A Appendix

A.1 Glossary of Adversarial Attacks

We present a glossary of adversarial attacks considered in AdvGLUE in Table 6 and 7.

A.2 Additional Related Work

We discuss more related work about textual adversarial attacks and defenses in this subsection.

Textual Adversarial Attacks Recent research has shown deep neural networks (DNNs) are vulnerable to adversarial examples that are carefully crafted to fool machine learning models without disturbing human perception [16, 38, 33]. However, compared with a large amount of adversarial attacks in continuous data domain [53, 5, 11], there are a few studies focusing on the discrete text domain. Most existing gradient-based attacks on image or audio models are no longer applicable to NLP models, as words are intrinsically discrete tokens. Another challenge for generating adversarial text is to ensure the semantic and syntactic coherence and consistency.

Existing textual adversarial attacks can be roughly divided into three categories: word-level transformations, sentence-level attacks, and human-crafted samples. (i) Word-level transformations adopt different word replacement strategies during attack. For example, existing work [27, 10] applies character-level perturbation to carefully crafted typo words (*e.g.*, from "foolish" to "fo0lish"), thus making the model ignore or misunderstand the original statistical cues. Others adopt knowledge-based perturbation and utialize knowledge base to constrain the search space. For example, Zang et al. [56] uses sememe-based knowledge base from HowNet [40] to construct a search space for word substitution. Some [24, 27] use non-contextualized word embedding from GLoVe [39] or Word2Vec [32] to build synonym candidates, by querying the cosine similarity or euclidean distance between the original and candidate word and selecting the closet ones as the replacements. Recent work [13, 29] also leverages BERT to generate contextualized perturbations by masked language modeling. (ii) Different from the dominant word-level adversarial attacks, sentence-level adversarial attacks perform sentence-level transformation or paraphrasing by perturbing the syntactic structures based on human crafted rules [35, 42] or carefully designed auto-encoders [20, 49]. Sentence-level manipulations are generally more challenging than word-level attacks, because the perturbation space for syntactic structures are limited compared to word-level perturbation spaces that grow exponentially with the sentence length. However, sentence-level attacks tend to have higher linguistic quality than word-level, as both semantic and syntactic coherence are taken into considerations when generating adversarial sentences. (iii) Human-crafted adversarial examples are generally crafted in the human-in-the-loop manner [21, 37, 1] or use manually crafted templates to generate test cases [35, 42]. Our AdvGLUE incorporates all of the above textual adversarial to provide a comprehensive and systematic diagnostic report over existing state-of-the-art large-scale language models.

Defenses against Textual Adversarial Attacks To defend against textual adversarial attacks, existing work can be classified into three categories: (i) *Adversarial Training* is a practical method to defend against adversarial examples. Existing work either uses PGD-based attacks to generate adversarial examples in the embedding space of NLP as data augmentation [60], or regularizes the standard objective using virtual adversarial training [23, 30, 12]. However, one drawback is that the threat model is often unknown, which renders adversarial training less effective when facing unseen attacks. (ii) *Interval Bound Propagation* (IBP) [9] is proposed as a new technique to consider the worst-case perturbation theoretically. Recent work [19, 22] has applied IBP in the NLP domain to certify the robustness of models. However, IBP-based methods rely on strong assumptions of model architecture and are difficult to adapt to recent transformer-based language models. (iii) *Randomized Smoothing* [7] provides a tight robustness guarantee in ℓ_2 norm by smoothing the classifier with Gaussian noise. Ye et al. [55] adapts the idea to the NLP domain, and replace the Gaussian noise with synonym words to certify the robustness as long as adversarial word substitution falls into predefined synonym sets. However, to guarantee the completeness of the synonym set is challenging.

A.3 Task Descriptions, Statistics and Evaluation Metrics

We present the detailed label distribution statistics and evaluation metrics of GLUE and AdvGLUE benchmark in 8.

Table 6: **Glossary of adversarial attacks (word-level and sentence-level) in AdvGLUE.** For each adversarial attack, we provide a brief explanation and a corresponding example in AdvGLUE.

Perturbations	**Explanation**	**Examples** (~~Strikethrough~~ = Original Text, red = Adversarial Perturbation)
TextBugger (Word-level / Typo-based)	TextBugger first identifies the important words in each sentence and then replaces them with carefully crafted typos.	**Task:** QNLI **Question**: What was the population of the Dutch Republic before this emigration? **Sentence**: This was a ~~huge~~ hu ge influx as the entire population of the Dutch Republic amounted to ca. **Prediction:** False → True
TextFooler (Word-level / Embedding-similarity-based)	Embedding-similarity-based adversarial attacks such as TextFooler select synonyms according to the cosine similarity of word embeddings. Words that have high similarity scores will be used as candidates to replace original words in the sentences.	**Task:** QQP **Question 1**: I am getting fat on my lower body and on the ~~chest~~ torso, is there any way I can get fit without looking skinny fat? **Question 2**: Why I am getting skinny instead of losing body fat? **Prediction:** Not Equivalent → Equivalent
BERT-ATTACK (Word-level / Context-aware)	BERT-ATTACK uses pre-trained BERT to perform masked language prediction to generate contextualized potential word replacements for those crucial words.	**Task:** MNLI **Premise**: Do you know what this is? With a dramatic gesture she flung back the left side of her ~~coat~~ sleeve and exposed a small enamelled badge. **Hypothesis**: The coat that she wore was long enough to cover her knees . **Prediction:** Neutral → Contradiction
SememePSO (Word-level / Knowledge-guided)	Knowledge-guided adversarial attacks such as SememePSO use external knowledge base such as HowNet or WordNet to search for substitutions.	**Task:** QQP **Question 1**: What people who you've never met have ~~influenced~~ infected your life the most? **Question 2**: Who are people you have never met who have had the greatest influence on your life? **Prediction:** Equivalent → Not Equivalent
CompAttack (Word-level / Compositions)	CompAttack is a whitebox-based adversarial attack that integrates all other word-level perturbation methods in one algorithm to evaluate model robustness to various adversarial transformations.	**Task:** SST-2 **Sentence**: The primitive force of this film seems to ~~bubble~~ bybble up from the vast collective memory of the combatants. **Prediction:** Positive → Negative
SCPN (Sent.-level / Syntactic-based)	SCPN is an attack method based on syntax tree transformations. It is trained to produce a paraphrase of a given sentence with specified syntactic structures.	**Task:** RTE **Sentence 1**: He became a boxing referee in 1964 and became most well-known for his decision against Mike Tyson, during the Holyfield fight, when Tyson bit Holyfield's ear. **Sentence 2**: Mike Tyson bit ~~Holyfield's ear~~ in 1964. **Prediction:** Not Entailment → Entailment
T3 (Sent.-level / Syntactic-based)	T3 is a whitebox attack algorithm that can add perturbations on different levels of the syntax tree and generate the adversarial sentence.	**Task:** MNLI **Premise**: What's truly striking, though, is that Jobs ~~has~~ had never really let this idea go. **Hypothesis**: Jobs never held onto an idea for long. **Prediction:** Contradiction → Entailment
AdvFever (Sent.-level / Syntactic-based)	Entailment preserving rules proposed by AdvFever transform all the sentences satisfying the templates into semantically equivalent ones.	**Task:** SST-2 **Sentence**: ~~I'll bet the video game is~~ There exists a lot more fun than the film that goes by the name of i 'll bet the video game. **Prediction:** Negative → Positive
StressTest (Sent.-level / Distraction-based)	StressTest appends three true statements ("and true is true", "and false is not true", "and true is true" for five times) to the end of the hypothesis sentence for NLI tasks.	**Task:** RTE **Sentence 1**: Yet, we now are discovering that antibiotics are losing their effectiveness against illness. Disease-causing bacteria are mutating faster than we can come up with new antibiotics to fight the new variations. **Sentence 2**: Bacteria is winning the war against antibiotics and true is true. **Prediction:** Entailment → Not Entailment
CheckList (Sent.-level / Distraction-based)	CheckList adds randomly generated URLs and handles to distract model attention.	**Task:** QNLI **Question**: What was the population of the Dutch Republic before this emigration? https://t.co/DlI9kw **Sentence**: This was a huge influx as the entire population of the Dutch Republic amounted to ca. **Prediction:** False → True

Table 7: **Glossary of adversarial attacks (human-crafted) in AdvGLUE.** For each adversarial attack, we provide a brief explanation and a corresponding example in AdvGLUE.

Perturbations	**Explanation**	**Examples** (~~Strikethrough~~ = Original Text, red = Adversarial Perturbation)
CheckList (Human-crafted)	CheckList analyses different capabilities of NLP models using different test types. We adopt two capability tests: *Temporal* and *Negation*, which test if the model understands the order of events and if the model is sensitive to negations.	**Task:** SST-2 **Sentence**: I think this movie is perfect, but I used to think it was annoying. **Prediction:** Positive → Negative
StressTest (Human-crafted)	StressTest proposes carefully crafted rules to construct "stress tests" and evaluate robustness of NLI models to specific linguistic phenomena. Here we adopt the test cases focusing on *Numerical Reasoning*.	**Task:** MNLI **Premise**: If Anne' s speed were doubled, they could clean their house in 3 hours working at their respective rates. **Hypothesis**: If Anne' s speed were doubled, they could clean their house in less than 6 hours working at their respective rates. **Prediction:** Entailment → Contradiction
ANLI (Human-crafted)	ANLI is a large-scale NLI dataset collected iteratively in a human-in-the-loop manner. The sentence pairs generated in each round form a comprehensive dataset that aims at examining the vulnerability of NLI models.	**Task:** MNLI **Premise**: Kamila Filipcikova (born 1991) is a female Slovakian fashion model. She has modeled in fashion shows for designers such as Marc Jacobs, Chanel, Givenchy, Dolce & Gabbana, and Sonia Rykiel. And appeared on the cover of Vogue Italia two times in a row. **Hypothesis**: Filipcikova lives in Italy. **Prediction:** Neutral → Contradiction
AdvSQuAD (Human-crafted)	AdvSQuAD is an adversarial dataset targeting at reading comprehension systems. Examples are generated by appending a distracting sentence to the end of the input paragraph. We adopt the distracting sentences and questions in the *QNLI* format with labels "not answered".	**Task:** QNLI **Question**: What day was the Super Bowl played on? **Sentence**: The Champ Bowl was played on August 18th,1991. **Prediction:** False → True

SST-2 The Stanford Sentiment Treebank [43] consists of sentences from movie reviews and human annotations of their sentiment. Given a review sentence, the task is to predict the sentiment of it. Sentiments can be divided into two classes: positive and negative.

QQP The Quora Question Pairs (QQP) dataset is a collection of question pairs from the community question-answering website Quora. The task is to determine whether a pair of questions are semantically equivalent.

MNLI The Multi-Genre Natural Language Inference Corpus [52] consists of sentence pairs with textual entailment annotations. Given a premise sentence and a hypothesis sentence, the task is to predict whether the premise entails the hypothesis (entailment), contradicts the hypothesis (contradiction), or neither (neutral)

QNLI Question-answering NLI (QNLI) dataset consists of question-sentence pairs modified from The Stanford Question Answering Dataset [41]. The task is to determine whether the context sentence contains the answer to the question.

RTE The Recognizing Textual Entailment (RTE) dataset is a combination of a series of data from annual textual entailment challenges. Examples are constructed based on news and Wikipedia text. The task is to predict the relationship between a pair of sentences. For consistency, the relationship can be classified into two classes: entailment and not entailment, where neutral and contradiction are seen as not entailment.

We also show the detailed per-task model performance on AdvGLUE and GLUE in Table 9.

A.4 Implementation Details of Adversarial Attacks

TextBugger To ensure the small magnitude of the perturbation, we consider the following five strategies: (i) randomly inserting a space into a word; (ii) randomly deleting a character of a word; (iii) randomly replacing a character of a word with its adjacent character in the keyboard; (iv) randomly replacing a character of a word with its visually similar counterpart (*e.g.*, "0" v.s. "o", "1"

Table 8: The label distribution of AdvGLUE dataset. For SST-2, we report the label distribution as *"negative"*:*"positive"*. For QQP, we report the label distribution as *"not equivalent"*:*"equivalent"*. For QNLI, we report the label distribution as *"true"*:*"false"*. For RTE, we report the label distribution as *"entailment"*:*"not entailment"*. For MNLI, we report the label distribution as *"entailment"*:*"neutral"*:*"contradiction"*.

Corpus	Task	|Dev| (GLUE)	|Test| (GLUE)	|Dev| (AdvGLUE)	|Test| (AdvGLUE)	Evaluation Metrics
SST-2	sentiment	428:444	1821	72:76	590:830	acc.
QQP	paraphrase	25,545:14,885	390,965	46:32	297:125	acc./F1
QNLI	NLI/QA	2,702:2,761	5,463	74:74	394:574	acc.
RTE	NLI	146:131	3,000	35:46	123:181	acc.
MNLI	NLI	6,942:6,252:6,453	19,643	92:84:107	706:565:593	matched acc./mismatched acc.

Table 9: Model performance on AdvGLUE test set and GLUE dev set.

Models	Avg		SST-2		MNLI		RTE		QNLI		QQP	
	GLUE	AdvGLUE	GLUE	AdvGLUE	GLUE	AdvGLUE	GLUE	AdvGLUE	GLUE	AdvGLUE	GLUE	AdvGLUE
BERT(Large)	85.76	33.68	93.23	33.03	85.78/85.57	28.72/27.05	68.95	40.46	91.91	39.77	90.72/87.38	37.91/16.56
RoBERTa(Large)	91.44	50.21	95.99	58.52	89.74/89.86	50.78/39.62	86.60	45.39	94.14	52.48	91.99/89.37	57.11/41.80
T5(Large)	90.39	56.82	95.53	60.56	88.98/89.20	48.43/38.98	84.12	62.83	93.78	57.64	90.82/88.07	63.03/55.68
ALBERT(XXLarge)	91.87	59.22	95.18	66.83	89.29/89.88	51.83/44.17	88.45	73.03	95.26	63.84	92.26/89.49	56.40/32.35
ELECTRA(Large)	93.16	41.69	97.13	58.59	90.71	14.62/20.22	90.25	23.03	95.17	57.54	92.56	61.37/42.40
DeBERTa(Large)	92.67	60.86	96.33	57.89	90.95/90.85	58.36/52.46	90.25	78.94	94.86	57.85	92.29/89.69	60.43/47.98
SMART(BERT)	85.70	30.29	93.35	25.21	84.72/85.34	26.89/23.32	69.68	38.16	91.71	34.61	90.25/87.22	36.49/20.24
SMART(RoBERTa)	92.62	53.71	96.56	50.92	90.75/90.66	45.56/36.07	90.98	70.39	95.04	52.17	91.20/88.44	64.22/44.28
FreeLB(RoBERTa)	92.28	50.47	96.44	61.69	90.64	31.59/27.60	86.69	62.17	95.04	62.29	92.58	42.18/31.07
InfoBERT(RoBERTa)	89.06	46.04	96.22	47.61	89.67/89.27	50.39/41.26	74.01	39.47	94.62	54.86	92.25/89.70	49.29/35.54

v.s. "l"); and (v) randomly swapping two characters in a word. The first four strategies guarantee the word edit distance between the typo word and its original word to be 1, and that of the last strategy is limited to 2. Following the default setting, in Strategy (i), we only insert a space into a word when the word contains less than 6 characters. In Strategy (v), we swap characters in a word only when the word has more than 4 characters.

TextFooler Concretely, for the sentiment analysis tasks, we set the cosine similarity threshold to be 0.8, which encourages the synonyms to be semantically close to original ones and enhances the quality of adversarial data. For the rest of the tasks, we follow the default hyper-parameter to set the cosine similarity threshold to be 0.7. Besides, the number of synonyms for each word is set to 50 following the default setting.

BERT-ATTACK We follow the hyper-parameters from the official codebase, and set the number of candidate words to 48 and cosine similarity threshold to 0.4 in order to filter out antonyms using synonym dictionaries, as BERT masked language model does not distinguish synonyms and antonyms.

SememePSO We adopt the official hyper-parameters in which maximum and minimum inertia weights are set to 0.8 and 0.2, respectively. We also set the maximum and minimum movement probabilities of the particles to 0.8 and 0.2, respectively, following the default setting. Population size is set to 60 in every task.

CompAttack We follow the T3 [49] and C&W attack [5] and design the same optimization objective for adversarial perturbation generation in the embedding space as:

$$\mathcal{L}(\boldsymbol{e}^*) = ||\boldsymbol{e}^*||_p + c \cdot g(\boldsymbol{x}'), \tag{3}$$

where the first term controls the magnitude of perturbation, while $g(\cdot)$ is the attack objective function depending on the attack scenario. c weighs the attack goal against attack cost. CompAttack constrains the perturbation to be close to pre-defined perturbation space, including typo space (*e.g.*, TextBugger), knowledge space (*e.g.,* WordNet) and contextualized embedding space (*e.g.,* BERT embedding clusters) to make sure the perturbation is valid. We can also see from Table 3 that CompAttack overall has lower filter rate than other state-of-the-art attack methods.

SCPN We use the pre-trained SCPN models released by the official codebase. Following the default setting, we select the most frequent 10 templates from ParaNMT-50M corpus [51] to guide the generation process. We first parse sentences from GLUE dev set using Stanford CoreNLP. We used CoreNLP version 3.7.0 in our experiment, along with the Shift-Reduce Parser models.

Table 10: **Examples of AdvGLUE benchmark**.

Task	Linguistic Phenomenon	Samples (~~Strikethrough~~ = Original Text, red = Adversarial Perturbation)	Label → Prediction
SST-2	Typo (Word-level)	**Sentence**: The primitive force of this film seems to ~~bubble~~ bybble up from the vast collective memory of the combatants.	Positive → Negative
SST-2	Context-aware (Word-level)	**Sentence**: In execution , this clever idea is far ~~less~~ smaller funny than the original , killers from space.	Negative → Positive
SST-2	CheckList (Human-crafted)	**Sentence**: I think this movie is perfect, but I used to think it was annoying.	Positive → Negative
QQP	Embedding (Word-level)	**Question 1**: I am getting fat on my lower body and on the ~~chest~~ torso, is there any way I can get fit without looking skinny fat? **Question 2**: Why I am getting skinny instead of losing body fat?	Not Equivalent → Equivalent
QQP	Syntactic (Sent.-level)	**Question 1**: ~~Can I learn MMA at the age of 26?~~ You can learn MMA at 24? **Question 2**: Can I learn MMA at the age of 24?	Not Equivalent → Equivalent
QQP	CheckList (Human-crafted)	**Question 1**: Is Alfred Kennedy an analyst? **Question 2**: Is Alfred Kennedy becoming an analyst?	Not Equivalent → Equivalent
MNLI	Typo (Word-level)	**Premise**: uh-huh how about any ~~matching~~ mathcing programs **Hypothesis**: What about matching programs?	Entailment → Contradiction
MNLI	Distraction (Sent.-level)	**Premise**: You and your friends are not welcome here, said Severn. **Hypothesis**: Severn said the people were not welcome there and true is true.	Entailment → Contradiction
MNLI	ANLI (Human-crafted)	**Premise**: Kamila Filipcikova (born 1991) is a female Slovakian fashion model. She has modeled in fashion shows for designers such as Marc Jacobs, Chanel, Givenchy, Dolce & Gabbana, and Sonia Rykiel. And appeared on the cover of Vogue Italia two times in a row. **Hypothesis**: Filipcikova lives in Italy.	Neutral → Contradiction
QNLI	Distraction (Sent.-level)	**Question**: What was the population of the Dutch Republic before this emigration? https://t.co/DlI9kw **Sentence**: This was a huge influx as the entire population of the Dutch Republic amounted to ca.	False → True
QNLI	AdvSQuAD (Human-crafted)	**Question**: What day was the Super Bowl played on? **Sentence**: The Champ Bowl was played on August 18th,1991.	False → True
RTE	Knowledge (Word-level)	**Sentence 1**: In Nigeria, by far the most populous country in sub-Saharan Africa, over 2.7 million people ~~are~~ exist infected with HIV. **Sentence 2**: 2.7 percent of the people infected with HIV live in Africa.	Not Entailment → Entailment
RTE	Syntactic (Sent.-level)	**Sentence 1**: He became a boxing referee in 1964 and became most well-known for his decision against Mike Tyson, during the Holyfield fight, when Tyson bit Holyfield's ear. **Sentence 2**: Mike Tyson bit ~~Holyfield's ear~~ in 1964.	Not Entailment → Entailment

T3 We follow the hyper-parameters in the official setting where the scaling const is set to $1e4$ and the optimizing confidence is set to 0. In each iteration, we optimize the perturbation vector for at most 100 steps with learning rate 0.1.

AdvFever We follow the entailment preserving rules proposed by the official implementation. We adopt all 23 templates to transform original sentences into semantically equivalent ones. Many common sentence patterns in everyday life are included in these templates.

A.5 Examples of AdvGLUE benchmark

We show more comprehensive examples in Table 10. Examples are generated with different levels of perturbations and they all can successfully change the predictions of all surrogate models (BERT, RoBERTa and RoBERTa ensemble).

A.6 Fine-tuning Details of Large-Scale Language Models

For all the experiments, we are using a GPU cluster with 8 V100 GPUs and 256GB memory.

Table 11: The statistics of AdvGLUE in the human training phase.

Corpus	Pay Rate (per batch)	#/ Qualified Workers	Human Acc. (Avg.)	Human Acc. (vote)	Fleiss Kappa
SST-2	\$0.4	70	89.2	95.0	0.738
MNLI	\$1.0	33	80.4	85.0	0.615
RTE	\$1.0	66	85.8	92.0	0.602
QNLI	\$1.0	41	85.6	91.0	0.684
QQP	\$0.5	58	86.4	90.0	0.691

BERT (Large) For RTE, we train our model for 10 epochs and for other tasks we train our model for 4 epochs. Batch size for QNLI is set to 512, and for other tasks it is set to 256. Learning rates are all set to $2e-5$.

ELECTRA (Large) We follow the official hyper-parameter setting to set the learning rate to $5e-5$ and set batch size to 32. We train ELECTRA on RTE for 10 epochs and train for 2 epochs on other tasks. We set the weight decay rate to 0.01 for every task.

RoBERTa (Large) We train our RoBERTa for 10 epochs with learning rate $2e-5$ on each task. The batch size for QNLI is 32 and 64 for other tasks.

T5 (Large) We train our T5 for 10 epochs with learning rate $2e-5$ on each task. The batch size for QNLI is 32 and 64 for other tasks. We follow the templates in original paper to convert GLUE tasks into generation tasks.

ALBERT (XXLarge) We use the default hyper-parameters to train our ALBERT. For example, max training steps for SST-2, MNLI, QNLI, QQP, RTE, is 20935, 10000, 33112, 14000, 800 respectively. For MNLI and QQP, batch size is set to 32 and for other tasks batch size is set to 128.

DeBERTa (Large) We use the official hyper-parameters to train our DeBERTa. For example, learning rate is set to $1e-5$ across all tasks. For MNLI and QQP, batch size is set to 64 and for other tasks batch size is set to 32.

SMART For SMART(BERT) and SMART(RoBERTa), we use grid search to search for the best parameters and report the best performance among all trained models.

FreeLB (RoBERTa) For FreeLB, we test every parameter combination provided by the official codebase and select the best parameters for our training.

InfoBERT (RoBERTa) We set the batch size to 32 and learning rate to $2e-5$ for all tasks.

A.7 Human Evaluation Details

Human Training We present the pay rate and the number of qualified workers in Table 11. We also test our qualified workers on another non-overlapping 100 samples of the GLUE dev sets for each task. We can see that the human accuracy is comparable to [36], which means that most our selected annotators understand the GLUE tasks well.

Human Filtering The detailed filtering statistics of each stage is shown in Table 12. We can see that around $60-80\%$ of examples are filtered due to the low transferability and high word modification rate. Among the remaining samples, around $30-40\%$ examples are filtered due to the low human agreement rates (Human Consensus Filtering), and around $20-30\%$ are filtered due to the semantic changes which lead to the label changes (Utility Preserving Filtering).

Human Annotation Instructions We show examples of annotation instructions in the training phase and filtering phase on MNLI in Figure 2 and 3. More instructions can be found in `https://adversarialglue.github.io/instructions`. We also provide a

Table 12: Filter rates during data curation.

Tasks	Metrics	Word-level Attacks					Average
		SememePSO	TextFooler	TextBugger	CombAttack	BERT-ATTACK	
SST-2	Transferability	58.85	63.56	64.87	53.58	66.87	61.54
	Fidelity	14.65	11.06	22.40	19.93	12.03	16.01
	Human Consensus	10.53	10.56	2.27	9.92	7.09	8.07
	Utility Preserving	6.68	5.43	0.51	3.20	3.82	3.93
	Filter Rate	90.71	90.62	90.04	86.63	89.81	89.56
MNLI	Transferability	44.16	43.15	42.58	35.08	41.80	41.36
	Fidelity	36.57	45.94	37.71	38.14	38.60	39.39
	Human Consensus	10.37	6.38	5.51	11.15	9.78	8.64
	Utility Preserving	4.49	2.08	1.32	11.07	5.91	4.97
	Filter Rate	95.59	97.55	87.12	95.45	96.10	94.36
RTE	Transferability	55.32	67.38	41.96	54.20	60.94	55.96
	Fidelity	19.83	7.79	42.18	23.17	14.25	21.44
	Human Consensus	8.08	7.91	3.55	7.64	8.44	7.12
	Utility Preserving	8.69	6.13	0.60	5.70	8.54	5.93
	Filter Rate	91.93	89.21	88.29	90.72	92.16	90.46
QNLI	Transferability	63.36	70.67	59.24	55.47	69.15	63.58
	Fidelity	17.73	13.01	25.31	23.53	13.17	18.55
	Human Consensus	10.06	9.80	6.84	9.98	9.36	9.21
	Utility Preserving	3.48	2.41	1.50	4.94	4.10	3.29
	Filter Rate	94.63	95.89	92.89	93.92	95.78	94.62
QQP	Transferability	42.96	58.60	55.09	44.83	51.97	50.69
	Fidelity	45.61	29.35	26.46	30.99	37.77	34.04
	Human Consensus	4.38	4.69	5.19	10.08	3.94	5.66
	Utility Preserving	3.79	3.86	3.16	7.93	4.60	4.67
	Filter Rate	96.73	96.50	89.90	93.83	98.28	95.05

FAQ document in each task description page `https://docs.google.com/document/d/1MikHUdyvcsrPqE8x-N-gHaLUNAbA6-Uvy-iA5gkStoc/edit?usp=sharing`.

A.8 Discussion of Limitations

Due to the constraints of computational resources, we are unable to conduct a comprehensive evaluation of all existing language models. However, with the release of our leaderboard website, we are expecting researchers to actively submit their models and evaluate against our AdvGLUE benchmark to have a systematic understanding of model robustness. We are also interested in the adversarial robustness of large-scale auto-regressive language models under the few-shot settings, and leave it as a compelling future work.

In this paper, we follow ANLI [37] and generate adversarial examples against surrogate models based on BERT and RoBERTa. However, there are concerns [2] that such adversarial filtering may not be able to fairly benchmark the model robustness, as participants may top the leaderboard by producing different errors from our surrogate models. We note that such concerns can be solved given systematic data curation. As shown in our main benchmark results, we observe we successfully select the adversarial examples with high adversarial transferability that can unveil the vulnerabilities shared across models of different architectures. Specifically, we observe a huge performance gap in ELECTRA (Large) that is pre-trained with different data and shown less robust than one of surrogate model RoBERTa (Large).

Finally, we emphasize that our AdvGLUE benchmark mainly focuses on robustness evaluation. Thus AdvGLUE can also be considered as a supplementary diagnostic test set besides the standard GLUE benchmark. We suggest that participants should evaluate their models against both GLUE benchmark and our AdvGLUE to understand both model generalization and robustness. We hope our work can help researchers to develop models with high generalization and adversarial robustness.

A.9 Website

We present the diagnostic report on our website in Figure 4.

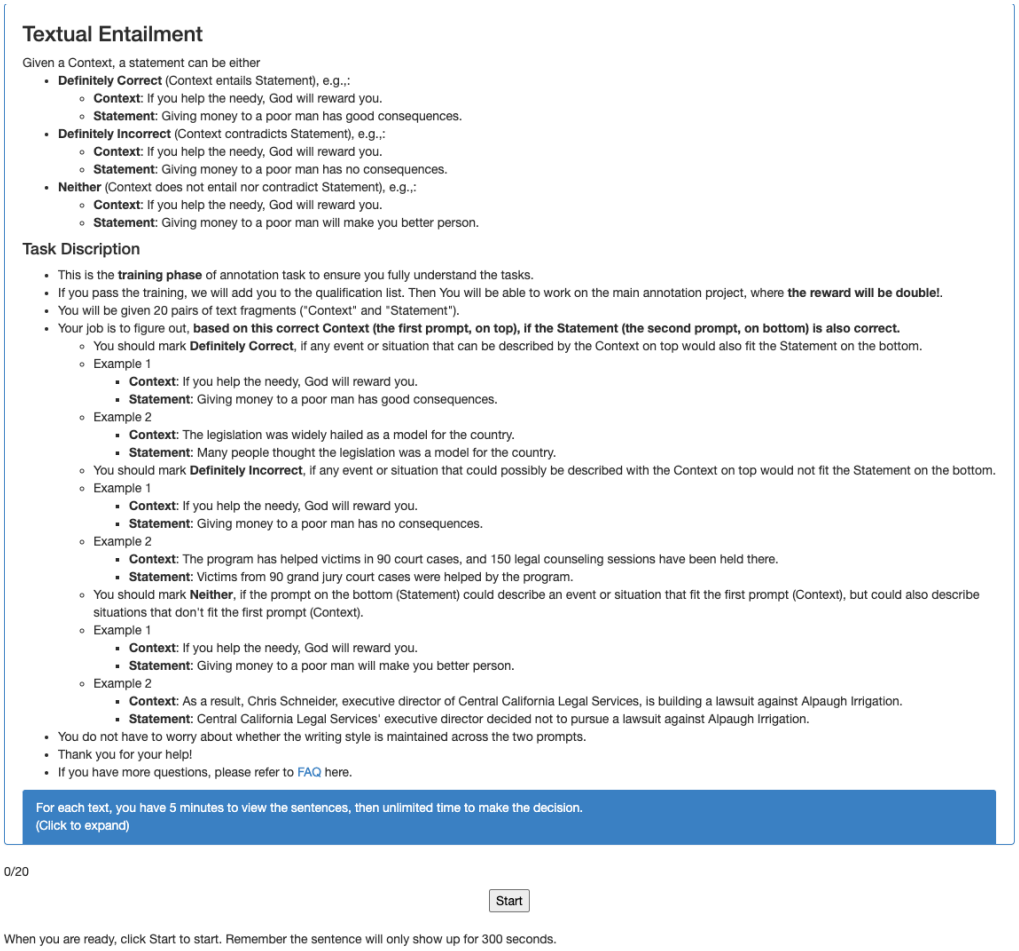

Figure 2: Human annotation instructions (training phase) for MNLI.

B Data Sheet

We follow the documentation frameworks provided by Gebru et al. [14].

B.1 Motivation

For what purpose was the dataset created? While recently a lot of methods (SMART, FreeLB, InfoBERT, ALUM) claim that they can improve the model robustness against adversarial attacks, the adversary setup in these methods (i) lacks a unified standard and is usually different across different methods; (ii) fails to cover comprehensive linguistic transformation (typos, synonymous substitution, paraphrasing, etc) to recognize to which levels of adversarial attacks models are still vulnerable. This motivates us to build a unified and principled robustness benchmark dataset and evaluate to which extent the state-of-the-art models have progressed so far in terms of adversarial robustness.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? University of Illinois at Urbana-Champaign (UIUC) and Microsoft Corporation.

B.2 Composition/collection process/preprocessing/cleaning/labeling and uses:

The answers are described in our paper as well as website `https://adversarialglue.github.io`.

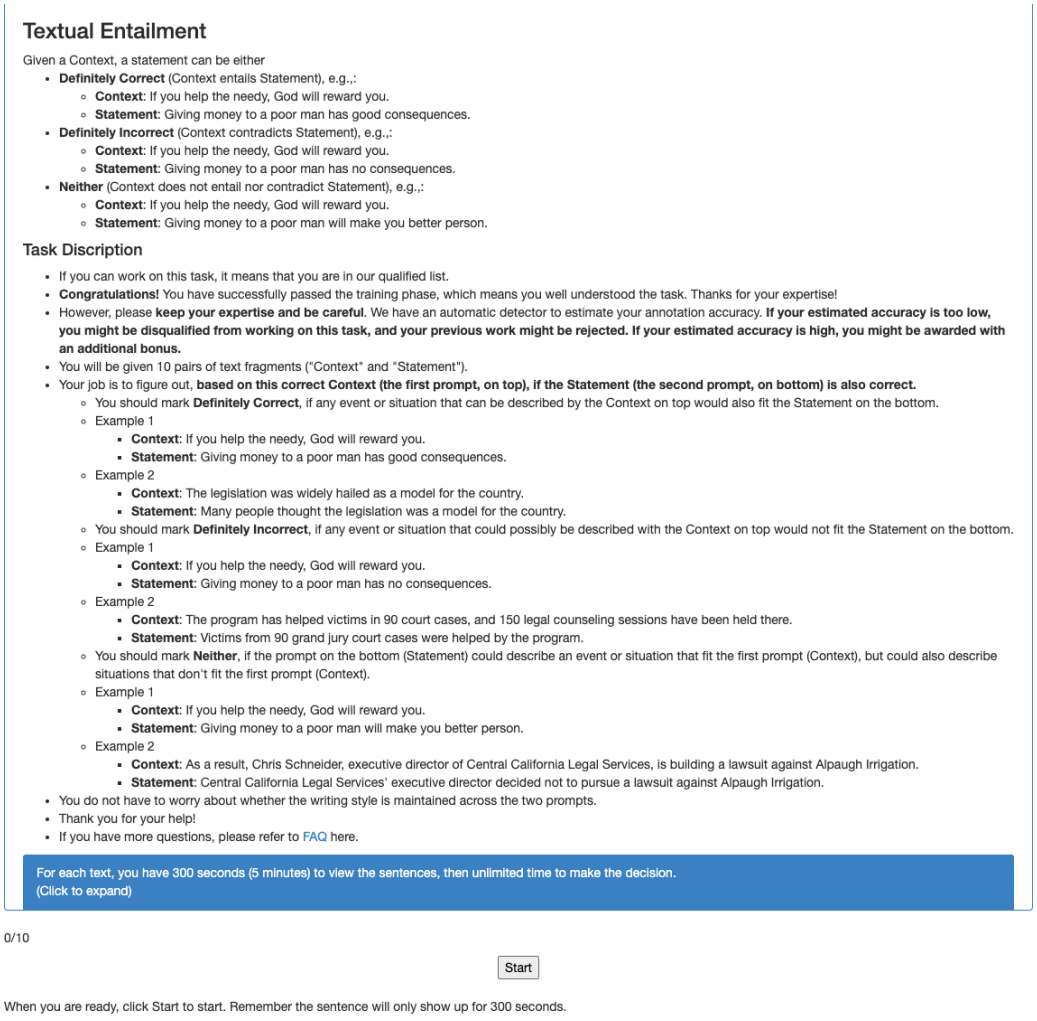

Figure 3: Human annotation instructions (filtering phase) for MNLI.

B.3 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? The dev set is released to the public. The test set is hidden and can only be evaluated by an automatic submission API hosted on CodaLab.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? The dev set is released on our website `https://adversarialglue.github.io`. The test set is hidden and hosted on CodaLab.

When will the dataset be distributed? It has been released now.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? Our dataset will be distributed under the CC BY-SA 4.0 license.

B.4 Maintenance

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? Boxin Wang (`boxinw2@illinois.edu`) and Chejian Xu (`xuchejian@zju.edu.cn`) will be responsible for maintenance.

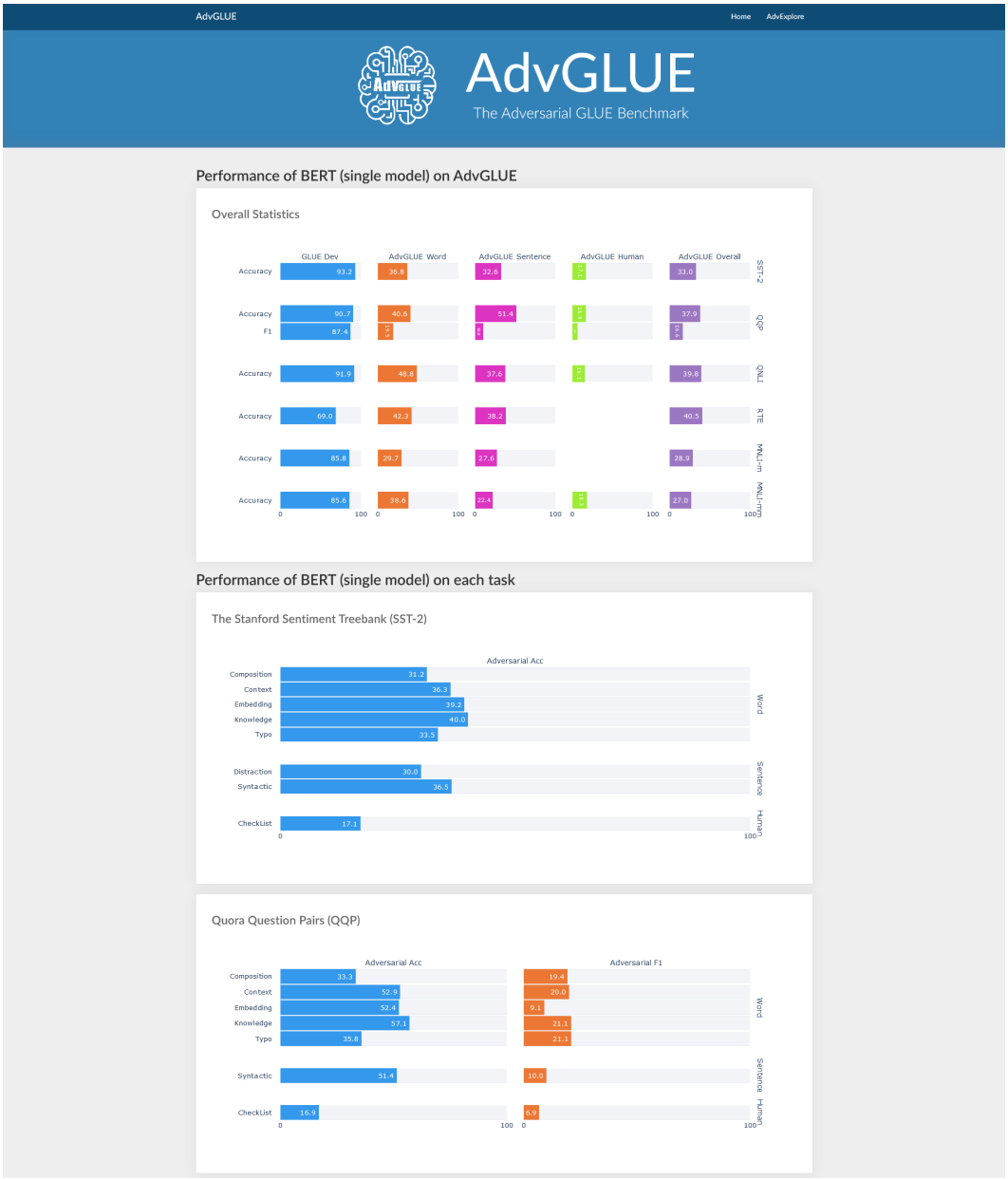

Figure 4: An example of model diagnostic report for BERT (Large).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? Yes. If we include more tasks or find any errors, we will correct the dataset and update the leaderboard accordingly. It will be updated on our website.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? They can contact us via email for the contribution.