataset.

Remark: Finding an OOD dataset for large models like ChatGPT is difficult due to the unavailability of its training data. Consider these datasets as 'out-of-example' datasets since they did not show up in ChatGPT's training data. Additionally, distribution shift may happen at different dimensions: not only across domains, but also across time. Thus, even if ChatGPT and other LLMs may already use similar datasets like medical diagnosis and product review, our selected datasets are still useful for OOD evaluation due to temporal distribution shift. Finally, we must admit the limitation of these datasets and look forward to brand new ones for more thorough evaluation.

4 Experiment

4.1 Zero-shot Classification

4.1.1 Setup

We compare the performance of ChatGPT on AdvGLUE classification benchmark with the following existing popular foundation models: DeBERTa-L (He et al., 2020), BART-L (Lewis et al., 2020), GPT-J-6B (Wang and Komatsuzaki, 2021), Flan-T5 (Raffel et al., 2020; Chung et al., 2022), GPT-NEOX-20B (Black et al., 2022), OPT-66B (Zhang et al., 2022a), BLOOM (Scao et al., 2022), and GPT-3 (text-davinci-002 and text-davinci-003) [7]. The latter two are from OpenAI API service and the rest are on Huggingface model hub. The notation '-L' means '-large', as we only evaluate the large version of these models. The detailed information of these models are introduced in Appendix D.

For adversarial classification tasks on AdvGLUE and ANLI, we adopt attack success rate (ASR) as the metric for robustness. The metric details are listed in Appendix B. For OOD classification tasks, F1-score (F1) is adopted as the metric. As mentioned before, we only perform zero-shot evaluation. Thus, we simply run all models on a local computer with plain GPUs, which could be the case in most downstream applications.[8] Note that we use the NLI-fine-tuned version of DeBERTa-L and BART-L on natural language inference tasks to perform zero-shot classification since they are not originally designed for text classification. For other models, we adopt the prompt-based paradigm to

[6]Considering ChatGPT is reported to be trained on a substantial corpus of internet language data as of 2021, identifying an out-of-distribution dataset poses a difficulty. Furthermore, we concern that previous natural language processing datasets predating 2022 may have been assimilated by ChatGPT, so we only utilize datasets that are recently released.

[7]Note that the classification task may be unfavorable to the generative models since we did not limit their output space as discriminative models like DeBERTa-L do.

[8]Even the local computer is not that "plain" since it requires at least 1 A100 GPU with 80 GB of memory.

Table 2: 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	<u>57.7</u>	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)	<u>40.5</u>	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	<u>58.3</u>	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	<u>28.2</u>	54.6	45.3	35.8	68.8	57.5	18.9
text-davinci-003 (175 B)	44.6	55.1	<u>44.6</u>	<u>38.5</u>	<u>34.6</u>	62.9	57.3	<u>19.6</u>
ChatGPT (175 B)	**39.9**	**18.0**	**32.2**	**34.5**	**24.7**	**55.3**	**60.6**	**20.2**

get answers for classification by inputting prompts. All prompts used in this paper are presented in Appendix E. Note that we manually processed some outputs since the outputs of some generative LLMs are not easy to control.

4.1.2 Results

The classification results of adversarial and OOD robustness are shown in Table 2.

First, **ChatGPT shows consistent improvements on adversarial datasets.** It outperforms all counterparts on all adversarial classification tasks. However, we see that there is still room for improvement since the absolute performance is far from perfection. For instance, the ASRs on SST-2 and ANLI are 40% and 55.3%, respectively, indicating much room for improvement. This could be due to the reason that they are trained on clean corpus and some adversarial texts are washed out from 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. Such adversarial vulnerability poses a major threat to various applications of foundation models, which we will further discuss in Section 5.1 and Appendix C.1. In addition to foundation models, we are surprised to find that some small models also achieve great performance on adversarial tasks while it has much less parameters than the strong models (e.g, DeBERTa-L on QQP and QNLI tasks). This indicate that fine-tuning on relevant tasks can still improve the performance. Furthermore, Flan-T5 also achieves comparable performance to most larger models. Since Flan-T5 is also trained via instruction tuning, this implies the efficacy of such training approach in prompting-based NLP tasks.

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. More discussions on OOD are presented in Section 5.2 and Appendix A.2 shows some informal analysis from the perspective of OOD theory.

Third, on the DDXPlus dataset, **ChatGPT is better at understanding diaglogue-related texts compared with other LLMs.** The DDXPlus benchmark presents a formidable challenge for many models. The majority of models perform at a level akin to random chance, with the exception of the davinci series and ChatGPT, which exhibit exceptional performance. One plausible explanation for the superior performance of these three models may be their substantial increase in the number of model parameters. This substantial increase in parameter count may enable the model to learn more complex representations and subsequently result in an improvement of performance. Another possible reason for the success of ChatGPT is its ability to understand the conversational context of DDXPlus, which consists of doctor-posed diagnostic questions and patient responses. ChatGPT has demonstrated superior performance in understanding conversational context compared to previous models, which likely contributes to its improved performance on this dataset.

Finally, it is worth noting that due to the critical nature of the healthcare field, **ChatGPT does not provide definitive answers in medical-related questions but instead offers informed suggestions and analysis, followed by a recommendation for further offline testing and consultation to ensure accurate diagnosis.** When the provided information is insufficient to make a judgment, ChatGPT will acknowledge this and offer an explanation, demonstrating its responsible approach to medical-related inquiries. This highlights the benefits of using ChatGPT for medical-related inquiries compared to search engines, as it can provide comprehensive analysis and suggestions without requiring the users to have medical expertise, while also being responsible and cautious in its responses. This suggests a promising future for the integration of ChatGPT in computer-aided diagnosis systems.

4.2 Zero-shot Machine Translation

4.2.1 Setup

We further evaluate the adversarial robustness of ChatGPT on an English-to-Chinese (En $\rightarrow$ 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 and Thottingal, 2020) and Trans-OPUS-MT-EN-ZH[9] 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.[10]

4.2.2 Results

The results of zero-shot machine translation are shown in Table 3. 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.

Table 3: 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	<u>59.28</u>
text-davinci-003	**30.60**	**40.01**	**61.88**
ChatGPT	<u>26.27</u>	<u>37.29</u>	58.95

4.3 Case Study

Table 4 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. Table 6 presents some cases of ChatGPT on OOD inputs. 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.

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

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

Table 4: 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

5 Discussion

5.1 Adversarial Attack Remains a Major Threat

As discussed in experiments, dealing with adversarial inputs still remains challenging to large foundation models. 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.

As for those who cannot train large models and only use them in downstream tasks, such threat still exists due to the defect inheritance of pre-trained models. In this case, how to achieve perfect fine-tuning or adaptation performance on downstream tasks while certainly reducing the defect inheritance remains a major challenge. Luckily, some pioneering work (Zhang et al., 2022b; Chin et al., 2021) might provide solutions. This represents a novel and emerging direction for future research. However, as foundation models grow larger that go beyond the capabilities of most researchers, reducing the defects through fine-tuning could be impossible. An open question rises for both model owners and downstream users on how to defend the adversarial attack.

In addition to adversaries in training data, prompts can also be attacked (Maus et al., 2023), which requires further knowledge and algorithms to deal with. This is currently a challenging problem due to the sensitivity of prompting to LLMs.

5.2 Can OOD Generalization be Solved by Large Foundation Models?

Larger models like ChatGPT and text-davinci-003 have the potential to achieve superior performance on OOD datasets with better prompt engineering, inspiring us to think of the problem: is OOD generalization solved by these giant models? The huge training data and parameters are a double-edged sword: overfitting vs. generalization. It is also intuitive to think that OOD data is unseen during training, so adding it into training set is enough, which is what these large models did. Is the "unreasonable effectiveness of data" (Sun et al., 2017) real? However, as the model sizes are becoming larger, it still remains unknown when and why LLMs will overfit.

Another possible reason is the training data of ChatGPT and text-davinci-003 actually encompass similar distributions to our test sets even if they are collected after 2021. Flipkart is for product review and DDXPlus is for medical diagnosis, which in fact are common domains widely existing on the Internet. Thus, they could be not OOD to these models, that could lead to overfitting. New datasets from long-tailed domains are in need for more fair evaluations.

Finally, our analysis does not show that ID-OOD performances are always positively correlated (Miller et al., 2021), but can sometimes inversely correlated (Teney et al., 2022). Regularization and other techniques should be developed to improve the OOD performance of language models.

5.3 Beyond NLP Foundation Models

Adversarial and OOD robustness do not only exist in natural language, but also in other domains. In fact, most research comes from machine learning and computer vision communities. Researchers in computer vision area could possibly think: can we solve OOD and adversarial robustness in image data by training a vision foundation model? For instance, the recent ViT-22B (Dehghani et al., 2023) scales vision Transformer (Dosovitskiy et al., 2020) to 22 billion parameters by training it on the 4 billion JFT dataset (Zhai et al., 2022) (a larger version of the previous JFT-300M dataset (Sun et al., 2017)), which becomes the largest vision foundation model to date. ViT-22B shows superior performance on different image classification tasks. However, it does not show "emergent abilities" (Wei et al., 2022) with the increment of parameters as other LLMs. Not only LLMs, the robustness in other areas also remains to be solved.

Back to theory, algorithms, and optimization areas, which foundational research areas in artificial intelligence. Will the large foundation models disrupt these areas? First, we should acknowledge that the success of foundation models should also attribute to these areas, e.g., most LLMs adopt the Transformer (Vaswani et al., 2017) and other advanced learning and training research. Second, the success of foundation models shed light on these areas: can we solve the problems like adversarial and OOD by developing new theories, algorithms, and optimization methods? Such research could offer valuable contribution to foundation models, e.g., improve the data and training efficiency and efficacy. Finally, researchers in these areas should not be dis-encouraged since the advance of scientific research should be diverse and not restricted to those done with rich computing resources.

6 Limitation

This paper offers a preliminary empirical study on the robustness of large foundation models, which has the following limitations.

First, we only perform zero-shot classification using ChatGPT and other models. Results of these models could change if we perform fine-tuning or adaptation given enough computing resources. But as we stated in introduction, it is expensive and un-affordable to perform further operation on today's latest foundation models, we believe zero-shot evaluation is reasonable.

Second, it seems controversial to evaluate large foundation models on small datasets in this work. However, since the training data of ChatGPT and some large models remains unclear, it is difficult to find larger datasets. Especially, ChatGPT is trained on huge datasets on the Internet as of 2021, making it more difficult to find appropriate datasets for thorough evaluation. We do believe more datasets can be used for such evaluation.

Third, we did most evaluations on text classification and only minor evaluations on machine translation. It is well-known that ChatGPT and other foundation models can do more tasks such as generation. Again, because of lack of appropriate datasets, evaluating generation performance is also difficult. We also admit that introducing more proper prompts could improve its performance.

Fourth, it is worth noting that ChatGPT is mainly designed to be a chatbot service rather than a tool for text classification. Our evaluations are mainly for classification, which have nothing to do with the robustness of ChatGPT for online chatting experience. We do hope every end-user can find ChatGPT helpful in their lives.

Finally, we could further provide detailed synopsis by conducting experiments on data before 2021 as comparisons and analyzing more OOD cases to see why ChatGPT succeeds or fails. Other experiments include detailed ablation study using different language models and investigation of induced outputs by LLMs through prompts. These can be done in future work. Another claim is that ChatGPT is not perfect for adversarial tasks. But we also need to develop certain metrics to show 'how good' is the performance.

7 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.

Acknowledgement

This paper received attentions from many experts since its first version was released on ArXiv. Authors would like to thank all who gave constructive feedback to this work.

Disclaimer

Potential Ethics and Societal Concerns raised by ChatGPT Robustness The increasing popularity of ChatGPT and other chatbot services certainly face some new concerns from both ethics and society. The purpose of this paper is to show that ChatGPT can be attacked by adversarial and OOD examples using existing public dataset, but not to attack it intentionally. We hope that this will not be leverage by end-users. Finally, we also hope the community can realize the importance of robustness research and develop new technologies to make our systems more secure, robust, and responsible.

ChatGPT usage Some authors in this paper are from mainland China where ChatGPT is currently unavailable. In order to conduct this research without disobeying local laws and OpenAI service terms, Hao Chen, who is one of our coauthors and lives in U.S., did all experiments related to ChatGPT and OpenAI. All experiments on ChatGPT are based on its Feb 13 version. Further updates of ChatGPT may lead to change of the results in this paper.

The contribution of each author Jindong led the project, designed experiments, wrote the code framework, and wrote the paper. Xixu and Wenxin shared equal contributions. Xixu was in charge of processing, experimenting, and writing about the DDXPlus and ANLI datasets. Wenxin designed all prompts to generative models and wrote about this part. Hao did the machine translation experiments, wrote necessary codes, and was in charge of code organization and reproducibility. Runkai helped polish the paper and organized case study. Other authors actively participated in this project from day one, reviewed the paper carefully, and provided valuable comments to improve this work.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Amos Azaria. Chatgpt usage and limitations. 2022.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

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.

Som Biswas. Chatgpt and the future of medical writing, 2023.

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.

Ting-Wu Chin, Cha Zhang, and Diana Marculescu. Renofeation: A simple transfer learning method for improved adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3243–3252, 2021.

Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. Chatgpt goes to law school. *Available at SSRN*, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

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.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. *arXiv preprint arXiv:2302.05442*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Roberto Gozalo-Brizuela and Eduardo C Garrido-Merchan. Chatgpt is not all you need. a state of the art review of large generative ai models. *arXiv preprint arXiv:2301.04655*, 2023.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023.

Philipp Hacker, Andreas Engel, and Marco Mauer. Regulating chatgpt and other large generative ai models. *arXiv preprint arXiv:2302.02337*, 2023.

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.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, et al. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *arXiv preprint arXiv:2212.14882*, 2022.

Mohammad Khalil and Erkan Er. Will chatgpt get you caught? rethinking of plagiarism detection. *arXiv preprint arXiv:2302.04335*, 2023.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=e2TBb5y0yFf`.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

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*, pages 7871–7880, 2020.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), 2022.

Muneer M Alshater. Exploring the role of artificial intelligence in enhancing academic performance: A case study of chatgpt. *Available at SSRN*, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Natalie Maus, Patrick Chao, Eric Wong, and Jacob Gardner. Adversarial prompting for black box foundation models. *arXiv preprint arXiv:2302.04237*, 2023.

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*, pages 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.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

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*, pages 4885–4901, 2020b.

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

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*, 2023.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. In *Advances in neural information processing systems*, pages 8735–8745, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

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*, pages 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.

Yiqiu Shen, Laura Heacock, Jonathan Elias, Keith D Hentel, Beatriu Reig, George Shih, and Linda Moy. Chatgpt and other large language models are double-edged swords, 2023.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

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*, pages 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*, pages 843–852, 2017.

Teo Susnjak. Chatgpt: The end of online exam integrity? *arXiv preprint arXiv:2212.09292*, 2022.

Wilbert Tabone and Joost de Winter. Using chatgpt for human–computer interaction research: A primer. 2023.

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.

Eva AM van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L Bockting. Chatgpt: five priorities for research. *Nature*, 614(7947):224–226, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. 2019. In the Proceedings of ICLR.

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.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 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*, pages 1112–1122. Association for Computational Linguistics (ACL), 2018.

Yaqi Xie, Chen Yu, Tongyao Zhu, Jinbin Bai, Ze Gong, and Harold Soh. Translating natural language to planning goals with large-language models. *arXiv preprint arXiv:2302.05128*, 2023.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. Glue-x: Evaluating natural language understanding models from an out-of-distribution generalization perspective. *arXiv preprint arXiv:2211.08073*, 2022.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022.

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*, 2022a.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*, 2021.

Ziqi Zhang, Yuanchun Li, Jindong Wang, Bingyan Liu, Ding Li, Yao Guo, Xiangqun Chen, and Yunxin Liu. Remos: reducing defect inheritance in transfer learning via relevant model slicing. In *Proceedings of the 44th International Conference on Software Engineering*, pages 1856–1868, 2022b.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867*, 2023.

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.

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 and 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 and 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., 2022a) 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 [11] 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 5.

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

Table 5: 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 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

E.2 OOD Case Study

We list some of the OOD examples for case study in Table 6.