GLUE-X: Evaluating Natural Language Understanding Models from an Out-of-Distribution Generalization Perspective

Linyi Yang[1,2][*], **Shuibai Zhang**[1,4,*], **Libo Qin**[3], **Yafu Li**[1], **Yidong Wang**[1], **Hanmeng Liu**[1], **Jindong Wang**[5], **Xing Xie**[5], **Yue Zhang**[1,2] [†]

[1] School of Engineering, Westlake University
[2] Institute of Advanced Technology, Westlake Institute for Advanced Study
[3] School of Computer Science and Engineering, Central South University
[4] University of Electronic Science and Technology of China
[5] Microsoft Research Asia

Abstract

Pre-trained language models (PLMs) are known to improve the generalization performance of natural language understanding models by leveraging large amounts of data during the pre-training phase. However, the out-of-distribution (OOD) generalization problem remains a challenge in many NLP tasks, limiting the real-world deployment of these methods. This paper presents the first attempt at creating a unified benchmark named GLUE-X for evaluating OOD robustness in NLP models, highlighting the importance of OOD robustness and providing insights on how to measure the robustness of a model and how to improve it. The benchmark includes 15 publicly available datasets for OOD testing, and evaluations are conducted on 8 classic NLP tasks over 21 popularly used PLMs. Our findings confirm the need for improved OOD accuracy in NLP tasks, as significant performance degradation was observed in all settings compared to in-distribution (ID) accuracy.

1 Introduction

Pre-trained Language Models (PLMs) (Qiu et al., 2020; Bommasani et al., 2021) have achieved competitive performance across standard NLP benchmarks (Blasi et al., 2022), such as GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a). However, recent studies (Gururangan et al., 2018; Ribeiro et al., 2019; Kaushik et al., 2020; Ribeiro et al., 2020; Ruder, 2021) show concerns that models are yet not close to achieving proper natural language understanding, essential questions being raised about their robustness (Srivastava et al., 2020; Wang et al., 2021b) and underlying sensitivity to systematic biases (Niven and Kao, 2019; Sagawa et al., 2020). Such issues manifest

in the performance decay, especially for out-of-distribution (OOD) generalization when the test distribution differs from training (Arora et al., 2021; Malinin et al., 2021; Hupkes et al., 2022).

OOD generalization has been systemically studied for Computer vision (CV) and artificial general intelligence (AGI) (Koh et al., 2021; Srivastava et al., 2022), for which large evaluation datasets are available. While sharing the same aspirational goal, existing evaluations (Kaushik and Lipton, 2018; Min et al., 2019; Gardner et al., 2020) and methods (Hendrycks et al., 2020; Bommasani et al., 2021) for OOD generalization of NLP contains only one or a few tasks (Wu et al., 2021; Wang and Culotta, 2021; Howard et al., 2022; Lu et al., 2022), which do not adequately capture limitations of existing models, resulting in inflated test accuracy (Tu et al., 2020; Ribeiro et al., 2020). Thus it remains a gap in evaluating models in a unified way by executing a range of text classification tasks.

To facilitate research in this direction, we introduce the GLUE-X benchmark for evaluating the out-of-distribution performance of PLMs. GLUE-X expands upon previous multi-task benchmarks (Zheng et al., 2022; Xu et al., 2020, 2021) by including test data from multiple domains, covering eight standard tasks in GLUE, with an average 2 test domains for each task, allowing comprehensive cross-distribution evaluations. Specifically, GLUE-X focuses on domain generalization, where a model trained on a source domain can be directly generalized to target domains without any labeled or unlabeled data from target domains. It also enables the analysis of two main factors affecting the cross-domain generalization performance, namely the pre-trained language model (e.g., architecture, size, etc.) and different training strategies (e.g., fine-tuning, prompt-tuning (Chen et al., 2022), linear probing (Wu et al., 2020), and domain-generalization training (Wang et al., 2023)).

Using GLUE-X, we evaluate the performance

---

[*] Equal contribution. Random order. Shuibai Zhang finished this work at Westlake University as an intern.
[†]{yanglinyi, zhangyue}@westlake.edu.cn
Yue Zhang is the corresponding author.

of *21* pre-trained language models in a unified setting and under the same experimental conditions. In addition, we consider *3* tuning strategies designed for improving single-source domain generalization: linear probing (Tripuraneni et al., 2020; Wu et al., 2020), fine-tuning, and the linear probing then fine-tuning method (LP-FT) (Kumar et al., 2022). Finally, we analyze the internal causes of OOD robustness at the feature level by measuring the rationale overlap between human and model predictions (Lei et al., 2016).

Results show that the average accuracy of PLMs on cross-domain evaluations falls significantly short of human performance, even for the highest-performing model (80.1% – human versus 74.6% – model). In contrast to the GLUE leaderboard, where over 20 single-model results outperform human baselines, none of the backbones included in GLUE-X is able to surpass human performance under the same evaluation setting. These findings suggest the importance of cross-distribution evaluation for natural language processing. In addition, evidence shows that the superior performance of PLMs on GLUE may be relatively superficial and less useful as a performance indicator in practice.

Detailed analysis shows that (1) no one backbone can significantly outperform the others across all tasks, which is consistent with the conclusion (Wenzel et al., 2022) in the computer vision; (2) surprisingly, the influence of model architectures is somehow more significant than the model parameters towards the OOD robustness; (3) the ID and OOD performance holds a linear correlation in most cases for text classifications; (4) in terms of the tuning strategy, we show that linear probing and then fine-tuning can slightly improve the OOD performance compared to standard fine-tuning.

To our knowledge, we are the *first* to systemically evaluate natural language understanding systems for cross-distribution generalization on genuine data compared to human performance. More importantly, we make datasets of cross-domain evaluations for all typical text classification tasks, which allows us to report OOD results under the same experimental conditions. We open-source the codebase and datasets [1]. The GLUE-X leaderboard is available at https://gluexbenchmark.com/.

---

[1] https://github.com/YangLinyi/GLUE-X

2 Related Work

Benchmarking Robustness to OOD. Recent work (Ibrahim et al., 2022) finds that today's leading PLMs are not robust to changing domains, where some OOD test samples varied during training. In particular, pre-trained transformers can rely heavily on spurious patterns (artefacts) (Gururangan et al., 2018; Kaushik et al., 2020; Tu et al., 2020). For this reason, the standard held-out accuracy can overestimate the performance (Sagawa et al., 2020; Kaushik et al., 2021), and evaluating the OOD robustness is crucial for real-world applications, which require models to hold good transferability. Consequently, there is a rising concern about improving dataset and benchmark development. Recent work introduces new out-of-distribution benchmarks for graphs (Gui et al., 2022), optical character recognition (OCR) (Larson et al., 2022), computer vision (CV) (Ibrahim et al., 2022), time series tasks (Gagnon-Audet et al., 2022), and artificial general intelligence (AGI) (Koh et al., 2021; Srivastava et al., 2022). However, evaluating the out-of-distribution generalization in a multi-task setting has received relatively little attention for NLP.

There is a line of work focusing on the development of challenge datasets, representing as Adversarial NLI (Nie et al., 2020), Dynabench (Kiela et al., 2021), Contrastive Set (Gardner et al., 2020), and AdvGLUE (Wang et al., 2021a) where examples are created to be difficult for current models via an iterative, adversarial, and human-and-model-in-the-loop procedure. However, these datasets focus on robustness and stability issues rather than generalization and the artifact. In contrast, GLUE-X contains both off-the-shelf and self-collected datasets to implement cross-distribution tests.

Prior work (Wenzel et al., 2022) observed that OOD performance holds a linear correlation with ID accuracy in CV based on 172 publicly available datasets and 31k networks, while their relationship is largely dataset-dependent. However, this conclusion has been found somewhat controversial, as Teney et al. (2022) argue that the selection of datasets influences the OOD performance.

Existing Benchmarks for NLU. There have been different types of leaderboards towards evaluating natural language understanding (NLU) systems. Examples of building challenging benchmarks in recent years include GLUE (Wang

et al., 2019b), SuperGLUE (Wang et al., 2019a), FewGLUE (Schick and Schütze, 2020), FEVER (Petroni et al., 2021), FewNLU (Zheng et al., 2022), and AdvGLUE (Wang et al., 2021a). In particular, FewGLUE and FewNLU focus on the few-shot learning challenge. The performance decay of NLP models has been found in real-world deployment because of the arises of the OOD generalization challenge as well as robustness issues, such as adversarial robustness. Similar to our work, other benchmarks, such as AdvGLUE (Wang et al., 2021a), leverage the training set extracted from GLUE for each task. Differently, we consider evaluating OOD performance in a general multi-task setting, where the test data arise from one or more different distributions.

Domain Generalization (DG) (Wang et al., 2022a) aims to learn a generalized model that is robust to unseen distributions using training data from multiple domains (Balaji et al., 2018; Dou et al., 2019; Vu et al., 2022; Varshney et al., 2022). We focus on the single-source DG (Huang et al., 2020; Krueger et al., 2021; Wang et al., 2022a) setting, which is a popular setting for measuring the OOD robustness in NLP (Hendrycks et al., 2020), and aligns with the GLUE leaderboard. As stated by a recent taxonomy and review towards generalisation research in NLP (Hupkes et al., 2022), current work does not provide standardized data or procedures for generalization testing, while we use GLUE-X as the first attempt towards this goal.

3 Data and Settings

The goal of GLUE-X is to provide an extension of GLUE with the same training data but multifarious OOD test sets.

3.1 Overview of GLUE-X

The evaluation in GLUE-X is intrinsically related to the domain generalization task considering a practical and challenging setting, where a model trained on multiple source domains can be directly generalized to a target domain without any labeled or unlabeled data from the target domain (Muandet et al., 2013). We articulate the following tasks and datasets in GLUE-X.

Tasks. As a benchmark styled after GLUE (Wang et al., 2019b), we consider eight tasks in GLUE-X: Sentiment Analysis (*SST-2*), Natural Language Inference (*MNLI, QNLI*), Textual Entailment (*RTE*),

Task	ID	OOD	Size
Sentiment Analysis	SST-2	IMDB	50,000
		Yelp	598,000
		Amazon	4,000,000
		Flipkart	205,041
Linguistic Acceptability	CoLA	Grammar Test	304,277
Textual Similarity	STSB	SICK	9,840
Natural Language Inference	MNLI (matched)	MNLImis	9,832
		SNLI	570,152
		SICK	9,840
Question Answering NLI	QNLI	NewsQA (Reconstructed)	119,525
Textual Entailment	RTE	SciTail	26,527
		HANS	60,000
Paraphrase	MRPC	QQP	404,276
		Twitter	16,777
	QQP	MRPC	4,076
		Twitter	16,777

Table 1: Data statistic of GLUE-X, which describes the source and size for OOD tests over different tasks.

Paraphrase (*MRPC, QQP*), Textual Similarity (*STS-B*) and Linguistic Acceptability (*CoLA*). [2]
Datasets. GLUE-X follows the same in-domain training data and evaluation metrics as GLUE (Wang et al., 2019b). To construct the out-of-domain test, we adopt popular datasets extracted from different domains while keeping the same prediction labels as the original tasks in GLUE. The detailed data statistics are shown in Table 1.

3.2 Dataset Curation

We construct test sets for each task under the requirement that they share the same label types with the training set. To this end, GLUE-X contains 15 OOD datasets, including publicly available datasets (Amazon, HANS, etc) and newly collected/re-constructed datasets (Grammar Test). In particular, we select the OOD datasets for each task, including sentiment analysis – IMDB (Maas et al., 2011), Yelp (Zhang et al., 2015), Amazon (Kaushik et al., 2020), and Flipkart (Vaghani and Thummar, 2023); linguistic acceptability – Grammar Test; textual similarity – SICK (Zhang et al., 2018); NLI – MNLI-Mismatched (Williams et al., 2017), SNLI (Bowman et al., 2015), and SICK (Zhang et al., 2018); Textual Entailment – RTE; Paraphrase – MRPC and QQP (Bentivogli et al., 2009; Dolan and Brockett, 2005; Wang et al., 2017;

---

[2]The WNLI task is not included in GLUE-X since there is no sufficient in-domain data for constructing OOD tests (Wang et al., 2022b; Yang et al., 2019, 2022).

McCoy et al., 2019). Regarding the QNLI task, we convert instances from NewsQA (Trischler et al., 2017) to the consistent data format of QNLI for conducting the OOD evaluation. The detailed description of the newly collected dataset, Grammar Test, can be found in Appendix A.

SICK contains multiple labels, including textual similarity, also used as an OOD test set of the textual similarity task. We rounded floating number labels of textual similarity to integers from 0 to 4, converting it into a five-class dataset to align with other classification tasks in GLUE-X. In addition, MRPC and QQP are leveraged as OOD datasets of each other as the paraphrasing task.

3.3 Metrics

We first average metrics to get a score for those tasks with multiple metrics. Following GLUE and SuperGLUE, we then report the score of NLU models by averaging the scores of all tasks as the OOD performance. For rankings, in addition to the robustness rank by considering the decreased ratio between OOD and ID performance, we adopt Friedman rank (Friedman, 1940) over multiple tasks:

$$\text{rank}_F = \frac{1}{n} \sum_{i=1}^{n} \text{rank}_i, \quad (1)$$

where n is the number of tasks (e.g., $n = 8$ in Table 3) and rank_i is the rank of the performance in the i-th task considering in the GLUE-X. We report the robustness ranking in terms of the decreased ratio of OOD performance and Friedman rank.

3.4 Post-hoc Analysis

In addition to quantitative analysis, we choose two tasks, sentiment analysis, and natural language inference, for post-hoc feature analysis (Lei et al., 2016). We adopt the sensitivity of contextual decomposition technique (Jin et al., 2019; Yang et al., 2021), which removes part of inputs from the sequence text to evaluate a model's sensitivity to them, thereby allowing for identifying important features. The output is the overlap between rationales by models and humans, which to some extent represents the trust of models (Jacovi and Goldberg, 2020; Yang et al., 2020).

Formally, given a phrase p starting with the negative limitations in the k-th document $\mathcal{D}^{(k)}$, we sample the documents which contain the same phrase p to alleviate the influence by chance when there

Type Backbone	Training (GPU Hours)	Inference (GPU Hours)	Total Hours
BERT-large	440	240	680
T5-large	792	420	1,212
ALBERT-base	165	120	285

Table 2: The training and testing cost of GLUE-X.

are multiple shreds of evidence saturating the prediction. The window size of the phrase p is limited to 3. Taking sentiment analysis for example, given *"This movie was so unbelievably bad"* if we only remove the non-causal word *movie*, the prediction is not expected to change for a robust model.

The importance score is computed as follows:

$$\phi(\mathbf{p}, \widehat{\mathcal{D}^{(k)}}) = \mathbb{E}_{\widehat{\mathcal{D}^{(\beta)}}}[l\left(\widehat{\mathcal{D}^{(\beta)}}; \widehat{\mathcal{D}}\right) - \\ l\left(\widehat{\mathcal{D}^{(\beta)}} \backslash \mathbf{p}; \widehat{\mathcal{D}}\right)] \quad (2)$$

where $\mathcal{D}^{(\beta)}$ denotes the resulting text after masking out a single token (phrase) starting with the negative pronoun (un-, non-, etc.) in the length of N surrounding the phrase $\mathbf{p}$. We use $l\left(\widehat{\mathcal{D}^{(\beta)}} \backslash \mathbf{p}; \widehat{\mathcal{D}}\right)$ to represent the model prediction logits of the ground-truth class after replacing the masked-out context. $\backslash \mathbf{p}$ indicates the operation of masking out the phrase p in a given document.

3.5 Models and Training Strategies

Models. To ensure that our results are relevant for both researchers and practitioners, we consider both top-performing model backbones and cost-efficient methods: *Discriminative Models* – BERT-base, BERT-large (Devlin et al., 2018), RoBERTa-base, RoBERTa-large (Liu et al., 2019), XLNet-base, XLNet-large (Yang et al., 2019); *Generative Models* – BART-base, BART-large (Lewis et al., 2020), T5-small, T5-base, T5-large (Raffel et al., 2020), GPT2, GPT2-medium, GPT2-large (Radford et al., 2019); *Generative and Discriminative Models* – ELECTRA-small, ELECTRA-base, ELECTRA-large (Clark et al., 2020); *Cost-Efficient Models* – ALBERT-base (Lan et al., 2020), and DistilBERT-base (Sanh et al., 2019). We also report the results of GPT-3 (Brown et al., 2020) and GPT-3.5 (OpenAI, 2023) through in-context learning. We follow the official implementations of several pre-trained language models from Huggingface[3] to reproduce results on GLUE using the validation set and test these models on GLUE-X.

---

[3]https://huggingface.co/models

The hyper-parameters of each model are selected by using grid search and can be found in Appendix. **Fine-tuning Strategies.** We investigate the efficacy of different fine-tuning strategies for OOD generalization. In particular, we consider three paradigms: standard fine-tuning, fine-tuning only the head (linear probing), and linear probing then fine-tuning. The detailed training cost and inference speed estimated by a single V100 are shown in Table 2, in which we evaluate the performance using the in- and out-of-domain test data, recording the training cost in GLUE and GLUE-X. We use 50 NVIDIA Tesla V100 GPU cards and 8 NVIDIA A100 GPU cards and spend 10,000+ GPU hours based on the estimation with a single V100 card.

4 Experiments

We explore the facets of OOD generalization in NLP using GLUE-X, highlighting discrepancies to previous findings and discussing their implications.

4.1 Human Annotation

We employ human annotators to give predictions on OOD datasets and identify rationales.
Predictions. We use a crowd-sourcing company to recruit editors and annotators to give predictions on 15 OOD datasets. To fairly compare human performance with models, we simulate the models' OOD testing during the manual annotation process. Specifically, annotators are given essential instructions and a few examples from the in-domain dataset that gently guide them to annotate. Then they are asked to label instances from unseen OOD datasets, typically collected from other domains. 1,000 testing samples are used to obtain the human performance for each OOD dataset.

We employ multiple labelers to annotate the same data point (1,000 samples for each dataset) during the annotation to ensure the high quality of the crowdsourcing work. All annotators have an undergraduate degree in English or a PhD in an English-speaking country. In particular, we employ ten people to annotate the SICK dataset as same as the original data (Zhang et al., 2018). We employ two annotators for labeling the same instance for the other datasets. After the trial phase of data annotation, we set the Inter-Annotator Agreement (IAA) score threshold for each task depending on the difficulty level. Finally, the average IAA over the 15 OOD datasets is 0.857, indicating acceptable agreement.

Rationale Marking. Following Kaushik et al. (2020) and Kaushik et al. (2021), we use extractive explanations for marking rationales that support classification decisions. Inspired by Kaushik et al. (2021) and Lertvittayakumjorn and Toni (2021), we leverage the rationale marking annotated by humans to compare with rationale selected by models on sentiment analysis and natural language inference (NLI) tasks. We ask two labelers to annotate sampled instances from IMDB, Yelp, and Amazon datasets for the sentiment analysis task. At the outset, annotators were given instructions and examples that gently guided them to annotate rationales. Only adjectives, adverbs, nouns, and verbs were considered rationale candidates. Besides, rationales were required to carry complete semantic information. We sampled 6,000 instances for each dataset randomly. Using F1 score, the IAA for IMDB, Yelp and Amazon are 0.874, 0.871, and 0.840, respectively. For NLI, we use the explanation dataset, e-SNLI (Camburu et al., 2018), to assert the models' trust.

4.2 Prediction Results

Overall Performance on GLUE-X. We report the average score of different models sorted in descending order representing the overall performance in Table 3. In addition to the overall performance, we provide the Friedman Rank for in- and out-of-domain results. From Table 3, we observe that all pre-trained models involved in GLUE-X show significant performance decay under the OOD test compared to the ID performance (**20.05%** decay in average). The results also suggest no significant difference in the OOD robustness between generative and discriminative models for text classification. We also provide the results of GPT-3 with in-context learning in Appendix G since it leverages a different training strategy.

Model-level Analysis. On the model level, we observe that ELECTRA-large achieves the best performance for both ID (**89.18%**) and OOD (**74.62%**) tests. Lightweight models, BERT-base, GPT-2, and DistilBERT-base, are in the bottom three on GLUE-X with the lowest OOD performance. In contrast, the base-size ELECTRA and ALBERT achieve comparable generalization results. Moreover, by comparing the Friedman rank of OOD and ID tests in Table 3, we observe that the fluctuation of the OOD F-rank is slightly lower than the ID F-rank, which hints that the uncertainty of performance

Pre-trained Models	Avg GLUE-X	Avg GLUE	Avg Δ↓	F-Rank OOD	F-Rank ID	Rank Δ↓	PARAM (M)
ELECTRA-large (Clark et al., 2020)	**74.62**	**89.18**	**16.33**	**2.13**	**2.25**	**1**	334.09
T5-large (Raffel et al., 2020)	72.81	87.70	16.98	2.38	3.00	2	737.67
RoBERTa-large (Liu et al., 2019)	71.62	87.83	18.46	4.00	3.00	3	355.36
BART-large (Lewis et al., 2020)	70.38	87.05	19.15	5.00	3.63	6	406.29
T5-base (Raffel et al., 2020)	70.05	85.92	18.47	5.88	6.13	4	222.90
XLNet-large (Yang et al., 2019)	69.69	86.75	19.67	6.00	4.63	8	360.27
RoBERTa-base (Liu et al., 2019)	68.73	85.27	19.40	7.00	6.63	7	124.65
ELECTRA-base (Clark et al., 2020)	67.78	85.92	21.11	9.63	8.63	15	108.89
GPT2-large (Radford et al., 2019)	66.46	83.57	20.47	10.88	11.50	10	**774.03**
BART-base (Lewis et al., 2020)	65.89	83.04	20.65	11.00	11.00	12	139.42
BERT-large (Devlin et al., 2018)	65.80	83.26	20.97	11.38	10.38	14	335.14
T5-small (Raffel et al., 2020)	65.43	80.35	18.57	12.63	15.00	5	60.51
ALBERT-base (Lan et al., 2020)	65.30	82.58	20.93	12.88	13.25	13	11.68
ELECTRA-small (Clark et al., 2020)	65.06	81.50	20.17	13.88	16.13	9	13.48
GPT2-medium (Radford et al., 2019)	65.03	81.84	20.54	12.88	13.63	11	354.82
XLNet-base (Yang et al., 2019)	64.57	82.26	21.50	12.75	12.13	16	116.72
BERT-base (Devlin et al., 2018)	64.10	82.08	21.91	13.88	13.88	17	109.48
DistilBERT-base (Sanh et al., 2019)	61.94	80.21	22.78	17.75	17.38	18	66.36
GPT2 (Radford et al., 2019)	61.16	79.30	22.88	18.13	17.88	19	124.44

Table 3: Overall performance sorted by the GLUE-X performance. The average accuracy shown in the table is the mean average score of the OOD performance for each task. The average Δ↓ indicates the decreased ratio from the average ID accuracy to OOD accuracy. We also provide the Friedman rank (Friedman, 1940) for OOD and ID tests (shown as F-Rank). The robustness rank is sorted by the average ratio of performance decay in ascending order.

has been decreased on GLUE-X by using a large amount of the test data.

The Performance of Compressed Models. The results of GLUE-X suggest that OOD generalization still faces fundamental challenges, especially for lightweight models. For example, we find that compressed models (e.g., DistilBERT-base) show relatively low performance compared to others. Differently, the OOD performance of ALBERT-base (11M parameters) is significantly higher than DistilBERT-base (**65.30% vs. 61.94%**), even better than several moderate-sized models (BERT-large, GPT2-medium, and XLNet-base).

4.3 Discussion

Human vs. Model. The average performance decay between in- and out-of-domain tests of humans (87.10% – ID vs. 80.14% – OOD) is significantly lower than models, even for the best-performing model with the lowest performance decay (**7.82% vs. 16.33%**), as shown in Table 4. Regarding the average OOD performance, the human baseline is also much higher than the models, with at least an **6.69%** increase (80.14% vs. 74.62%)[4]. Such

[4]Note that the human performance of RTE, MRPC, and QQP is still adjusted and will be updated in the next version.

a large performance gap indicates that PLMs cannot achieve competitive results with humans on GLUE-X. More specifically, the human baseline outperforms the state-of-the-art results on five of eight tasks. It is noteworthy that we control OOD evaluations of humans in the same experimental setting with models by testing on unseen samples.

OOD Robustness. As shown in Table 4, we suggest that there is no silver bullet towards the OOD robustness, given that no single model can consistently outperform others over all tasks on GLUE-X. For example, ELECTRA-large can only achieve the best performance on four of eight tasks. We also find that the generalization for the CoLA dataset is the most challenging task for models since the test set holds the biggest difference with training data. In contrast, models tend to perform better on the relatively easy dataset, such as sentiment analysis (SST-2). For example, the best-performing model ELECTRA-large can achieve a **94.67%** accuracy on SST-2 yet only a **37.85%** Matthew's Corr on CoLA. Besides, we also observe that the distribution shift between the ID and OOD datasets largely influences the OOD generalization results. In particular, the performance decay on the OOD test is exacerbated by the increase of distribution shifts,

Model	SST-2 OOD	MNLI OOD	QNLI OOD	RTE OOD	MRPC OOD	QQP OOD	STS-B OOD	CoLA OOD	Avg OOD	Avg Δ↓
Human Performance	*92.36*	*84.13*	81.10	*86.53*	*79.31*	*78.46*	80.28	58.98	*80.14*	*7.82*
ELECTRA-large	**94.67**	76.94	80.44	78.74	69.96	**77.24**	81.14	**37.85**	74.62	16.33
T5-large	93.83	76.36	**81.72**	**81.52**	**72.66**	72.26	77.86	26.30	72.81	16.98
RoBERTa-large	93.07	77.28	79.67	77.84	65.19	76.11	77.91	25.90	71.62	18.46
BART-large	93.51	76.09	80.45	71.94	70.56	73.41	76.03	21.06	70.38	19.15
T5-base	93.52	73.76	80.29	73.85	70.90	73.20	74.98	19.94	70.05	18.47
XLNet-large	93.75	**77.59**	79.98	76.29	65.07	65.25	76.86	22.76	69.69	19.67
RoBERTa-base	92.77	74.21	79.55	67.82	66.89	70.90	74.90	22.81	68.73	19.40
ELECTRA-base	90.41	75.33	78.10	74.00	59.49	66.55	77.10	21.23	67.78	21.11
GPT2-large	91.60	73.62	75.77	70.54	62.05	70.16	70.61	17.32	66.46	20.47
BART-base	91.49	74.32	78.77	63.28	66.56	67.36	72.46	12.90	65.89	20.65
BERT-large	91.47	73.33	78.79	64.97	62.06	69.26	69.76	16.77	65.80	20.97
T5-small	89.40	70.67	77.44	63.96	71.63	67.24	72.58	10.55	65.43	18.57
ALBERT-base	89.97	70.95	77.31	65.70	61.08	68.04	72.17	17.18	65.30	20.93
ELECTRA-small	89.40	70.57	75.02	65.89	59.64	65.14	72.36	22.48	65.06	20.17
GPT2-medium	91.42	72.82	77.70	66.60	57.75	67.53	69.16	17.26	65.03	20.54
XLNet-base	91.54	74.75	76.87	63.47	62.34	65.76	68.29	13.53	64.57	21.50
BERT-base	89.36	70.92	78.31	59.54	61.83	67.49	67.68	17.66	64.10	21.91
DistilBERT-base	87.06	70.27	74.27	58.76	61.63	64.96	66.18	12.37	61.94	22.78
GPT2	82.46	69.67	76.41	60.55	58.90	64.79	66.26	10.22	61.16	22.88

Table 4: Detailed OOD performance for each task in GLUE-X. Evaluation metrics for each task are the same as GLUE (the average results are reported for those tasks considering two metrics). The best performance is shown in bold. Human evaluation is simulated in a similar OOD setting by receiving instructions from ID samples while predicting data from OOD datasets. The human baseline is shown in italics if it beats the best-performing model.

as shown in Appendix C.

Model Architectures vs. Parameter Size. The rightmost column of Table 4 demonstrates the decreased ratio representing the model robustness to some extent. Although the large-sized model, such as T5-large, and RoBERTa-large, can surpass the corresponding base-sized models in terms of the lower decreased ratio, empirical evidence from Table 4 also shows that model types could be more influential than the parameter size towards the OOD performance. Specifically, as shown in Table 4, results of the same architecture with different parameters are closer to the results of similar parameter-size models based on different architectures. For instance, the decreased ratio of T5 architectures pre-training with different parameter sizes (T5: 16.98% – large (737M); 18.47% – base (223M); 18.57% – small (61M)) are close to each other, similar to RoBERTa (18.46% –large vs. 19.40% – base). It hints that designing model architectures and training methods could be one of the future directions for improving OOD robustness.

Robustness vs. Trust. The average results of the rationale overlap between models and humans for three sentiment analysis tasks are shown in Table 5, indicating the trust measurement. As shown in the table, somehow surprisingly, we find that the best-performing discriminative model – RoBERTa-large

Model	Precision	Recall	F1
ELECTRA-small	**15.17**	42.48	**20.54**
ELECTRA-large	14.61	**43.96**	20.23
BERT-large	12.97	40.33	18.14
ALBERT-base	13.12	39.43	18.10
T5-large	12.78	39.09	17.74
ELECTRA-base	12.87	37.73	17.66
T5-base	12.69	37.98	17.42
BART-base	12.16	36.76	16.79
BERT-base	11.97	36.17	16.53
T5-small	12.03	35.41	16.49
BART-large	11.68	36.23	16.25
XLNet-large	11.63	36.38	16.18
RoBERTa-base	10.68	34.29	14.94
DistilBERT-base	10.56	31.87	14.55
GPT2-large	10.24	31.54	14.10
GPT2-medium	10.20	30.68	13.94
XLNet-base	9.86	30.90	13.68
GPT2	9.63	28.53	13.11
RoBERTa-large	7.93	27.88	11.45

Table 5: The average F1 score of the rationale overlap on three sentiment analysis tasks sorts the table.

(see Table 3) achieves the lowest rationale overlap between humans and models. While RoBERTa-large can achieve a relatively high overlap with humans on the NLI task (see Appendix D). This can be because the rationale overlap is largely influenced by datasets.

It is noteworthy that small-sized models can achieve relatively higher rationale overlaps than
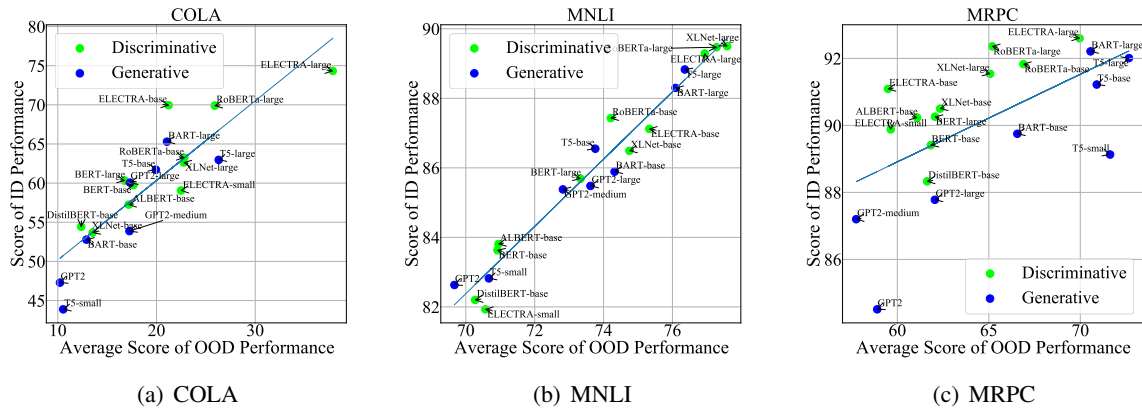
(a) COLA   (b) MNLI   (c) MRPC

Figure 1: Scatter figures that illustrate the correlation between ID and OOD performance for different tasks.
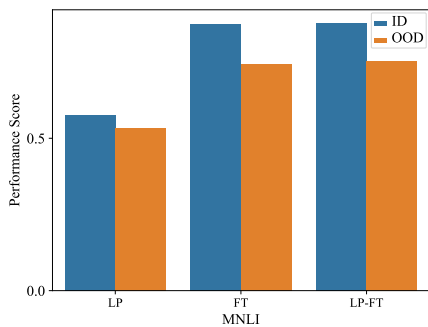

Figure 2: The influence of different tuning strategies on the task of MNLI, including Linear probing, Fine-tuning, and Linear probing then Fine-tuning (LP-FT). The results are based on RoBERTa-base.

large-sized models, which is generally consistent with the results reported by previous work (DeYoung et al., 2020). For instance, ELECTRA-small achieves the highest F1 score with only 13.48M parameters. In addition, the models pre-trained with the same architectures usually achieve similar performance (e.g., ELECTRA-small and ELECTRA-large, GPT2-medium and GPT2-large).

ID vs. OOD Performance. We show the correlation of three tasks between the in- and out-of-domain results in Figure 1 (the full results can be found at Appendix F). Unsurprisingly, we observe that the in-domain performance is always higher than the out-of-domain performance. Specifically, we find that the OOD performance is much lower than the ID performance in the task of COLA. In contrast, the gap between ID and OOD performance based on SST-2 and MNLI is relatively lower than others. We suppose this is partially influenced by the distribution shift between the in- and out-of-domain datasets.

Regarding the type of pre-trained models, we show that discriminative models show a stronger linear correlation when compared to generative models (19 data points). From the task perspective, we observe that datasets largely influence the correlation between ID and OOD. For instance, ID and OOD performance are inversely correlated on MRPC yet almost correlated on other tasks, hinting that the inverse correlation is possible for the specific task when the size of test samples is limited.

The Influence of Tuning Methods. Taking MNLI as an example, we compare the results of RoBERTa-base using three different training strategies in Figure 2. As found by previous work (Kumar et al., 2022), fine-tuning can do worse than linear probing in the presence of a large distribution shift in CV. However, as shown in Figure 2, we find that linear probing methods show relatively low accuracy for both ID and OOD tests, which is different from the conclusion in CV. This can be because freezing pre-trained features hinders the generalization of NLP tasks that are generally more complex than the OOD generalization in CV. While the LP-FT can be relatively helpful for improving the OOD robustness of NLP models in terms of the slight performance improvement compared to the standard fine-tuning method. For this reason, there is still much room to improve in designing methodologies of domain generalization that can improve the OOD robustness for text classification. In addition to tuning methods discussed in GLUE-X, the recently emerging trend of the development of large-scale language models (LLMs) represented by ChatGPT is worth paying attention to. In particular, how to appropriately define the OOD generalization for LLMs is still under-explored since the

pre-training corpus of these models is not disclosed yet (Wang et al., 2023).

5 Conclusion

We constructed GLUE-X, an OOD robustness benchmark for natural language understanding tasks that aim to enable fair evaluation over multiple datasets from multiple domains in a consistent setting. With GLUE-X, we evaluate 21 pre-trained models on 8 classification tasks, providing analysis using 3 different tuning strategies and post-hoc analysis for gaining internal causes for the OOD robustness. We conclude that (1) current PLMs still have a lag much behind human-level towards the OOD robustness; (2) the ID and OOD performance usually hold a linear correlation in most cases, while the coefficiency of the correlation is primarily related to the selection of OOD datasets; (3) stronger architectures can bring decent performance benefit, especially for the OOD performance.

Limitation

Our primary focus is on the OOD robustness of text classification tasks. However, there are other NLP tasks that the community should not ignore. GLUE-X currently does not include language generation tasks such as machine translation, summarization, and dialogue. Moreover, extending the current GLUE-X to more real-world datasets from different domains is of great importance. We aim to make GLUE-X a continuously maintained project.

Ethics Statement

This paper honors the ACL Code of Ethics. Public available datasets are used to establish the GLUE-X leaderboard. No private data was used. All annotators from the crowdsourcing company have received enough labor fees corresponding to their amount of annotated instances. The code and data are open-sourced under the CC-BY-NC-SA license.

Acknowledgement

References

Udit Arora, William Huang, and He He. 2021. Types of out-of-distribution texts and how to detect them. *ArXiv*, abs/2109.06827.

Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. 2018. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TACL*.

Damian E Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Annual Conference of the Association for Computational Linguistics (ACL)*, Dublin, Ireland.

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. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

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

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Yulong Chen, Yang Liu, Li Dong, Shuohang Wang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2022. Adaprompt: Adaptive model training for prompt-based nlp. *arXiv preprint arXiv:2202.04824*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. In *ACL*.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. 2019. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32.

Milton Friedman. 1940. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92.

Jean-Christophe Gagnon-Audet, Kartik Ahuja, Mohammad-Javad Darvishi-Bayazi, Guillaume Dumas, and Irina Rish. 2022. Woods: Benchmarks for out-of-distribution generalization in time series tasks. *arXiv preprint arXiv:2203.09978*.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. 2022. GOOD: A graph out-of-distribution benchmark. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*.

Phillip Howard, Gadi Singer, Vasudev Lal, Yejin Choi, and Swabha Swayamdipta. 2022. Neuro-counterfactuals: Beyond minimal-edit counterfactuals for richer data augmentation. *arXiv preprint arXiv:2210.12365*.

Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. 2020. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, pages 124–140. Springer.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. 2022. State-of-the-art generalisation research in nlp: a taxonomy and review. *arXiv preprint arXiv:2210.03050*.

Mark Ibrahim, Quentin Garrido, Ari Morcos, and Diane Bouchacourt. 2022. The robustness limits of sota vision models to natural variation. *arXiv preprint arXiv:2210.13604*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. 2019. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In *International Conference on Learning Representations*.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2020. Learning the difference that makes a difference with counterfactually augmented data. *International Conference on Learning Representations (ICLR)*.

Divyansh Kaushik and Zachary C Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015.

Divyansh Kaushik, Amrith Setlur, Eduard Hovy, and Zachary C Lipton. 2021. Explaining the efficacy of counterfactually augmented data. *International Conference on Learning Representations (ICLR)*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Stefan Larson, Gordon Lim, Yutong Ai, David Kuang, and Kevin Leach. 2022. Evaluating out-of-distribution performance on document image classifiers. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-based human debugging of nlp models: A survey. *arXiv preprint arXiv:2104.15135*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jinghui Lu, Linyi Yang, Brian Mac Namee, and Yue Zhang. 2022. A rationale-centric framework for human-in-the-loop machine learning. *arXiv preprint arXiv:2203.12918*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Andrey Malinin, Neil Band, Yarin Gal, Mark Gales, Alexander Ganshin, German Chesnokov, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, Vyas Raina, Denis Roginskiy, Mariya Shmatova, Panagiotis Tigas, and Boris Yangel. 2021. Shifts: A dataset of real distributional shift across multiple large-scale tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on International Conference on Machine Learning-Volume 28*, pages I–10.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. 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, Online. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

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

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2021. Kilt: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

Sebastian Ruder. 2021. Challenges and opportunities in nlp benchmarking.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. 2020. An investigation of why over-parameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Timo Schick and Hinrich Schütze. 2020. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. 2020. Robustness to spurious correlations via human annotations. In *International Conference on Machine Learning*, pages 9109–9119. PMLR.

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

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. 2020. On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*, 33:7852–7862.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Nirali Vaghani and Mansi Thummar. 2023. Flipkart product reviews with sentiment dataset.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022. Investigating selective prediction approaches across several tasks in iid, ood, and adversarial settings. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1995–2002.

Thuy Vu, Shahram Khadivi, Dinh Phung, and Gholamreza Haffari. 2022. Domain generalisation of nmt: Fusing adapters with leave-one-domain-out training. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 582–588.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021a. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*.

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

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2021b. Measure and improve robustness in nlp models: A survey. *arXiv preprint arXiv:2112.08313*.

Yile Wang, Linyi Yang, Zhiyang Teng, Ming Zhou, and Yue Zhang. 2022b. Pre-training a graph recurrent network for language representation.

Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *AAAI*.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4144–4150.

Florian Wenzel, Andrea Dittadi, Peter Vincent Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, et al. 2022. Assaying out-of-distribution generalization in transfer learning. *arXiv preprint arXiv:2207.09239*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Sen Wu, Hongyang R Zhang, and Christopher Ré. 2020. Understanding and improving information transfer in multi-task learning. In *International Conference on Learning Representations*.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. *arXiv preprint arXiv:2101.00288*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

Liang Xu, Xiaojing Lu, Chenyang Yuan, Xuanwei Zhang, Huilin Xu, Hu Yuan, Guoao Wei, Xiang Pan, Xin Tian, Libo Qin, et al. 2021. Fewclue: A chinese few-shot learning evaluation benchmark. *arXiv preprint arXiv:2107.07498*.

Linyi Yang, Eoin Kenny, Tin Lok James Ng, Yi Yang, Barry Smyth, and Ruihai Dong. 2020. Generating plausible counterfactual explanations for deep transformers in financial text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6150–6160, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Linyi Yang, Jiazheng Li, Padraig Cunningham, Yue Zhang, Barry Smyth, and Ruihai Dong. 2021. Exploring the efficacy of automatically generated counterfactuals for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 306–316, Online. Association for Computational Linguistics.

Linyi Yang, Lifan Yuan, Leyang Cui, Wenyang Gao, and Yue Zhang. 2022. Factmix: Using a few labeled in-domain examples to generalize to cross-domain named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5360–5371.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Li Zhang, Steven R Wilson, and Rada Mihalcea. 2018. Multi-label transfer learning for multi-relational semantic similarity. *arXiv preprint arXiv:1805.12501*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657.

Yanan Zheng, Jing Zhou, Yujie Qian, Ming Ding, Chonghua Liao, Li Jian, Ruslan Salakhutdinov, Jie Tang, Sebastian Ruder, and Zhilin Yang. 2022. FewNLU: Benchmarking state-of-the-art methods for few-shot natural language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 501–516, Dublin, Ireland. Association for Computational Linguistics.

A Data Collection

We derive the CoLA-OOD dataset from the Public High School English Exam, which contains 304,277 examples. The original multi-choice fill-in tests are converted into COLA-style, with correct answers as positive examples and incorrect answers as negative examples. The golden answer is given by the English teacher who is a native speaker or holds an English Teaching degree. We collect the data from publicly available internet resources, and the original open-access materials can be found from https://www.koolearn.com/shiti.

The input of the CoLA-OOD dataset, Grammar Test, is a text span containing a QA pair or a few sentences. The ground truth of the output is to decide whether the grammar of the sentence is acceptable or not. For example, given the sentence 'Is there a post office near here? Yes, there isn't .', the label is unacceptable since there is a grammar error existing in the input. Otherwise, for a sentence without any grammar errors, 'The young man is the CEO of the company, In other words, he is in charge of the company.', the corresponding label is acceptable.

B Training Details

We illustrate the cross-domain evaluation settings used for GLUE-X in Figure 3. Notably, the source domain only contains a single dataset, while target domains can include more than one dataset from multiple domains.

Tuning Strategies: Fine-tuning; Linear Probing; LP-FT

	Source Domain: Fine-Tuning	Source Domain: Linear-Probing	Source Domain: LP-FT
	Target Domains: Zero-shot Test	Target Domains: Zero-shot Test	Target Domains: Zero-shot Test

Figure 3: The demonstration of training and testing settings used for cross-domain evaluations in GLUE-X.

Regarding the training, we performed the grid search for each task, kept the best-performing checkpoint in ID datasets, and tested their performance on their corresponding OOD datasets. The hyperparameters used by these weights can be seen in Table 9.

C Domain Distributions

We evaluate distribution shifts between different datasets regarding Maximum Mean Discrepancy

Model	F1	Precision	Recall
ELECTRA-base	**34.98**	**31.06**	52.94
RoBERTa-large	34.89	30.57	**53.95**
XLNet-large	34.73	30.64	53.32
ELECTRA-large	34.37	30.34	52.67
RoBERTa-base	33.78	30.06	51.36
T5-large	33.70	29.41	52.00
GPT2-large	33.37	29.95	49.51
XLNet-base	33.16	29.42	49.95
GPT2-medium	33.06	29.46	49.65
BERT-large	32.96	29.45	49.88
DistilBERT-base	32.71	29.40	48.86
GPT2	32.36	29.10	48.29
ALBERT-base	32.34	28.93	49.09
BART-base	32.31	28.90	49.51
T5-base	32.29	28.50	49.30
ELECTRA-small	31.96	28.68	47.83
BERT-base	31.52	28.30	47.26
BART-large	31.31	28.04	47.80
T5-small	30.93	27.45	47.30

Table 6: The rationale overlap based on e-SNLI sorted by descending order of the F1 score.

(MMD) and Word Overlap Rate. MMD distance focuses on the semantic distribution shift between datasets, while Word Overlap Rate pays more attention to superficial similarity.

C.1 Word Overlap

The similarity between datasets of In-distribution datasets and Out-of-distribution datasets are shown in Figure 4.

C.2 MMD Distance

The MMD distance between ID and OOD datasets is shown in Figure 5 for each task including in GLUE-X. When computing the MMD distance between two datasets, we ensure that the same number of sentences are sampled and fed into PLMs (e.g. RoBERTA-base) to extract their semantic features. We sample multiple times to get an average MMD sample score to estimate MMD distance of two datasets. The calculation of MMD is shown as follows:

$$\text{MMD}^2[\mathcal{F}, X, Y] = \frac{1}{m(m-1)} \sum_{i \neq j}^{m} k\left(x_i, x_j\right)$$

$$+ \frac{1}{n(n-1)} \sum_{i \neq j}^{n} k\left(y_i, y_j\right) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k\left(x_i, y_j\right)$$

$$(3)$$

$\mathcal{F}$ is a MMD function class, i and j represents the batch of instances sampled from different distributions. m and n are the size of i and j.
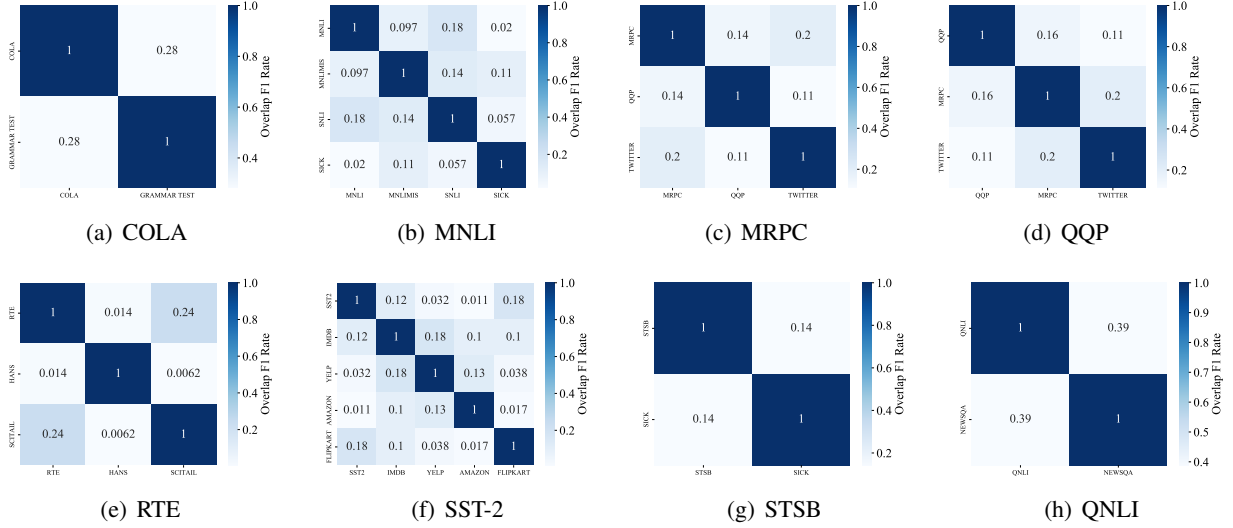
(a) COLA   (b) MNLI   (c) MRPC   (d) QQP

(e) RTE   (f) SST-2   (g) STSB   (h) QNLI

Figure 4: The word-level overlap between the training set and test set for each task.

(a) COLA   (b) MNLI   (c) MRPC   (d) QQP
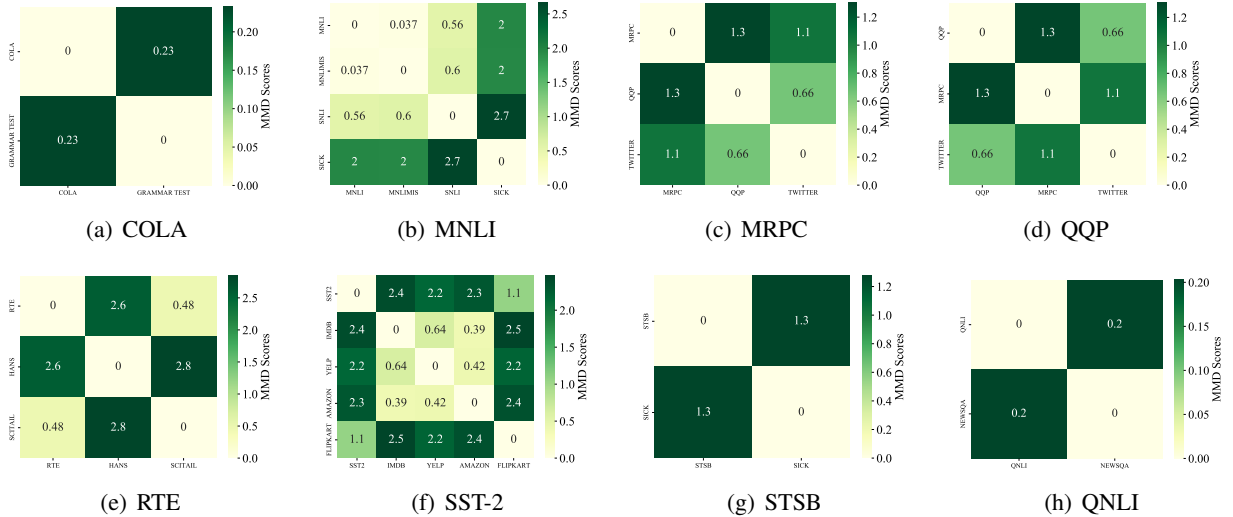
(e) RTE   (f) SST-2   (g) STSB   (h) QNLI

Figure 5: The MMD Scores between the training set and test set for each task. Lower MMD score means the higher correlation between datasets.

D Rationale Overlap

In order to measure the difference between rationales detected by PLMs and humans, we define precision as the percentage of the predicted rationales that also exist in the human annotation and recall as the percentage of words in the human annotation that also exist in the predicted rationales. We calculate the F1 score as an evaluation metric of overlap.

We show the evaluation of rationale overlap between models and humans on the e-SNLI dataset (Camburu et al., 2018) in Table 6. We find that the performance gap between different models is not very large (varying from 30.93 to 34.98). Models show a higher rationale overlap with humans

based on e-SNLI than sentiment analysis datasets. This can be because the average length of instances in e-SNLI is generally shorter than that in sentiment analysis datasets. In particular, the base-sized ELECTRA has achieved the highest F1 score (34.98%) among these models.

E The In-domain Evaluation Results

Following (Wang et al., 2019b), we report the in-domain evaluation results in Table 8. We generally find that ELECTRA-large achieves the best average performance over seven tasks. Note that we report the results by evaluating models on the validation set provided by GLUE.

Model	SST-2	MNLI	QNLI	RTE	MRPC	QQP	STS-B	CoLA	Avg	Avg $\Delta\downarrow$
Humans (OOD)	*92.36*	*84.13*	*81.10*	*83.47*	*84.70*	*85.43*	*80.28*	*58.98*	*80.14*	*7.82*
GPT-3 (ID)	93.68	69.27	79.20	80.20	79.21	72.15	88.10	50.13	76.49	-
GPT-3.5 (ID)	95.75	72.25	82.78	82.71	73.36	75.69	89.55	54.99	78.39	-
GPT-3 (OOD)	92.33	61.50	79.00	71.03	59.55	55.41	73.74	27.31	64.98	11.51
GPT-3.5 (OOD)	**95.92**	66.01	75.84	66.15	58.43	67.96	74.01	30.77	66.90	**11.49**
ELECTRA-large (OOD)	95.14	**76.94**	**80.44**	**78.74**	**69.96**	**77.24**	**81.14**	**37.85**	**74.62**	16.33

Table 7: OOD performance of GPT-3 and GPT3.5 using in-context learning compared with human performance and ELECTRA-large. We randomly select a single instance for each label. GPT-3 refers to text-davinci-003, and GPT-3.5 denotes the gpt-3.5-turbo.

F The Correlation between ID and OOD Performance

In general, we find that the overall performance of ID and OOD tests shows a linear correlation for both discriminative and generative models. In addition to the overall performance, we look at task-level performance at a more granular level in Figure 6. As shown in Figure 6, we find that the linear correlation does not exist for every task. For example, the linear correlation is extremely weak for MRPC and QQP, with relatively low OOD accuracy. While the linear correlation becomes significant on STSB and QNLI.

G The ID and OOD Performance of GPT-3 and GPT-3.5

Settings. The performance of GPT-3 and GPT-3.5 is shown in Table 7, where we report the classification results based on 1,000 instances for each task. The training strategy of GPT-3 is simulated to keep the same as human evaluation. We feed the model with some in-domain instances as instructions before testing on the OOD dataset. To achieve this,

we adopt the official API for calculating the in-domain performance of GPT-3 (text-DaVinci-003) based on 1,000 sampled ID instances. We leverage the in-context learning following (Ouyang et al., 2022) to calculate its OOD results on GLUE-X.

Results. In Table 7, it is interesting to see that the performance decay ratio of GPT-3 caused by the domain generalization is similar to GPT-3.5 while significantly larger than Humans (11.49% – GPT 3.5 vs. 11.51% – GPT-3 vs. 7.82% – Humans), indicating that there is much room for improvement in the OOD robustness, even for state-of-the-art LLMs. Meanwhile, it can be seen that the OOD performance of GPT-3.5 is still far behind the human performance (66.90% – GPT-3 vs. 80.14% – Humans), and slightly lower than ELECTRA-large (69.68%). Notably, the results of GPT-3/3.5 should be **treated with caution and just for reference** because we are not sure if datasets of GLUE-X are already included in the training corpus of GPT-3/3.5. Also, the OOD performance listed in Table 7 cannot be compared with PLMs fairly, as we only adopt instructions to evaluate it not fine-tuning the model like other PLMs in GLUE-X.

| Model	SST-2	MNLI	QNLI	RTE	MRPC	QQP	STSB	COLA	Average	Parameters
ELECTRA-large	97.25	89.29	93.65	88.45	92.60	89.84	88.06	74.33	89.18	334.09
RoBERTa-large	95.87	89.47	93.45	84.48	92.36	90.43	86.68	69.90	87.83	355.36
T5-large	95.41	88.83	94.34	89.89	92.01	90.59	87.58	62.97	87.70	737.67
BART-large	95.76	88.30	94.20	83.39	92.21	90.41	86.81	65.29	87.05	406.29
XLNet-large	96.44	89.50	93.32	84.12	91.54	90.06	86.36	62.63	86.75	360.27
T5-base	94.50	86.55	93.12	83.39	91.22	90.06	86.79	61.71	85.92	222.90
ELECTRA-base	91.51	87.12	92.09	80.14	91.09	89.36	86.07	69.95	85.92	108.89
RoBERTa-base	94.27	87.43	92.48	76.53	91.83	89.77	86.59	63.25	85.27	124.65
GPT2-large	94.50	85.48	91.21	75.45	87.78	89.34	84.75	60.06	83.57	774.03
BERT-large	93.46	85.69	91.84	70.76	90.26	89.78	83.97	60.32	83.26	335.14
BART-base	93.69	85.89	91.65	76.17	89.75	89.52	84.87	52.78	83.04	139.42
ALBERT-base	92.09	83.81	90.98	73.29	90.23	88.70	84.30	57.25	82.58	11.68
XLNet-base	94.15	86.49	91.36	68.59	90.50	89.39	83.94	53.67	82.26	116.72
BERT-base	92.89	83.63	91.05	66.79	89.41	89.40	83.71	59.75	82.08	109.48
GPT2-medium	94.27	85.38	90.81	70.04	87.20	89.42	83.75	53.87	81.84	354.82
ELECTRA-small	91.28	81.93	88.69	68.59	89.88	88.98	83.61	59.06	81.50	13.48
T5-small	91.97	82.82	90.77	70.40	89.13	89.07	84.74	43.88	80.35	60.51
DistilBERT-base	91.17	82.20	89.27	65.34	88.33	88.63	82.28	54.43	80.21	66.36
GPT2	90.94	82.63	88.78	69.31	84.51	88.63	82.31	47.29	79.30	124.44

Table 8: Detailed results of the in-domain test on each task sorted by the average performance.

| Model	SST2	MNLI	QNLI	RTE	MRPC	QQP	STSB	COLA
ELECTRA-large	2e-05/64	2e-05/16	5e-05/64	5e-05/32	2e-05/16	2e-05/64	2e-05/64	2e-05/128
RoBERTa-large	2e-05/32	2e-05/64	2e-05/32	2e-05/32	2e-05/16	2e-05/64	2e-05/16	2e-05/32
T5-large	1e-4/16	1e-4/32	1e-4/32	1e-4/32	1e-4/32	1e-4/16	1e-4/64	1e-4/32
BART-large	2e-05/32	2e-05/16	2e-05/32	3e-05/32	2e-05/32	2e-05/32	2e-05/30	2e-05/32
XLNet-large	3e-05/64	2e-05/64	3e-05/32	1e-05/32	1e-05/16	2e-05/32	2e-05/16	2e-05/16
T5-base	1e-4/32	1e-4/16	1e-4/32	1e-4/8	1e-4/16	1e-4/16	3e-4/16	1e-4/32
ELECTRA-base	1e-4/32	5e-05/64	5e-05/64	5e-05/16	5e-05/16	5e-05/32	5e-05/16	5e-05/32
RoBERTa-base	2e-05/32	2e-05/32	2e-05/32	2e-05/32	3e-05/32	2e-05/32	3e-05/32	2e-05/16
GPT2-large	2e-05/32	2e-05/32	2e-05/32	2e-05/16	3e-05/32	2e-05/32	2e-05/32	2e-05/32
BERT-large	2e-05/32	2e-05/32	2e-05/16	2e-05/16	2e-05/16	2e-05/64	2e-05/64	3e-05/16
BART-base	2e-05/32	2e-05/32	2e-05/32	2e-05/16	3e-05/32	2e-05/16	2e-05/16	2e-05/32
ALBERT-base	2e-05/32	2e-05/32	2e-05/32	2e-05/32	2e-05/16	2e-05/32	2e-05/32	2e-05/32
XLNet-base	3e-05/32	2e-05/32	2e-05/32	1e-05/16	2e-05/16	2e-05/32	2e-05/32	1e-05/32
BERT-base	2e-05/32	3e-05/32	2e-05/32	3e-05/32	3e-05/32	2e-05/32	2e-05/16	3e-05/32
GPT2-medium	2e-05/32	2e-05/16	3e-05/32	3e-05/32	3e-05/32	2e-05/16	3e-05/32	3e-05/32
ELECTRA-small	5e-05/64	5e-05/64	5e-05/32	5e-05/64	5e-05/32	5e-05/32	5e-05/32	5e-05/64
T5-small	1e-4/16	1e-4/16	1e-4/32	3e-4/16	1e-4/16	1e-4/16	3e-4/32	3e-4/32
DistilBERT-base	3e-05/16	2e-05/32	3e-05/32	2e-05/16	2e-05/16	2e-05/16	2e-05/16	2e-05/16
GPT2	2e-05/32	2e-05/32	3e-05/32	2e-05/32	2e-05/32	2e-05/32	3e-05/32	3e-05/32

Table 9: The hyper-parameter setting for each task, including the learning rate and batch size.
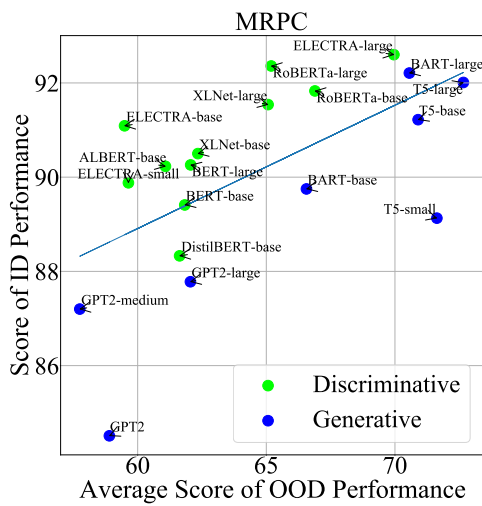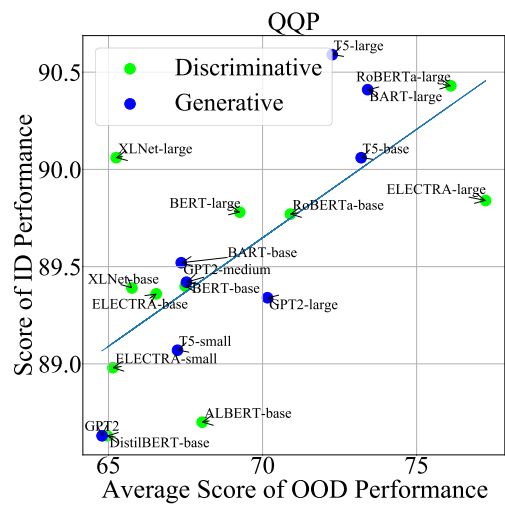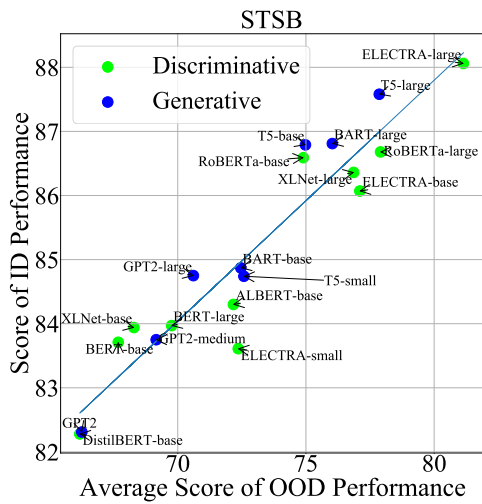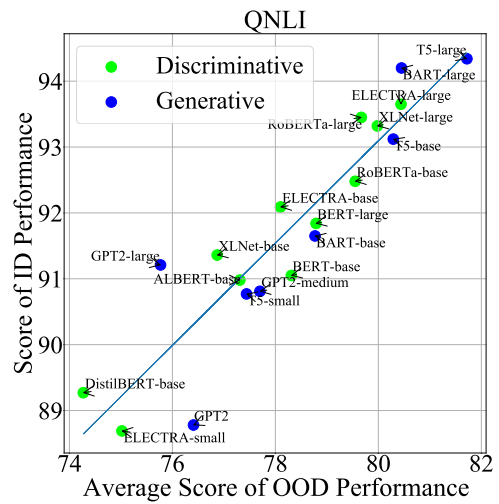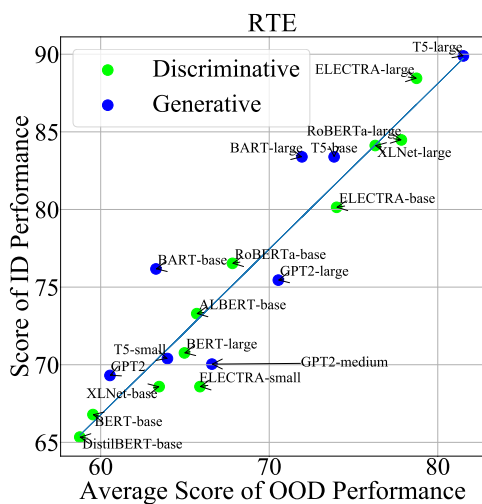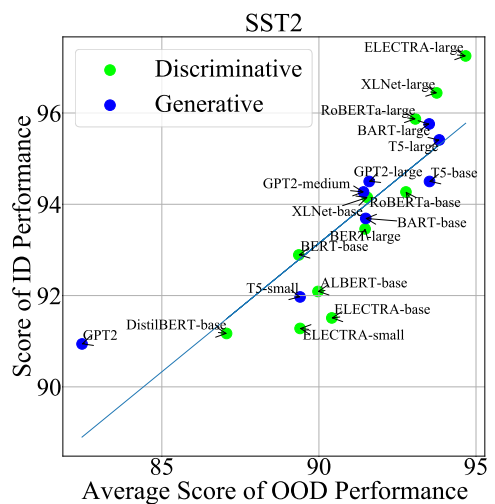
Figure 6: The correlation between the ID and OOD performance for each task involving in GLUE-X.

A For every submission:

☑ A1. Did you describe the limitations of your work?
Six.

☑ A2. Did you discuss any potential risks of your work?
Seven.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
One.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Left blank.

☑ B1. Did you cite the creators of artifacts you used?
Section 3.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 3.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 3.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 3.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3.

C ☑ Did you run computational experiments?

Section 4.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3 (Table 2).

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3.

D ☑ **Did you use human annotators (e.g., crowdworkers) or research with human participants?**
Section 4.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 4.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 4.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 4.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.