Whispers of Doubt Amidst Echoes of Triumph in NLP Robustness

Ashim Gupta **Rishanth Rajendhran** **Nathan Stringham**
Vivek Srikumar **Ana Marasović**
Kahlert School of Computing
University of Utah
ashim@cs.utah.edu

Abstract

Are the longstanding robustness issues in NLP resolved by today's larger and more performant models? To address this question, we conduct a thorough investigation using 19 models of different sizes spanning different architectural choices and pretraining objectives. We conduct evaluations using (a) OOD and challenge test sets, (b) CheckLists, (c) contrast sets, and (d) adversarial inputs. Our analysis reveals that not all OOD tests provide further insight into robustness. Evaluating with CheckLists and contrast sets shows significant gaps in model performance; merely scaling models does not make them sufficiently robust. Finally, we point out that current approaches for adversarial evaluations of models are themselves problematic: they can be easily thwarted, and in their current forms, do not represent a sufficiently deep probe of model robustness. We conclude that not only is the question of robustness in NLP as yet unresolved, but even some of the approaches to measure robustness need to be reassessed.

1 Introduction

The versatility and consumer growth of commercial LLMs like ChatGPT give the impression that robustness evaluations such as out-of-domain and stress testing are no longer relevant. We argue that they remain important. For many applications, the broad range of skills offered by general-purpose models — from writing a clinical note to drawing unicorns in LaTeX (Bubeck et al., 2023) — is not needed. Specializing moderate-scale models with finetuning still works better when a model must perform a specific task.[1] Some NLP research embraces this view, but takes it to an extreme. Bowman (2022) bemoans the fact that the 2018 BERT-base model continues to be a baseline of choice, despite the feasibility of finetuning larger and better-pretrained models on a single GPU today. The hustle to innovate upon general-purpose models and the disregard of stronger baselines make it unclear where the field stands in terms of established robustness evaluations. We seek to addresses this gap.

We first point to popular experimental setups that, while of interest, are not adequate choices for certain robustness studies. Specifically, we check 177 ACL publications that include keywords in Figure 1 and document used train-test splits. We end up with 101 train-test splits after filtering. We use them to finetune and evaluate 19 models that differ in (i) transformer type (encoder-/decoder-only, encoder-decoder), (ii) model size (60M to 13B), and (iii) pretraining objectives (LM/MLM only, additional multitask pretraining). See Table 2 for a model overview. We also assess few-shot in-context learning on these splits with Mistral-7B.

We find that for 24% of 62 train/OOD-test combinations finetuned models show no OOD issues; Mistral-7B for 40%. We discourage performing OOD tests with these splits and urge to continue assessing the suitability of data splits as backbone models become better. A few challenge sets continue to break models finetuned on MNLI, QNLI, and SNLI. Yet, sentiment classifiers are robust to stress tests. If the fragility of NLI models stems from data issues (Bowman and Dahl, 2021), future stress tests should focus on models trained on better data (e.g., Liu et al., 2022; Kamoi et al., 2023).

We turn to other insights from this work that reflect on different themes in NLP robustness. Using the CheckList methodology (Ribeiro et al., 2020), we show that highly accurate models still struggle with the most basic task phenomena. Larger models help, do not fully resolve the issue; there is scant evidence that increasing size further would be beneficial. Next, we consider contrast set eval-

[1] For example, see this blogpost (retrieved Nov 15, 2023): Fine-Tuning Llama-2: A Comprehensive Case Study for Tailoring Models to Unique Applications.

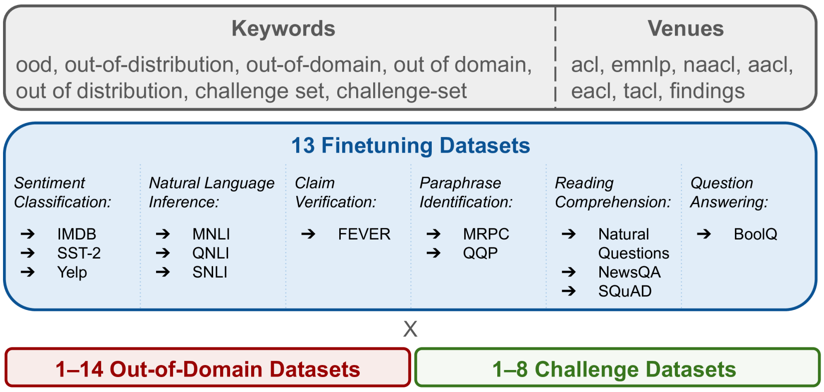

Keywords
ood, out-of-distribution, out-of-domain, out of domain, out of distribution, challenge set, challenge-set
Venues
acl, emnlp, naacl, aacl, eacl, tacl, findings
13 Finetuning Datasets
Sentiment Classification:
→ IMDB
→ SST-2
→ Yelp
Natural Language Inference:
→ MNLI
→ QNLI
→ SNLI
Claim Verification:
→ FEVER
Paraphrase Identification:
→ MRPC
→ QQP
Reading Comprehension:
→ Natural Questions
→ NewsQA
→ SQuAD
Question Answering:
→ BoolQ
X
1–14 Out-of-Domain Datasets
1–8 Challenge Datasets

Figure 1: Finetuning and evaluation datasets determined by analyzing train-test splits in *ACL/EMNLP publications from 2020–2022 (§2). Individual train-test splits are reported in Table 7. They represent the most common data setups for studying two popular aspects of NLP robustness.

	Model	Size	PT
Encoder-Only	RoBERTa-Base	124M	MLM
	RoBERTa-Large	355M	
	DeBERTa-v3-Base	184M	MLM
	DeBERTa-v3-Large	435M	
Decoder-Only	OPT-125M	125M	LM
	OPT-350M	331M	
	OPT-1.3B	1.3B	
	OPT-2.7B	2.7B	
	OPT-6.7B	6.7B	
	OPT-13B	12.8B	
	GPT-2	124M	LM
	GPT-2-Medium	354M	
	GPT-2-Large	774M	
	GPT-2-XL	1.6B	
Encoder-Decoder	T5-Small	60M	text-to-text MLM + MTL
	T5-Base	222M	
	T5-Large	737M	
	T5-XL (3B)	2.8B	
	T5-XXL (11B)	11.3B	

Figure 2: Finetuning models.

uations (Gardner et al., 2020), which test whether a model is consistently accurate in a set of subtly different yet differently labeled examples. We show that these evaluations continue to expose major model weaknesses, raising the question of why such informative evaluations are not more widely used. Finally, we stress a crucial finding — some established robustness evaluations can themselves be fragile. We demonstrate this by revealing that the success of adversarial attacks is exaggerated, and we define a more reliable metric for assessing the success of adversarial attacks. But the overarching lesson is broader than the new metric: be cautious about assuming that prevailing evaluation methods are well-designed.

As the landscape of LLMs evolves, new challenges arise such as preventing the generation of text that assists unlawful or harmful activities. Despite the importance of addressing these new difficulties, our research demonstrates that many long-standing robustness issues in NLP remain relevant and unresolved. If we cared about these challenges before, why should we stop now?

2 Where the Wild Robustness Setups Are

The goals of this section are to determine how models fare when their robustness is examined in two scenarios. First, when the data source differs between training and testing (*OOD evaluation*). Second, inputs are designed to challenge models beyond their normal operational capacity (*challenge set* or *stress* test; Naik et al., 2018). To this end, we examine experimental setups that have been seen in recent ACL publications.

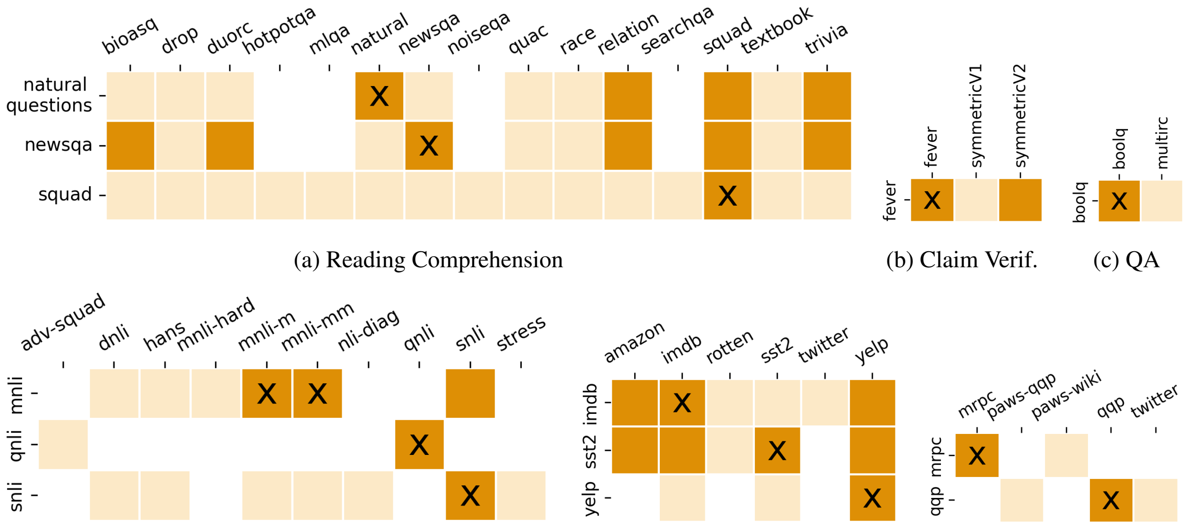
Common OOD/Challenge Data Splits. We manually collate train-test splits mentioned in 177 *ACL publications from the years 2020–2022, containing one of the keywords listed in Figure 1.[2] From these, we select 13 training datasets (spanning 6 task types) that appear in at least 4 papers reporting an OOD or challenge set evaluation (Figure 1, middle). After we filter splits where those training datasets occur, 101 splits remain; they are listed in Table 6 in the Appendix. Some splits are used both for OOD and challenge set testing.

Models. We use selected training datasets to separately finetune 19 models in Table 2.[3] We report the average performance across 3 runs with 3 seeds. In addition to finetuned models, we assess how robust few-shot in-context learning is in these settings using the base Mistral-7B model (Jiang et al., 2023). To the best of our knowledge (per the release information accompanying the model), its instruction data is not contaminated with standard NLP training sets via collections such as (Super-)NaturalInstructions (Mishra et al., 2022; Wang et al., 2022b) or T0 (Sanh et al., 2022). Thus, we expect that testing Mistral-7B on selected train-test splits aligns with the principle of OOD evaluation.

Q1: Are all commonly used OOD splits still suitable? No, 24% (15 out of 62) of the splits involving an OOD test set present a potentially in-

[2] We used Semantic Scholar (https://www.semanticscholar.org/) for this effort.

[3] Please refer to A in the appendix for more details.



(a) Reading Comprehension

(b) Claim Verif.

(c) QA

(d) Natural Language Inference

(e) Sentiment Classification

(f) Paraphrase Identification

Figure 3: Each colored square shows whether one of the top-5 models in-domain (when evaluated on the test set of the dataset in the row) has OOD performance that is at most 3% lower than in-domain. If yes, we use fuller color, and we deem that train-test split questionable for studying OOD. We mark in-domain splits with X.

adequate approach for examining the performance of finetuned models under distributional shifts, and 40% for few-shot in-context learning with Mistral-7B.[4] In Figure 3, train-test splits where some top-5 in-domain finetuned model (among 19) obtains performance that does not drop more than 3% OOD are marked with fuller colors. A comparable heatmap for Mistral-7B is given in Figure 8 (Appendix). If in-domain and OOD performance are close, then the model behaves as an ideal robust classifier (Taori et al., 2020). The existence of such models raises the question about the need to address alleged OOD issues. We focus on the top-5 models to exclude those with poor in-domain performance. In Figures 8g–8h in the Appendix, we visualize classification accuracy for train/OOD-tests that Mistral-7B and finetuned models handle similarly in-domain and OOD.

Q2: Which Data is Still "Stressful"? To answer this question, we consider models that are at least 85% accurate on their in-domain test sets: these highly accurate models are candidates for deployment, and require stress-testing. Table 1 shows the best accuracy on each challenge set (and their subparts) among these models, and also the difference between that model's in-domain and challenge set accuracy. The discussion below focuses on the NLI task — sentiment classifiers are robust and the model trained for QQP is not robust on PAWS-

Test (Category)	**Accuracy**	**Diff**
C-IMDB	94.5	2.1
IMDB Contrast (all)	99.6	-4.3
IMDB Contrast (contrast)	95.7	0.3
IMDB Contrast (original)	99.6	-4.3
ANLI (r1)	66.0	26.1
ANLI (r2)	53.5	38.6
ANLI (r3)	49.6	42.5
Breaking NLI	97.9	-6.7
HANS (all)	98.9	-6.8
HANS (constituent)	71.2	21.1
HANS (lexical overlap)	98.9	-6.8
HANS (subsequence)	70.3	21.9
MNLI-hard (val matched)	88.4	3.0
MNLI-hard (val mismatched)	88.2	3.3
NLI Diagnostics (min–max)	30.0–95.0	61.5 / -3.5
Stress Test (min–max)	76.8–90.6	14.4 / 0.9
SNLI CAD	82.9	9.2
SNLI-hard	85.5	6.6
PAWS-QQP	50.9	42.2

Table 1: The max. challenge set accuracy among models with 85+% in-domain performance. The last column shows the difference between accuracy in-domain performance and in the challenge set (higher means poorer generalization). Shaded rows highlight datasets that remain challenging for associated models.

[4] The number of total splits for Mistral-7B is 52 not 62 because we encountered issues with NewsQA.

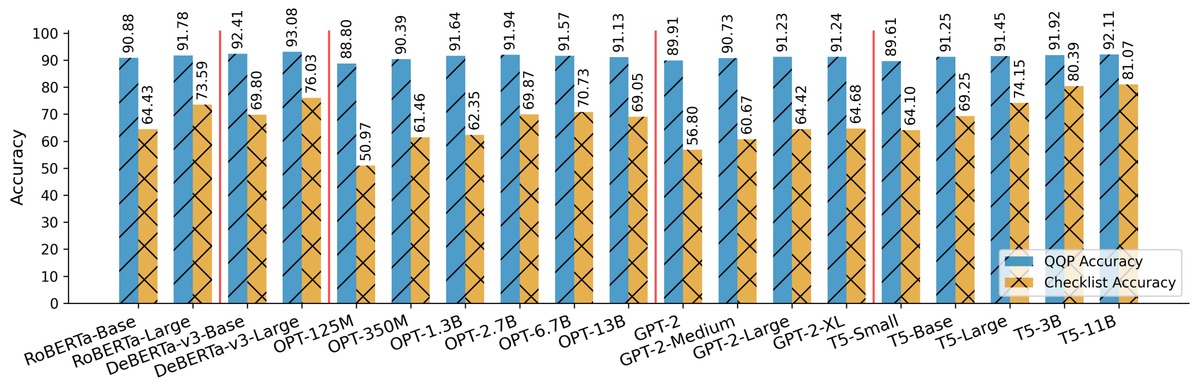


Figure 4: Accuracy on the QQP (duplicate questions identification) standard test set vs. CheckList accuracy.

	RoBERTa-L	DeBERTa-L	OPT-6.7B	GPT-2-XL	T5-11B
% Total Tests with 95+% accuracy	45.3	45.3	34.0	26.4	45.3
↳ **& No 3+% drops from scaling**	45.3	45.3	24.5	26.4	32.1
↳ **& Equally robust across model sizes**	32.1	34.0	13.2	22.6	26.4
% 10+% gains from scaling, no 3+% drops	28.3	20.8	9.4	7.5	15.1
% Total Tests with <60% accuracy	28.3	22.6	32.1	35.8	15.1
% Major scaling complications (10+% drops)	1.9	1.9	37.7	18.9	22.6

Table 2: Analysis of CheckList results for identifying duplicate questions (QQP) with best models from each type.

QQP.

We find that for models finetuned on MNLI, QNLI, or SNLI, the following datasets are still challenging: ANLI, HANS, SNLI CAD, SNLI-hard, and some of the tests in the NLI Diagnostic and Stress Test collections. None of the models reach 85% accuracy on these challenge sets. Although one model achieved 85.5% accuracy on SNLI-hard, we still include it due to the notable disparity between its in-domain and challenge set accuracies. We provide a breakdown of NLI Diagnostics and Stress Test results in Table 5 (Appendix). We also observe that when a challenge set ceases to be "stressful", various model types across different sizes demonstrate robustness. This is illustrated with "Breaking NLI" in Figure 9 (Appendix).

Mistral-7B has shown poor in-domain accuracy, which explains its low performance on specific stress tests like Breaking NLI, HANS, MNLI-hard, and SNLI-hard; see Table 3 (Appendix). Stress test failures of Mistral-7B happen with ANLI, PAWS-QQP, and SNLI-CAD.

3 Highly Accurate Models Still Stumble On The Basics

It has been shown that models, even when specialized for specific tasks and showing low error rates on standard datasets, might still fall short on the most basic task-related skills. We examine whether this has changed with the availability of models based on different transform types, pretraining objectives, and scaled up to 100 times larger sizes.

Background CheckList (Ribeiro et al., 2020) is a methodology for testing whether models have capabilities that are expected for a given task. It involves three test types: (i) minimum functionality tests (MFTs), (ii) invariance tests (INVs), and (iii) directional expectation tests (DIRs). MFTs check that the model works on simple examples, akin to unit testing in software engineering. INVs confirm that a model's prediction remains unchanged upon a minor label-preserving input change. For cases where such modifications alter labels, DIRs validate that the model changes its predictions. Tables 8–9 show examples of 53 tests for the task of identifying duplicate questions. Ribeiro et al. (2020) finetune the base BERT/RoBERTa on SST-2 (Socher et al., 2013), QQP, and SQuAD (Rajpurkar et al., 2016).[5] Their models achieve accuracy above 90, but CheckList reveals that these seemingly accurate models often lack key capabilities.

Q1: Are we at a stage where accurate models meet the expectations for their capabilities? No. Compare QQP accuracy with CheckList accuracy of 19 models finetuned for QQP in Fig. 4. There is still a notable discrepancy: QQP accuracy is consistently high among models (lowest is 89%), but CheckList accuracy varies and is substantialy lower, even as low as 51% (OPT-125M). It is reasonable to expect that these seemingly highly performing models should excel at relatively simple CheckList tests, yet they achieve accuracy higher than 95% for at most 45% of the total tests; see Table 2, row 1. Additionally, the same table shows that for a non-negligible fraction of tests, the accuracy is lower than 60%. Moreover, no model achieves over 60% accuracy on tests `{6,12,23,26,40,42,48}` that span 5 capabilities.

Q2: Are models more capable as they become larger? Generally, they are, but there are limits and irregularities. Looking at how checklist accuracy changes with model size for each model group (separated with vertical lines in Fig. 4), we see improvements that level off at suboptimal accuracies.

When focusing on specific tests, it is evident that accuracy often does not monotonically improve as the model size increases; see Figures 10–14 (Appendix). A 10% accuracy drop occurs for several tests when scaling from one model size to another (Table 2, row 6). In the same table, we report the fractions of the test for which 95+% accuracy is obtained without notable (3+%) drops during scaling (row 2) and, additionally, without notable (3+%) increases either (row 3). RoBERTa-Large, DeBERTa-Large, and GPT-2-XL achieve this commendable accuracy without scaling complications in the process, unlike T5-11B and OPT-6.7B. However, as evident from rows 2 and 3, a substantial fraction of such cases is when the model accuracy is flat across different sizes, not where scaling helps.

This does not negate the advantages of scaling altogether that are apparent from the superior checklist accuracy of, e.g., T5-3B/11B compared to their smaller counterparts. In Table 2, we also report the fraction of tests for which the accuracy of the smallest and largest model versions differ by 10% or more, while no notable drops were observed during scaling. In summary, scaling helps but is not a holistic solution, and how to finetune specialized models that have the necessary skills to do a given task robustly remains open.

4 Better Evaluation Paradigms Exist

Evaluation with sets of mutually dependent examples has been shown to reveal the fragility of models and provide a more accurate account of their true performance. Nevertheless, models' reasoning abilities continue to be assessed based on their performance on benchmarks such as MMLU (Hendrycks et al., 2021) that consist solely of i.i.d. test sets. This raises the question of whether evaluations that go beyond the i.i.d. assumption no longer offer a more reliable assessment, or if their existence is simply overlooked.

Background Motivated by the ongoing challenge of producing artifacts-free NLP benchmarks, Gardner et al. (2020) propose to evaluate with sets of examples that are minimally different from each other, but typically labeled differently. They report *contrast set consistency* which measures how often a model correctly addresses every example in a set. Models that succeed on standard benchmarks only because they exploit data shortcuts can do this rarely. Gardner et al. (2020) create contrast sets for 10 datasets, one of which is multimodal. Ravichander et al. (2022) introduce CondaQA, another contrastive dataset.

Experimental Setup Instead of training 10 (unimodal datasets) × 19 (models) × 3 (seeds) = 570 models, we reexamine the contrast set evaluation with Flan-T5-11B (Chung et al., 2022) whose instruction finetuning includes training data of all datasets in question except CondaQA. We prompt Flan-T5 with an instruction and 8 demonstrations; see Tables 14 - 17 in the Appendix. Since Flan models are instruction finetuned with chain-of-thought (CoT) prompting (Wei et al., 2022) and self-consistency (Wang et al., 2023), we also try these reasoning-boosting additions to get the highest consistency possible.[6]

Q1: Is there still a notable gap between original test sets and their contrastive counterparts? Yes, Fig. 5 shows that Flan-T5-11B's performance based on the established measurements and only

[5] `https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs`

[6] Explanations are written by one of the authors.

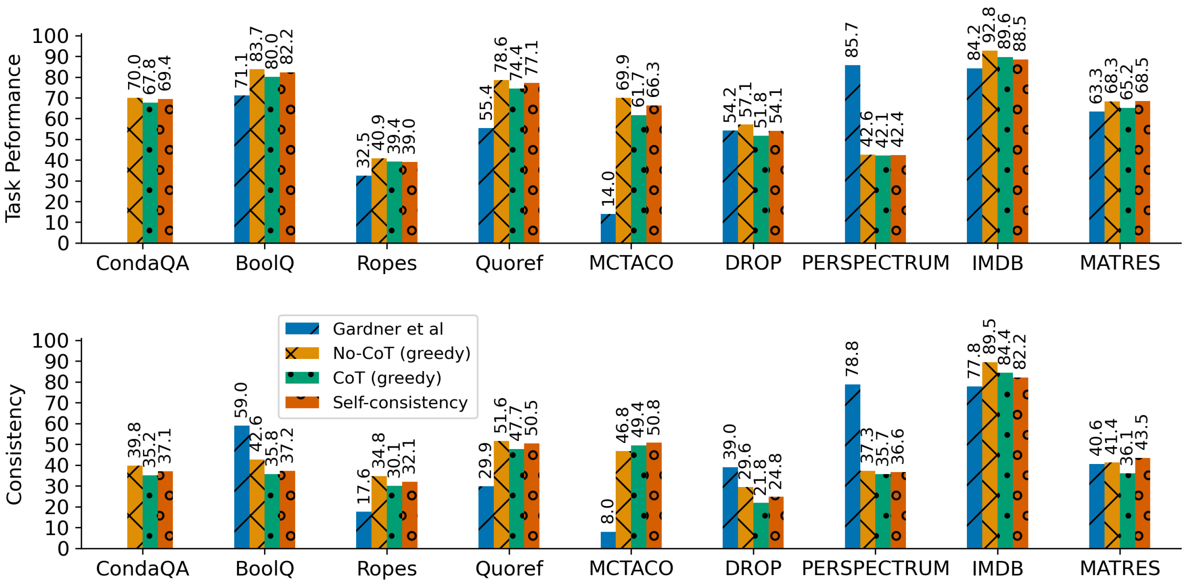


Figure 5: Flan-T5-11B performance with standard measures (accuracy, F1, token-F1) vs. contrast set consistency. Instruction finetuning data of Flan-T5-11B includes the entire training data for each task except CondaQA. Our prompts include an instruction, 8 examples, and optionally explanations for chain-of-thought prompting and self-consistency decoding.

the original instances in contrast sets is notably higher than its consistency for each contrast set.[7] Surprisingly, the benefits of including explanations are negligible — relative to the standard greedy decoding, chain-of-thoughts do not improve consistency in a single case, and self-consistency improves it only for MCTACO and MATRES.

Gardner et al. (2020) demonstrate the feasibility of creating a high-quality contrast set with 1K examples from an existing dataset in just a week's work by an expert. Considering this, along with the confirmed benefits of contrast datasets in robust reasoning evaluation, the development of contrastive versions of popular benchmarks such as BBH (Suzgun et al., 2023) and MMLU seems beneficial. Having in mind the challenges in conducting evaluations across multiple benchmarks, each with dozens of sub-tasks, we deem it is more strategic to direct a portion of resources to evaluations that extend beyond benchmarks i.i.d. with test sets than having more of the same.

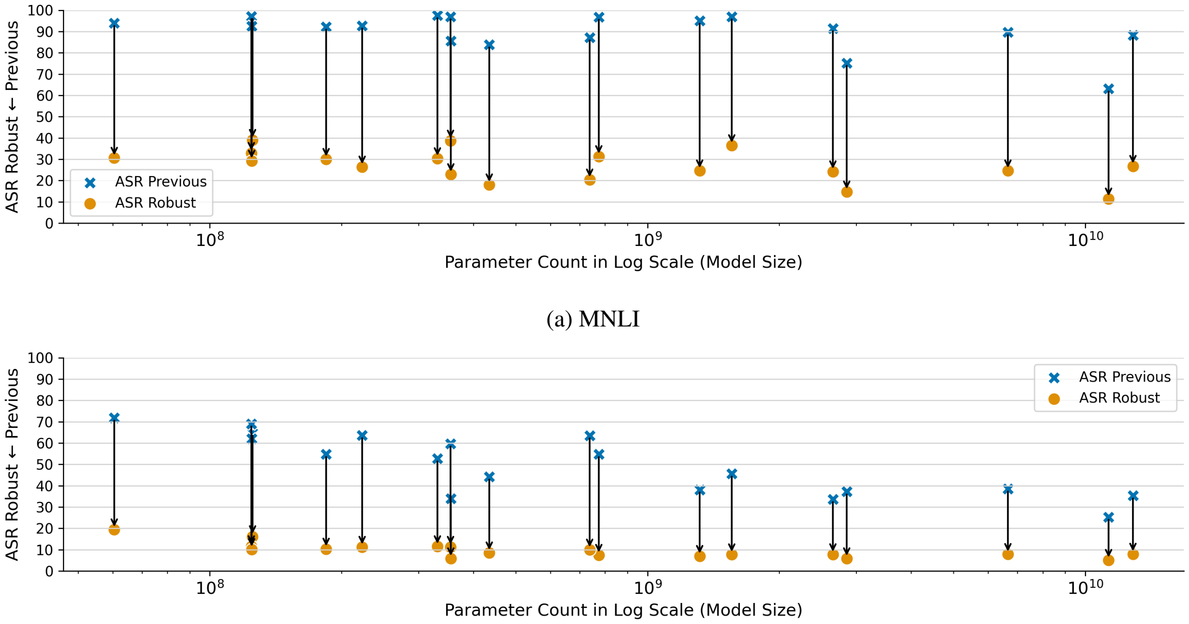
Q2: Has consistency improved amid persisting gaps? Only in some cases. The consistency values for Ropes, Quoref, MCTACO, and IMDB show large gains, yet, they remain low across the board (except for IMDB). Moreover, although Flan-T5-11B's standard task performance exceeds the results reported by Gardner et al. (2020) across almost all tasks, these improvements are not uniformly mirrored in increased consistency as evident by BoolQ and DROP. In summary, building models that are consistently robust in local neighborhoods remains a challenge.

5 Evaluating the Evaluators: The True Success of Adversarial Attacks

Despite the prevalence of the term "adversarial attacks" in 2020–2022 NLP publications, the key evaluation measurement is inadequate. While other sections of our work point to cases where models are now robust and where they remain fragile, here we bring to light a meta-issue: even some popular methods for evaluating robustness themselves need fundamental improvements to ensure our conclusions about model robustness are valid.

Background Adversarial attacks make imperceptible changes to task inputs that fool models into making mistakes. Malicious actors could use such attacks to their advantage without alerting model developers. In computer vision, "imperceptible" is the amount of noise added to the image representation that does not change its appearance (Goodfellow et al., 2015). Applying this idea to text is more complex because the nearest neighbors of noisy

[7] We do not report UD Parsing results. We do not get a reasonable performance with prompting.



(a) MNLI

(b) AGNews

Figure 6: The change in the attack success rate (ASR) as measured in prior work (1) vs. our robust modification (2). TextFooler is used to train the defense and DeepWordBug to fool 19 finetuned models in Table 2.

token embeddings could be tokens that change the original meaning. The challenge of applying attacks to text and the potential to cause harm with them, have made adversarial attacks and countermeasures, a major focus of NLP research.

Attacks We analyze the following commonly used attack methods for NLP models:

- TextFooler (Jin et al., 2020): Measures a token's importance with the change in the prediction score after removing it. Important tokens are replaced with possible synonyms found in an embedding space that have the same POS tag. Other attacks we study identify important tokens similar to TextFooler.
- BAE (Garg and Ramakrishnan, 2020): Replaces or extends important tokens with MLM (Devlin et al., 2019).
- TextBugger (Li et al., 2019): Combines character edits with embedding-based and heuristic token edits.
- PWWS (Ren et al., 2019): Replaces important tokens with WordNet synonyms with special care for named entities. Additionally, it constructs a priority order for candidate edits.
- DeepWordBug (Gao et al., 2018): Edits important tokens with 4 character-level heuristics.

All attacks edit until the prediction is altered and find edits in a black-box setting, i.e., model internals are not accessible.[8]

Defense Following Raina and Gales (2022), we train a binary classifier that distinguishes real from adversarial examples. We finetune BERT-base (Devlin et al., 2019) with a task's training examples and their adversarial versions generated by TextFooler. This defense (developed with TextFooler) is used against all attacks.

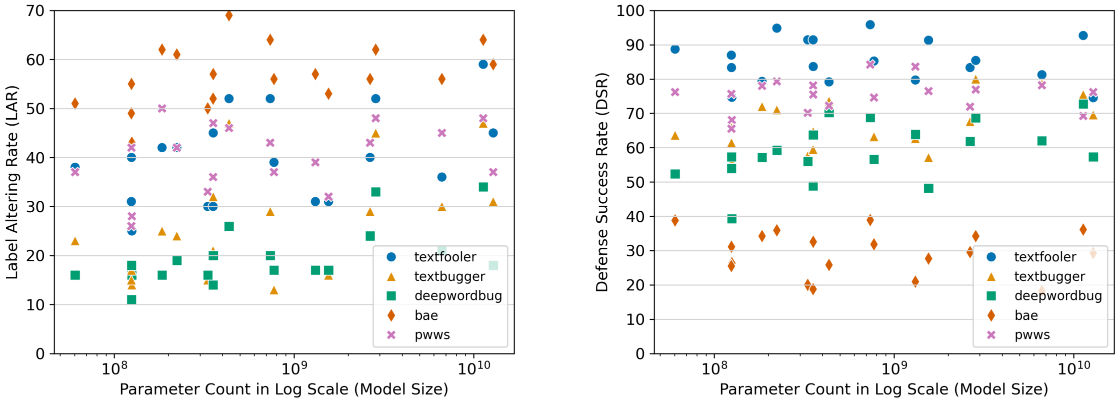
Since obvious attacking choices are to perturb tokens and/or characters, a stronger defense can be trained with a mix of attacks operating on different units (tokens, characters) and future attacks should be expected to elude such a defense.

Rigorously Determining Success of an Attack Attacks are evaluated with *attack success rate* — a fraction of instances for which an edit that alters the model's correct prediction is found:

$$\mathrm{ASR_{prev}} = \frac{1}{|\mathcal{C}|} \sum_{x \in \mathcal{C}} \mathbb{1}_{\{y_p(\mathrm{pert}(x)) \neq y_g(x)\}} \quad (1)$$

where $\mathcal{C} = \{x \in \mathcal{X} : y_p(x) = y_g(x)\}$, $y_g(\cdot)$ are gold labels, $y_p(\cdot)$ predicted, and pert is a method that alters original examples. However, this evaluation does not consider whether attacks are well formed or the effectiveness of any of the defenses

[8] TextBugger has a white-box setting version as well.



(a) LAR

(b) DSR

Figure 7: Label altering rate (LAR; 4) and defense success rate (DSR; 5) obtained in the MNLI setting. Higher values of these measurements contribute to worse effectiveness of attacks, i.e., lower attack success rate.

against them. We enhance the measurement such that this is accounted for.

Foremost, pert should not change the original label, i.e., $y_g(\text{pert}(x)) = y_g(x), \forall x \in \mathcal{X}$.[9] Moreover, we expect that an AI system has a defense that first detects whether a given example is an attack, i.e., $\text{detect}(\cdot) \in \{\texttt{real}, \texttt{attack}\}$. Our refined attack success rate is defined as:

$$\text{ASR}_{\text{our}} = \frac{1}{|\mathcal{C}|} \sum_{x \in \mathcal{C}} \text{AS}(x) \tag{2}$$

$$\text{AS}(x) = \begin{cases} 1, & \big(y_p(\text{pert}(x)) \neq y_g(x)\big) \wedge \\ & \big(\text{detect}(\text{pert}(x)) = \texttt{real}\big) \wedge \\ & \big(y_g(\text{pert}(x)) = y_g(x)\big) \\ 0, & \text{otherwise} \end{cases}$$

The challenge in properly calculating ASR is ensuring that the assumption that $y_g(\text{pert}(x)) = y_g(x)$ truly holds. In our initial analysis of perturbed examples, we found that this is often not fulfilled. Since assessing $y_g(\text{pert}(x))$ manually requires recruiting and training annotators, we suggest using a highly accurate model for the task.

A trivial way to reduce ASR is to label everything as an attack, making the system useless. Thus, ASR must be complemented with the *defense failure rate*, i.e., the rate at which the defense marks real examples as attacks:

$$\text{DFR} = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathbb{1}_{\{\text{detect}(x) = \texttt{attack}\}} \tag{3}$$

If the attack success rate or defense failure rate is high, we deem the attacker successful.

Q1: How successful are the attacks? Less successful than expected. We test the five attack methods to deceive the models in Table 2 finetuned for AGNews and MNLI. To evaluate attacks, we use a sample of 100 MNLI-matched validation examples and 100 AGNews test examples. We use GPT-4 (OpenAI, 2023) to determine $y_g(\text{pert}(x))$ and estimate that it has an error rate of 10-20%; see Appendix C for details. In Tables 18–27 (Appendix) we provide examples of produced adversarial examples. The DFR is only 6.61 on MNLI and 0.83 for AGNews, i.e., the system is still functional.[10]

In Figure 6, we show the ASR decline for the DeepWordBug which modifies characters, not tokens as TextFooler. The drops for the other four attacks are given in Figures 15–16 in the Appendix. Although DeepWordBug exhibits the smallest decline in effectiveness among attacks for MNLI, the drop is still substantial. DeepWordBug compromises at least 51.5% MNLI exmaples, and 20.3% AGNEWs examples, less than indicated by the prior ASR. As expected, the drop is more pronounced for TextFooler which was used for training the defense; see 15a and 16a). For instance, its actual success rate when attacking GPT2-XL (MNLI) falls to just 6%, a stark contrast to the 94.9% suggested by the prior ASR.

The prior work reports additional metrics: (i) the average percentage of original tokens/characters that are edited, and (ii) the average number of queries per sample. Smaller values of these two

[9] A few produce opposite-label examples with antonyms.

[10] We average DFR on the matched and mismatched MNLI evaluation data. We use the entire val/test MNLI/AGNews data for DFR.

metrics mean that the edit of the original instance is small (i.e., the edit is "imperceptible") and that the attack is efficient. Another condition that requires that x and $\mathrm{pert}(x)$ are paraphrases and/or that the edit size is minimal can be included in Eq (2). The former is easier to verify with GPT-4 because there is no universal threshold for the edit size. With this requirement, we expect ASR to reduce even more.

Q2: Why do attacks rarely succeed? To better isolate the factors that lead to lower ASR, we report two additional measurements. First, *label altering rate* — a rate at which the true label of the perturbed examples, $y_g(\mathrm{pert}(x))$, is changed:

$$\mathrm{LAR} = \frac{1}{|\mathcal{C}|} \sum_{x \in \mathcal{C}} \mathbb{1}_{\{y_g(\mathrm{pert}(x)) \neq y_g(x)\}} \tag{4}$$

Second, *defense success rate* — a rate at which the defense detects well-formed attacks:

$$\mathrm{DSR} = \frac{1}{|\mathcal{C}|} \sum_{\substack{x \in \mathcal{C} \\ \mathrm{A}(\mathrm{pert}(x))=1}} \mathbb{1}_{\{\mathrm{detect}(\mathrm{pert}(x))=\mathtt{attack}\}} \tag{5}$$

$$\mathrm{A}(\mathrm{pert}(x)) = \begin{cases} 1, & (y_g(\mathrm{pert}(x)) = y_g(x)) \wedge \\ & (y_p(\mathrm{pert}(x)) \neq y_g(x)) \\ 0, & \text{otherwise} \end{cases}$$

Attackers cannot fool models on original examples they cannot handle correctly, so DSR is checked only for examples in $\mathcal{C}$. Note that DSR and DFR do not sum to 100.[11]

Figure 7 shows LAR and DSR in the MNLI experimental setup and Figure 17 in the Appendix for AGnews. The LAR and DSR together underscore the need for multiple criteria to be met for an attack to be truly successful. The average LAR across 19 models for 3/5 attacks is 40%, even higher for `BAE`. These results demonstrate that it cannot be assumed that the label remains unchanged with perturbations. That said, `DeepWordBug` exhibits a lower rate of label mismatches, raising the question of its lower robust ASR. This is explained with a defense trained with `TextFooler`'s perturbations detecting 40–84% of attacks that were not part of its training; except `BAE`'s, which manage to bypass the defense more effectively. AGNews is less affected by label mismatching: LAR for AGNews ranges from 11% to 32%. However, the defense is effective: across models, it catches over 75% of perturbations from all attacks except `BAE`'s.

6 Related Work

A number of efforts in literature have discussed the question of robustness in NLP: through surveys (Wang et al., 2022a; Hupkes et al., 2022), new benchmarks (Ye et al., 2021; Yang et al., 2022; Yuan et al., 2023), toolkits (Goel et al., 2021), etc. While studies such as GLUE-X (Yang et al., 2022), and BOSS (Yuan et al., 2023) aim to establish new benchmarks for OOD assessment, our focus is not on creating a new benchmark. Instead, we identify OOD data splits that are no longer challenging. Our experiments, detailed in §2, reveal that sentiment classification models trained on SST-2 demonstrate robust generalization to examples from Amazon, Yelp, and IMDb, but struggle with examples from Rotten Tomatoes. This suggests that future evaluations should use this challenging train-test split. Notably, GLUE-X includes the three datasets already addressed for OOD evaluation but excludes Rotten Tomatoes.

HELM (Liang et al., 2022) advocates for a comprehensive evaluation of NLP models across dimensions like fairness, bias, robustness, and efficiency, emphasizing breadth and leaving room for more detailed exploration. For instance, its robustness assessments only involve small, semantics-preserving automatic transformations. In contrast, our evaluation is more comprehensive, encompassing domain generalization (§2), behavioral testing through Checklists (§3), consistency evaluations (via contrast sets in §4), and adversarial robustness (§5). Additionally, while Awadalla et al. (2022) explore distributional robustness for QA models trained on SQuAD (Rajpurkar et al., 2016), our analysis extends further, encompassing two diverse QA tasks: Reading Comprehension on SQuAD, NewsQA (Trischler et al., 2017), and NaturalQuestions (Kwiatkowski et al., 2019), as well as BoolQ (Clark et al., 2019).

We are not the first to discuss challenges with adversarial attack evaluations in NLP. For instance, Morris et al. (2020a) suggest additional constraints for filtering out adversarial examples. Our §5 proposal establishes a new protocol for systematically evaluating adversarial attacks. This involves careful monitoring of crucial metrics such as DFR, LAR, and DSR.

[11] Another measurement to consider is a rate at which the true label of the perturbed examples is unchanged, but the attacker does not fool the model into making a wrong prediction. However, given that $\mathrm{ASR}_{\mathrm{prev}}$ is high, we know that this rate is low and thus does not explain the drop in ASR.

7 Conclusions

This paper investigates NLP robustness from four perspectives: domain generalization, checklist-based behavioral analysis, contrast set consistency, and robustness to adversarial attacks. Notably, even the largest and most accurate models are inconsistent and succumb to fundamental task phenomena. Going beyond the current NLP model robustness landscape, our findings underscore a critical observation: the lack of robustness not only pertains to the models themselves but also encompasses deficiencies in existing methods for assessing their robustness. As progress unfolds within the field, careful consideration and continuous refinement of methodologies are imperative to ensure accurate and meaningful assessments of robustness in NLP models.

Acknowledgements

We thank Kyle Lo, Luca Soldaini, and members of the UtahNLP group for helpful feedback, Luca Soldaini for retrieving ACL publications for us, and users of the CHPC cluster who were patient with us.

References

Anas Awadalla, Mitchell Wortsman, Gabriel Ilharco, Sewon Min, Ian Magnusson, Hannaneh Hajishirzi, and Ludwig Schmidt. 2022. Exploring the landscape of distributional robustness for question answering models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5971–5987, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Samuel Bowman. 2022. The dangers of underclaiming: Reasons for caution when reporting how NLP systems fail. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7484–7499, Dublin, Ireland. Association for Computational Linguistics.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018*, pages 50–56. IEEE Computer Society.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages

1307–1323, Online. Association for Computational Linguistics.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online. Association for Computational Linguistics.

Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the NLP evaluation landscape. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 42–55, Online. Association for Computational Linguistics.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8018–8025. AAAI Press.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia. *CoRR*, abs/2303.01432.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. Textbugger: Generating adversarial text against real-world applications. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

John Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. 2020a. Reevaluating adversarial examples in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3829–3839, Online. Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020b. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Vyas Raina and Mark Gales. 2022. Residue-based natural language adversarial attack detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3836–3848, Seattle, United States. Association for Computational Linguistics.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

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

Abhilasha Ravichander, Matt Gardner, and Ana Marasović. 2022. CONDAQA: A contrastive reading comprehension dataset for reasoning about negation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8729–8755, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Priyanka Sen and Amir Saffari. 2020. What do models learn from question answering datasets? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2429–2438.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. 2020.

Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems*, volume 33, pages 18583–18599. Curran Associates, Inc.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

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.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022a. Measure and improve robustness in NLP models: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4569–4586, Seattle, United States. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022b. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. 2022. Glue-x: Evaluating natural language understanding models from an out-of-distribution generalization perspective. *arXiv e-prints*, pages arXiv–2211.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Revisiting out-of-distribution robustness in nlp: Benchmark, analysis, and llms evaluations. *arXiv preprint arXiv:2306.04618*.

A Additional Experimental Details

We fine-tune models across five model families with all the available sizes among them, giving us a total of 19 models per task. For models that are too big to train on a single GPU, we utilize DeepSpeed to perform multi-GPU training (Rajbhandari et al., 2020). For all the models except those that require DeepSpeed, we train three models with different random seeds (seeds 1, 2, and 3). In total, we have 51 models for each task.

A.1 Training Details

Learning Rate Search During preliminary experiments, we observed significant variance in task performance across different random seeds when using the default learning rates. We perform learning rate search across six learning rates to select the best learning rate: $[1e-4, 5e-4, 1e-5, 5e-5, 1e-6, 5e-6]$. To reduce the time required for learning rate search, we train each model to 1000 training steps and select the learning rate with the lowest training loss. We found this strategy works well and stabilizes training across all models we used.

Other Hyperparameters and Toolkits Used All classification models are trained with a batch size of 32 and the QA models are trained with a batch size of 8 and are determined based on the availability of the GPU resources. Wherever necessary, we employ gradient accumulation with multiple

GPUs to emulate these batch sizes. All QA models are trained with the sequence length of 512 and a document stride of 128. The maximum sequence lengths for classification tasks is dependent on each task as they vary in terms of the size of input text. All NLI and paraphrase identification models are trained with a sequence length of 256, while sentiment classification claim verification models, and topic classification (on AGNews) models are all trained with a sequence length of 512. During our preliminary experiments, we find that training for three epochs was sufficient and therefore we train all models for three epochs.

We train all our models using the Transformers library from Huggingface (Wolf et al., 2019) with the PyTorch backend (Paszke et al., 2019). For evaluating adversarial attacks, we use the TextAttack (Morris et al., 2020b) library. At the time of evaluation, the library did not support attacking text-to-text models like the T5 model for the classification tasks and is therefore implemented by ourselves.

Toolkit for Incontext Evaluations For few-shot in-context evaluations, we compared the popular lm-evaluation-harness (Gao et al., 2021) with llm-foundry and found `lm-evaluation-harness` to generally work better in reproducing the reported results.

QA Models - Span Classification vs Generative For question answering, we use the auto-regressive language models (i.e. OPT/GPT) in their generative form instead of span classification (Awadalla et al., 2022).

A.2 Pre-processing Evaluation Sets

For some evaluation sets, we did not find any publicly available data splits and therefore construct our own. For out-of-domain evaluation with qnli (the task of determining if a sentence answers a question or not), Swayamdipta et al. (2020) use the adversarial SQuAD data from Jia and Liang (2017). Since we did not find a publicly available version of this, we pre-process the data released by Jia and Liang (2017) [12] where we extract the last sentence from each passage along with the question to construct the evaluation instances. The adversarial instances are marked with the label `not-entailment`.

Similarly, the Amazon Reviews dataset [13] has been used in a number of out of domain evaluation settings for sentiment classification models. Different papers use different domains for training and evaluation. Therefore, we sample 10000 examples randomly from six genres (appliances, beauty, fashion, gift cards, magazines, and software).

Additionally, we found that the twitter paraphrase corpus was not available online[14] and thus we contact authors to get that data. The dataset provides paraphrase ratings on a scale of 1 to 6. We discard the examples with ratings 3 (recommened by authors) and classify those from 1-2 as `not-paraphrase` and from 4-6 as `paraphrase`.

The original QuAC dataset (Choi et al., 2018) contains question-answers in a multi-turn dialogue format and is therefore not directly applicable for reading comprehension. We use the converter script provided by Sen and Saffari (2020) to convert to the SQuAD format.[15]

Finally, the MultiRC dataset (Khashabi et al., 2018) which is used for out of domain evaluation with boolq, we extract only the yes/no questions from the train set as we find the validation set does not have enough of those yes/no questions.

B Llama2 7B vs Mistral 7B

For the in-context learning experiments we consider two high performing open-models, Llama2 7B (Touvron et al., 2023) and Mistral 7B (Jiang et al., 2023). To decide which to use we first compare the performance of the models on a subset of our evaluation sets including MNLI-mismatched, SST, QQP, and SQuAD. We find that Mistral 7B outperforms Llama2 across all tasks in both zero-shot and 8-shot settings. Additionally, Mistral 7B shows a consistent improvement going from zero-shot to 8-shot settings while Llama2 shows a drop in accuracy on MNLI-mismatched and a drop in f1-score for QQP. These results lead us to choose Mistral for our experiments.

C Robustness to Adversarial Attacks

We provide additional results that supplement the main body of the paper.

[12] `https://worksheets.codalab.org/worksheets/0xc86d3ebe69a3427d91f9aaa63f7d1e7d/`

[13] `https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/`

[14] `https://languagenet.github.io/`

[15] `https://github.com/amazon-science/qa-dataset-converter`

Test (Category)	**Accuracy**	**Diff**
C-IMDB	92.6	1.2
IMDB Contrast (all)	93.7	-4.3
IMDB Contrast (contrast)	92.8	1.0
IMDB Contrast (original)	94.5	-0.7
ANLI (r1)	48.0	11.7
ANLI (r2)	44.9	14.8
ANLI (r3)	45.6	14.1
Breaking NLI	77.8	-18.1
HANS (all)	56.2	3.5
HANS (constituent)	55.5	4.2
HANS (lexical overlap)	58.9	0.8
HANS (subsequence)	54.3	5.4
MNLI-hard (val matched)	55.0	4.7
MNLI-hard (val mismatched)	57.1	2.6
NLI Diagnostics (min–max)	32.1–71.4	27.6 / -11.7
Stress Test (min–max)	37.9–76.2	21.8 / 16.5
SNLI CAD	59.9	7.9
SNLI-hard	63.2	4.6
PAWS-QQP	57.2	21.2

Table 3: Incontext learning with Mistral-7B (8-shots). The last column shows the difference between accuracy in-domain performance and in the challenge set (higher means poorer generalization). Colored rows highlight datasets that are challenging for Mistral-7B.

- Figure 15–16 show the ASR drop for attacks not included in Figure 6.
- Figure 17 shows the LAR and DSR values in the AGNews experimental setup.
- Table 4 shows the label mismatch between a human (one of the authors in this case) and those assigned by GPT-4. Please refer to C.1 for discussion and anlaysis.
- Tables 18–27 provide examples produced by the five attack methods.

C.1 Analysis of Using GPT-4 to Determine Label Mismatch

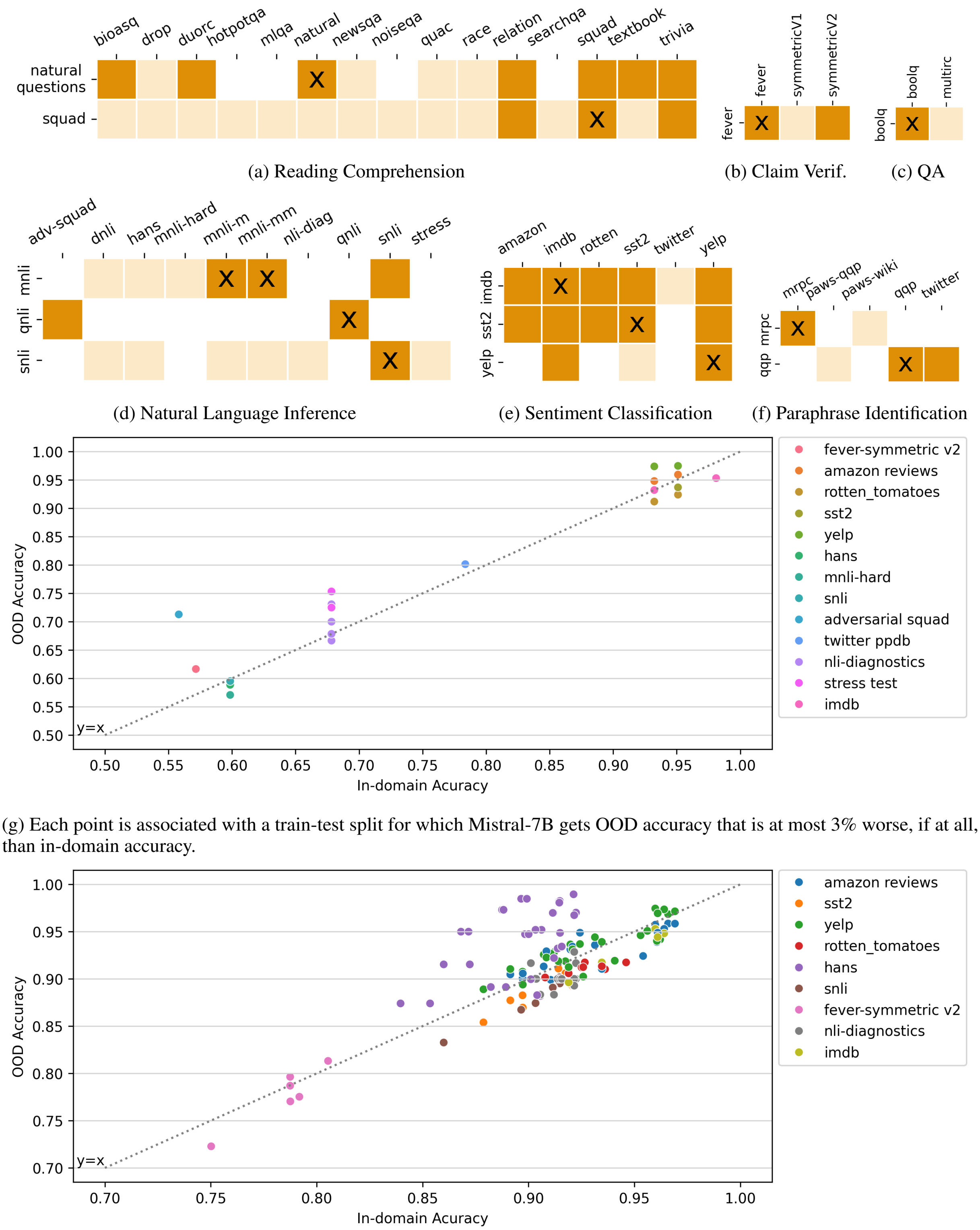
As mentioned in the main body of the text, we use GPT-4 (OpenAI, 2023) to determine the label of a perturbed example (i.e. $y_g(\text{pert}(x))$). To assess the accuracy of annotations from GPT-4, we manually evaluate 100 perturbed instances using DeepWordBug and TextFooler for both MNLI and AGNews datasets.

While annotating, we find that many of the examples generated by the attack algorithms are difficult to assign the labels to. Specifically, these cases arise when substitutions introduced by the attack algorithm either render the example incomprehensible or create difficulty in distinguishing between two potential labels (see for instance the first example in 27). Therefore, in addition to task labels, we also identify such bad/invalid examples and report them as the percentage of total annotated examples. For measuring the agreement between human assigned labels, and GPT-4 assigned labels, we report it as a percentage of examples that are assigned one of the task labels by the human annotator.

Dataset	Attack	Agreement (%)	Invalid (%)
MNLI	TextFooler	80.0	35.0
	DeepWordBug	84.3	17.0
AGNews	TextFooler	98.9	12.0
	DeepWordBug	95.8	4.0

Table 4: Agreement between GPT-4 labels and a human labeler, and % of examples classified by human as invalid.

The results are reported in table 4. We observe that, across all cases, GPT-4 has a higher agreement with human label (> 80 %). Additonally, DeepWordBug has a much lower rate of invalid examples than TextFooler. This makes sense because word-level substitutions made by TextFooler have a higher chance of destroying the meaning as compared to the character-level substitutions made by DeepWordBug.



(a) Reading Comprehension

(b) Claim Verif.

(c) QA

(d) Natural Language Inference

(e) Sentiment Classification

(f) Paraphrase Identification

(g) Each point is associated with a train-test split for which Mistral-7B gets OOD accuracy that is at most 3% worse, if at all, than in-domain accuracy.

(h) Each point is associated with a train-test split and a finetuned model. In this data split, the model gets OOD accuracy that is at most 3% worse, if at all, than in-domain accuracy.

Figure 8: Subfigures 8a–8f: Each colored square shows whether Mistral-7B in-domain (few examples and evaluation examples come from the dataset in the row) has OOD performance (few examples from the row and evaluation from the column) that is at most 3% lower than in-domain. If yes, we use fuller color, and we deem that train-test split questionable for studying OOD. We mark in-domain splits with X.

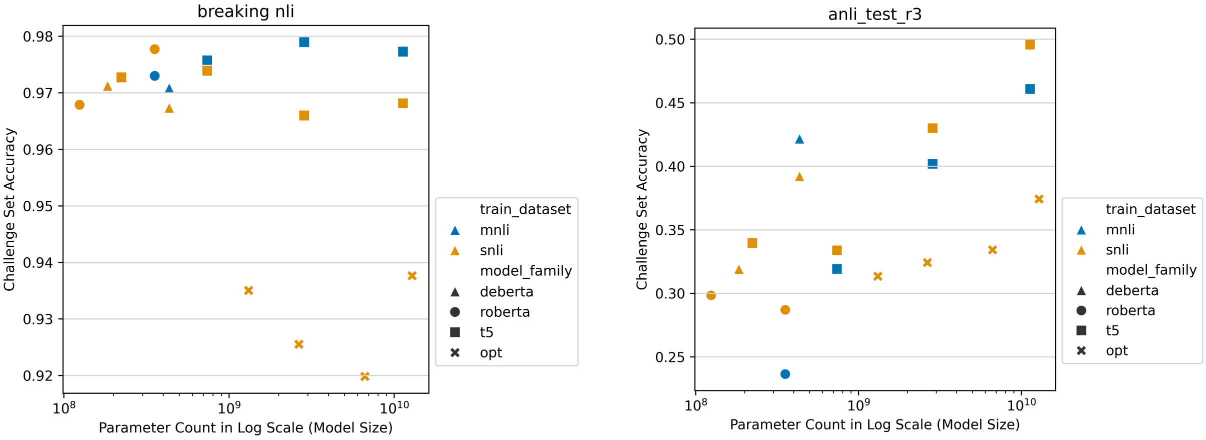


Figure 9: Breaking NLI and ANLI (r3) accuracy of all models with the in-domain accuracy accuracy of 85% or more. Breaking NLI shows that models of different types and sizes can get high challenge set accuracy, while ANLI shows that none of them can.

Test	Category	Sub-Category	Accuracy
NLI Diagnostics			95.0
NLI Diagnostics	acl	Domain	67.7
NLI Diagnostics	artificial	Domain	73.3
NLI Diagnostics	news	Domain	75.0
NLI Diagnostics	reddit	Domain	65.2
NLI Diagnostics	wikipedia	Domain	70.5
NLI Diagnostics	common sense	Knowledge	70.0
NLI Diagnostics	world knowledge	Knowledge	66.4
NLI Diagnostics	factivity	Lexical Semantics	68.6
NLI Diagnostics	lexical entailment	Lexical Semantics	76.4
NLI Diagnostics	morphological negation	Lexical Semantics	91.0
NLI Diagnostics	named entities	Lexical Semantics	70.4
NLI Diagnostics	quantifiers	Lexical Semantics	82.7
NLI Diagnostics	redundancy	Lexical Semantics	88.5
NLI Diagnostics	symmetry/collectivity	Lexical Semantics	73.8
NLI Diagnostics	conditionals	Logic	75.0
NLI Diagnostics	conjunction	Logic	87.5
NLI Diagnostics	disjunction	Logic	47.4
NLI Diagnostics	double negation	Logic	94.1
NLI Diagnostics	downward monotone	Logic	30.0
NLI Diagnostics	existential	Logic	73.3
NLI Diagnostics	intervals/numbers	Logic	63.2
NLI Diagnostics	negation	Logic	76.8
NLI Diagnostics	non-monotone	Logic	63.3
NLI Diagnostics	temporal	Logic	71.9
NLI Diagnostics	universal	Logic	94.4
NLI Diagnostics	upward monotone	Logic	79.4
NLI Diagnostics	active/passive	Predicate-Argument Structure	62.8
NLI Diagnostics	anaphora/coreference	Predicate-Argument Structure	74.7
NLI Diagnostics	coordination scope	Predicate-Argument Structure	72.5
NLI Diagnostics	core args	Predicate-Argument Structure	76.3
NLI Diagnostics	datives	Predicate-Argument Structure	85.0
NLI Diagnostics	ellipsis/implicits	Predicate-Argument Structure	81.4
NLI Diagnostics	genitives/partitives	Predicate-Argument Structure	95.0
NLI Diagnostics	intersectivity	Predicate-Argument Structure	63.0
NLI Diagnostics	nominalization	Predicate-Argument Structure	92.9
NLI Diagnostics	prepositional phrases	Predicate-Argument Structure	86.8
NLI Diagnostics	relative clauses	Predicate-Argument Structure	68.8
NLI Diagnostics	restrictivity	Predicate-Argument Structure	69.2
Stress Test	validation matched		90.6
Stress Test	validation matched	Antonym	89.1
Stress Test	validation matched	Length_Mismatch	89.5
Stress Test	validation matched	Negation	76.8
Stress Test	validation matched	Spelling Error (contentword_swap_perturbed)	89.7
Stress Test	validation matched	Spelling Error (functionword_swap_perturbed)	90.6
Stress Test	validation matched	Spelling Error (keyboard)	90.1
Stress Test	validation matched	Spelling Error (swap)	90.2
Stress Test	validation matched	Word_Overlap	83.1
Stress Test	validation mismatched		90.5
Stress Test	validation mismatched	Antonym	88.4
Stress Test	validation mismatched	Length Mismatch	90.2
Stress Test	validation mismatched	Negation	77.4
Stress Test	validation mismatched	Spelling Error (contentword_swap_perturbed)	89.6
Stress Test	validation mismatched	Spelling Error (functionword_swap_perturbed)	90.5
Stress Test	validation mismatched	Spelling Error (keyboard)	89.5
Stress Test	validation mismatched	Spelling Error (swap)	90.5
Stress Test	validation mismatched	Word Overlap	82.9

Table 5: The breakdown of individual tests in NLI Diagnostics and Stress Test.

Train	Eval	Metrics	Evaluation Types	Included?	Remarks
mnli	hans	Accuracy	challenge, domain	✓	
	mnli-hard	Accuracy	challenge, domain	✓	
	anli	Accuracy	challenge	✓	
	breaking nli	Accuracy	challenge	✓	
	nli-diagnostics	Matthew's Corr, Accuracy	challenge	✓	
	stress test	Accuracy	challenge	✓	
	snli	Accuracy	domain	✓	
	dnli	Accuracy	domain	✓	
	mnli-matched	Accuracy	in-domain	✓	
	mnli-mismatched	Accuracy	in-domain	✓	
snli	mnli	Accuracy	domain	✓	
	hans	Accuracy	challenge, domain	✓	
	anli	Accuracy	challenge	✓	
	snli-hard	Accuracy	challenge	✓	
	nli-diagnostics	Matthew's Corr, Accuracy	challenge, domain	✓	
	dnli	Accuracy	domain	✓	
	stress test	Accuracy	challenge, domain	✓	
	snli cad	Accuracy	challenge	✓	
	breaking nli	Accuracy	challenge	✓	
	snli	Accuracy	in-domain	✓	
qqp	twitter ppdb	Accuracy	domain	✓	Dataset not available at the link. Acquired via email.
	paws-qqp	Accuracy, AUC	challenge, domain	✓	
	qqp	Accuracy, F1	in-domain	✓	
yelp	imdb	Accuracy	domain	✓	
	sst2 yelp	Accuracy Accuracy	domain in-domain	✓ ✓	
sst2	imdb	Accuracy	domain	✓	
	imdb contrast	Accuracy	challenge	✓	
	c-imdb	Accuracy	challenge	✓	
	yelp	Accuracy	domain	✓	
	amazon reviews	Accuracy	domain	✓	Six Genres: appliances, beauty, fashion, gift_cards, magazines, software (10k examples each)
	rotten_tomatoes	Accuracy	domain	✓	
	sst2	Accuracy	in-domain	✓	
imdb	sst2	Accuracy	domain	✓	
	yelp	Accuracy	domain	✓	
	imdb contrast	Accuracy	challenge	✓	
	c-imdb	Accuracy	challenge	✓	
	twitter emotion	Accuracy	domain	✓	SemEval-2017 Task - 4 - Twitter Sentiment Analysis
	amazon reviews	Accuracy	domain	✓	Six Genres: appliances, beauty, fashion, gift_cards, magazines, software (10k examples each)
	rotten_tomatoes	Accuracy	domain	✓	
	imdb	Accuracy	in-domain	✓	
mrpc	paws-wiki mrpc	Accuracy, AUC Accuracy, F1	domain in-domain	✓ ✓	
fever	fever-symmetric v1	Accuracy	challenge, domain	✓	
	fever-symmetric v2	Accuracy	challenge, domain	✓	
	fever	Accuracy	in-domain	✓	Used as NLI
qnli	adversarial squad	Accuracy	domain	✓	Used in Dataset Cartography for the first time. Not Available. Created from the source
	qnli	Accuracy	in-domain	✓	

Table 6: Evaluation details

Train	Eval	Metrics	Evaluation Types	Included?	Remarks
squad	trivia qa	Exact Match, F1	domain	✓	
	natural questions	Exact Match, F1	domain, challenge	✓	
	newsqa	Exact Match, F1	domain, challenge	✓	
	hotpotqa	Exact Match, F1	domain	✓	
	searchqa	Exact Match, F1	domain	✓	
	bioasq	Exact Match, F1	domain	✓	
	textbook qa	Exact Match, F1	domain	✓	
	xquad	Exact Match, F1	domain	✗	English subset already included in SQuAD validation
	non-adversarial paraphrased	Exact Match, F1	challenge	✓	
	adversarial paraphrased	Exact Match, F1	challenge	✓	
	squad-hard	Exact Match, F1	challenge	✓	
	squad-implications	Exact Match, F1	challenge	✗	Generating set of implications requires each model's predictions.
	addonesent	Exact Match, F1	challenge	✓	The link on HF: https://huggingface.co/datasets/squad _adversarial gives loading error. The paper (https://aclanthology.org/2021.acl-srw.21.pdf) uses it wrongly. The examples come from 1000 dev examples. The original file contains these 1000 extra clean dev examples that need to be removed. So our count of these are 2560 and 787 instead of 3560 and 1787 as reported in the paper.
	addsent	Exact Match, F1	challenge	✓	The link on HF: https://huggingface.co/datasets/squad _adversarial gives loading error. The paper (https://aclanthology.org/2021.acl-srw.21.pdf) uses it wrongly. The examples come from 1000 dev examples. The original file contains these 1000 extra clean dev examples that need to be removed. So our count of these are 2560 and 787 instead of 3560 and 1787 as reported in the paper.
	quoref	Exact Match, F1	challenge	✓	
	qa_contrast	Exact Match, F1	challenge	✗	Mentioned in https://arxiv.org/pdf/2107.07150.pdf They mention: "For SQuAD, we evaluate a fine-tuned RoBERTa, the most downloaded model hosted on Huggingface,17 and use the QA impli- cation challenge set (Rajpurkar et al., 2016) as the human contrast set" Not clear what they are actually referring to.
	quac	Exact Match, F1	domain	✓	Original dataset has multi-turn dialogues. We convert to squad format using this: https://github.com/amazon-science/qa-dataset-converter
	drop	Exact Match, F1	domain	✓	
	duorc	Exact Match, F1	domain	✓	
	race	Exact Match, F1	domain	✓	
	relation extraction	Exact Match, F1	domain	✓	
	mlqa	Exact Match, F1	domain	✓	Using only en subset
	squad	Exact Match, F1	in-domain	✓	
newsqa	squad	Exact Match, F1	domain	✓	
	natural questions	Exact Match, F1	domain	✓	
	trivia qa	Exact Match, F1	domain	✓	
	quac	Exact Match, F1	domain	✓	Original dataset has multi-turn dialogues. We convert to squad format using this: https://github.com/amazon-science/qa-dataset-converter
	bioasq	Exact Match, F1	domain	✓	
	drop	Exact Match, F1	domain	✓	
	duorc	Exact Match, F1	domain	✓	
	race	Exact Match, F1	domain	✓	
	relation extraction	Exact Match, F1	domain	✓	
	textbook qa	Exact Match, F1	domain	✓	
	newsqa	Exact Match, F1	in-domain	✓	
natural questions	trivia qa squad	Exact Match, F1 Exact Match, F1	domain domain	✓ ✓	
	newsqa	Exact Match, F1	domain	✓	
	bioasq	Exact Match, F1	domain	✓	
	drop	Exact Match, F1	domain	✓	
	duorc	Exact Match, F1	domain	✓	
	race	Exact Match, F1	domain	✓	
	relation extraction	Exact Match, F1	domain	✓	
	textbook qa	Exact Match, F1	domain	✓	
	quac	Exact Match, F1	domain	✓	
	trec	Exact Match, F1	domain	✗	Only applicable for open-domain QA on naturalQuestions
	ambigqa	Exact Match, F1	domain	✗	Only applicable for open-domain QA on naturalQuestions
	natural questions	Exact Match, F1	in-domain	✓	
boolq	boolq contrast gardner	Exact Match or Accuracy	challenge	✓	other name boolq_contrast_gardner
	boolq cad	Exact Match or Accuracy	challenge	✓	
	multirc	Exact Match or Accuracy	domain	✓	Used only train set questions with yes/no answers
	boolq	Exact Match or Accuracy	in-domain	✓	

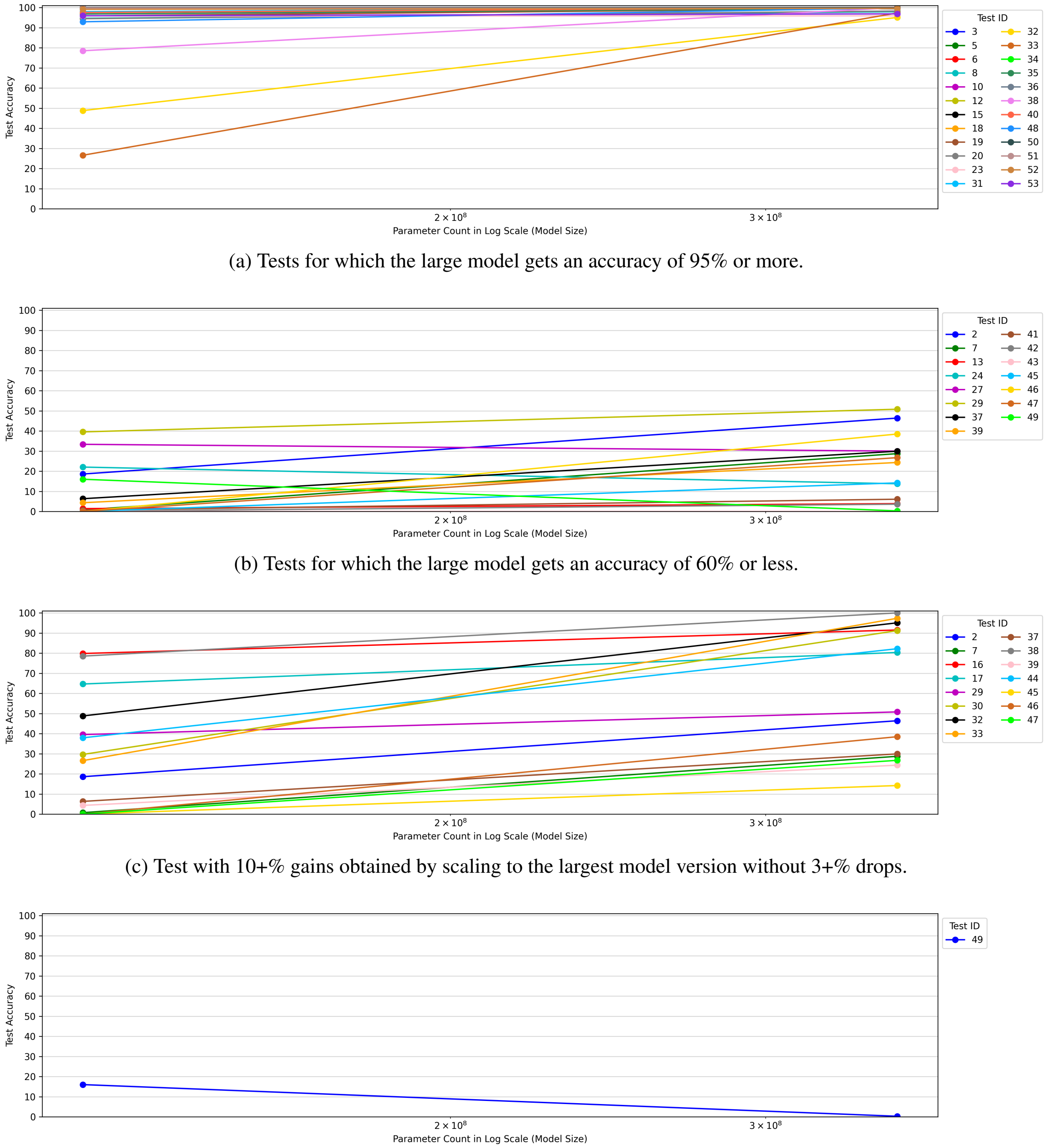
Table 7: Evaluation details

Test Id; Capability; Test Type; *Test Description*: Q1: `{question}` Q2: `{question}` ({label})			
1	Vocabulary	MFT	*Add an adjective (modifier)*: Q1: Is Adam Ward a historian? Q2: Is Adam Ward an aspiring historian? (not duplicates)
2		MFT	*Different adjectives*: Q1: Is Jason Price an immigrant? Q2: Is Jason Price Indian? (not duplicates)
3		MFT	*Different animals*: Q1: Can I feed my dog cereal? Q2: Can I feed my snake cereal? (not duplicates)
4		MFT	*Add irrelevant modifiers (examples with animals)*: Q1: Is that monkey up on the table? Q2: Is that monkey truly up on the table? (duplicates)
5		MFT	*Add irrelevant modifiers (examples with people)*: Q1: Is Melissa responding to Christina? Q2: Is Melissa really responding to Christina? (duplicates)
6		MFT	*Different irrelevant preamble*: Q1: My pet cat eats soy. Is it normal for animals to eat soy? Q2: My pet monkey eats soy. Is it normal for animals to eat soy? (duplicates)
7		MFT	*Preamble is relevant (different injuries)*: Q1: I hurt my hip last time I played football. Is this a common injury? Q2: I hurt my thigh last time I played football. Is this a common injury? (not duplicates)
8	Taxonomy	MFT	*How can I become more {synonym}?*: Q1: How can I become more religious? Q2: How can I become more spiritual? (duplicates)
9		INV	`(question, f(question))` *where* `f(question)` *replaces synonyms?*: Q1: I am a 32 year old single man, doing a govt job in India, not happy with my job and life, nothing much in my bank account, what should I do? Q2: I am a 32 year old single man, doing a govt job in India, not joyful with my job and life, nothing much in my bank account, what should I do? (duplicates)
10		INV	*Replace synonyms in pairs of duplicates from the dev set*: Q1: What is the secret of happy life? Q2: What's the bloody secret of a happy life? (duplicates) ‖ Q1: What is the secret of joyful life? Q2: What's the bloody secret of a happy life? (duplicates)
11		MFT	*How can I become more X ≠ How can I become less X*: Q1: How can I become less secular? Q2: How can I become more secular? (not duplicates)
12		MFT	*How can I become more X = How can I become less antonym(X)*: Q1: How can I become less hopeful? Q2: How can I become more hopeless? (not duplicates)
13	Robustness	INV	*Add one typo:* Q1: What are the best tisp for early-stage startups? Q2: What are your best tips for very early stage startups? (duplicates) ‖ Q1: What are the best tips for early-stage startups? Q2: What are your best tips for very early stage statrups? (duplicates) ‖ ..
14		INV	*Contractions*: Q1: What are the qualifications for being an FBI or CIA agent? Q2: What does it take to become an FBI agent? (not duplicates) ‖ Q1: What're the qualifications for being an FBI or CIA agent? Q2: What does it take to become an FBI agent? (not duplicates) ‖ ...
15		DIR	`(q, paraphrase(q))`: Q1: Do you think you can use another opertator's SIM in Jio SIM slot after using Jio SIM? Q2: Can you use another opertator's SIM in Jio SIM slot after using Jio SIM? (duplicates)
16		INV	*Product of* `paraphrases(q1) * paraphrases(q2)`: Q1: If you want to publish poetry on Quora, what should you do? Q2: Do you think you can post your poetry on Quora? (not duplicates)‖ Q1: In order to publish poetry on Quora, what should you do? Q2: Can you post my poetry on Quora? (not duplicates) ‖ ...
17	NER	MFT	*Same adjectives, different people*: Q1: Is Samuel Rogers Australian? Q2: Is Joshua James Australian? (not duplicates)
18		MFT	*Same adjectives, different people v2:* Q1: Is Eric Wilson Jewish? Q2: Is Victoria Wilson Jewish? (not duplicates)
19		MFT	*Same adjectives, different people v3*: Q1: Is Olivia Edwards Muslim? Q2: Is Olivia Reyes Muslim? (not duplicates)
20		INV	*Change same name in both questions:* Q1: Did Jesus keep the sabbath? Q2: When Jesus died on the cross did he do away with keeping the seventh year sabbath? (not duplicates) ‖ Q1: Did Kyle keep the sabbath? Q2: When Kyle died on the cross did he do away with keeping the seventh year sabbath? (not duplicates)
21		INV	*Change same location in both questions*: Q1: Why does the caste system persist in India? Q2: Do you support the caste system in India? ... (not duplicates)
22		INV	*Change same number in both questions*: Q1: How can I invest $100 into myself? Q2: What is the best way to invest $100 in todays market? (not duplicates) ‖ Q1: How can I invest $103 into myself? Q2: What is the best way to invest $103 in todays market? ... (not duplicates)
23		DIR	*Change first name in one of the questions*: Q1: What does Hillary Clinton think of high-skill immigration? Q2: What is Hillary Clinton's stance on high skilled immigration? (duplicates) ‖ Q1: What does Hillary Clinton think of high-skill immigration Q2: What is Diana Clinton's stance on high skilled immigration? (not duplicates) ‖ ...
24		DIR	*Change first and last name in one of the questions*: Q1: Would Hillary get women's vote just because she's a female? Q2: Are there a lot of women who will vote for Hillary Clinton just because she is a woman? (duplicates) ‖ Q1: Would Brooke get women's vote just because she's a female? Q2: Are there a lot of women who will vote for Hillary Clinton just because she is a woman? (not duplicates) ‖ ...
25		DIR	*Change location in one of the questions*: Q1: Why did India sign the Indus Water Treaty? Q2: Why did India signed Indus water treaty? (duplicates) ‖ Q1: Why did Nauru sign the Indus Water Treaty? Q2: Why did India signed Indus water treaty? (not duplicates) ‖ ...
26		DIR	*Change numbers in one of the questions*: Q1: What do you think of abolishing 500 and 1000 Rupee Currency notes by the Indian Government? Q2: Was the decision by the Indian Government to demonetize 500 and 1000 notes right or is it a big scam? (duplicates) ‖ Q1: What do you think of abolishing 500 and 931 Rupee Currency notes by the Indian Government? Q2: Was the decision by the Indian Government to demonetize 500 and 1000 notes right or is it a big scam? (not duplicates) ‖ ...
27		DIR	*Keep entities, fill in with gibberish*: Q1: What would have happened if Hitler hadn't declared war on the United States after Pearl Harbor? Q2: What would have happened if the United States split in two after the revolutionary war? (not duplicates) ‖ Q1: What would have happened if the United States split in two after the revolutionary war? Q2: What divided the United States in two after the revolutionary war? (not duplicates) ‖ ...

Table 8: Checklists tests 1–27 for QQP.

Test Id	**Capability**	**Test Type**	***Test Description*: Q1: `{question}` Q2: `{question}` ({label})**
28	Temporal	MFT	*Is person X ≠ Did person use to be X*: Q1: Is James Russell an actress? Q2: Did James Russell use to be an actress? (not duplicates)
29		MFT	*Is person X ≠ Is person becoming X*: Q1: Is Taylor Long an investigator? Q2: Is Taylor Long becoming an investigator? (not duplicates)
30		MFT	*What was person's life before becoming X ≠ [...] after becoming X*: Q1: What was Kyle Ross's life before becoming an academic? Q2: What was Kyle Ross's life after becoming an academic? (not duplicates)
31		MFT	*Do you have to X your dog before Y it ≠ [...] after Y it*: Q1: Do you have to weigh your dog before naming it? Q2: Do you have to weigh your dog after naming it? (not duplicates)
32		MFT	*Is it {ok, ...} to {smoke, ...} after ≠ before*: Q1: Is it reasonable to text before 7pm? Q2: Is it reasonable to text after 7pm? (not duplicates)
33	Negation	MFT	*How can I become a X person ≠ [...] a person who is not X*: Q1: How can I become an invisible person? Q2: How can I become a person who is not invisible? (not duplicates)
34		MFT	*Is it {ok, ...} to {smoke, ...} in country ≠ [...] not to [...]* Q1: Is it socially acceptable to preach in Tanzania? Q2: Is it socially acceptable not to preach in Tanzania? (not duplicates)
35		MFT	*What are things a {noun} should worry about ≠ [...] not worry about*: Q1: What are things an escort should worry about? Q2: What are things an escort should not worry about? (not duplicates)
36		MFT	*How can I become a X person = [...] a person who is not antonym(X)*: Q1: How can I become a smart person? Q2: How can I become a person who is not stupid? (duplicates)
37	Coref	MFT	*Simple coref (he and she)*: Q1: If Olivia and Donald were alone, do you think he would reject her? Q2: If Olivia and Donald were alone, do you think she would reject him? (not duplicates)
38		MFT	*Simple coref (his and her)*: Q1: If George and Jasmine were married, would his family be happy? Q2: If George and Jasmine were married, would Jasmine's family be happy? (not duplicates) || Q1: If George and Jasmine were married, would her family be happy? Q2: If George and Jasmine were married, would George's family be happy? (not duplicates)
39	SRL	MFT	*Who do X think = Who is the [...] according to X* Q1:Who do critics think is the brightest boxer in the world? Q2: Who is the brightest boxer in the world according to critics? (duplicates)
40		MFT	*Order does not matter for comparison*: Q1: Are dwarves warmer than men? Q2: Are men warmer than dwarves? (not duplicates)
41		MFT	*Order does not matter for symmetric relations*: Q1: Is Hannah engaged to Isabella? Q2: Is Isabella engaged to Hannah? (duplicates)
42		MFT	*Order does matter for asymmetric relations*: Q1: Is Elizabeth beating Adam? Q2: Is Adam beating Elizabeth? (not duplicates)
43		MFT	*Traditional SRL: active / passive swap*: Q1: Did Samuel miss the estate? Q2: Was the estate missed by Samuel? (duplicates)
44		MFT	*Traditional SRL: wrong active / passive swap* Q1: Did Michelle like the car? Q2: Was Michelle liked by the car? (not duplicates)
45		MFT	*Traditional SRL: active / passive swap with people*: Q1: Does Mary remember Adam? Q2: Is Adam remembered by Mary? (duplicates)
46		MFT	*Traditional SRL: wrong active / passive swap with people*: Q1: Does Michelle trust Angela? Q2: Is Michelle trusted by Angela? (not duplicates)
47	Logic	MFT	*A or B is not the same as C and D*: Q1: Is Emily Fisher an actress or an investor? Q2: Is Emily Fisher simultaneously an auditor and an organizer? (not duplicates)
48		MFT	*A or B is not the same as A and B*: Q1: Is Taylor King an educator or an accountant? Q2: Is Taylor King simultaneously an educator and an accountant? (not duplicates)
49		MFT	*A and / or B is the same as B and / or A*: Q1: Is Jennifer Flores an engineer and an editor? Q2: Is Jennifer Flores an editor and an engineer? (duplicates)
50		MFT	*a* `{nationality}` `{profession}` = *a* `{profession}` *and* `{nationality}`: Q1: Is Christina Nguyen a French nurse? Q2: Is Christina Nguyen a nurse and French? (duplicates)
51		MFT	*Reflexivity:* `(q,q)` *should be duplicate*: Q1: What does the following symbol mean Q2: What does the following symbol mean ? (duplicates)
52		INV	*Symmetry:* `f(a,b) = f(b,a)`: Q1: Which colleges come under the GMAT? Q2: Which all colleges come under GMAT in india? (not duplicates) || Q1: Which all colleges come under GMAT in india? Q2: Which colleges come under the GMAT? (not duplicates)
53		DIR	*Testing implications*: Q1: Why was Albert Einstein considered an atheist? Q2: Do atheists look down on Albert Einstein because he was religious? (not duplicates) || Q1: Why was Albert Einstein considered an atheist? Q2: Was Albert Einstein an atheist? (not duplicates) || ...

Table 9: Checklists tests 28–53 for QQP.



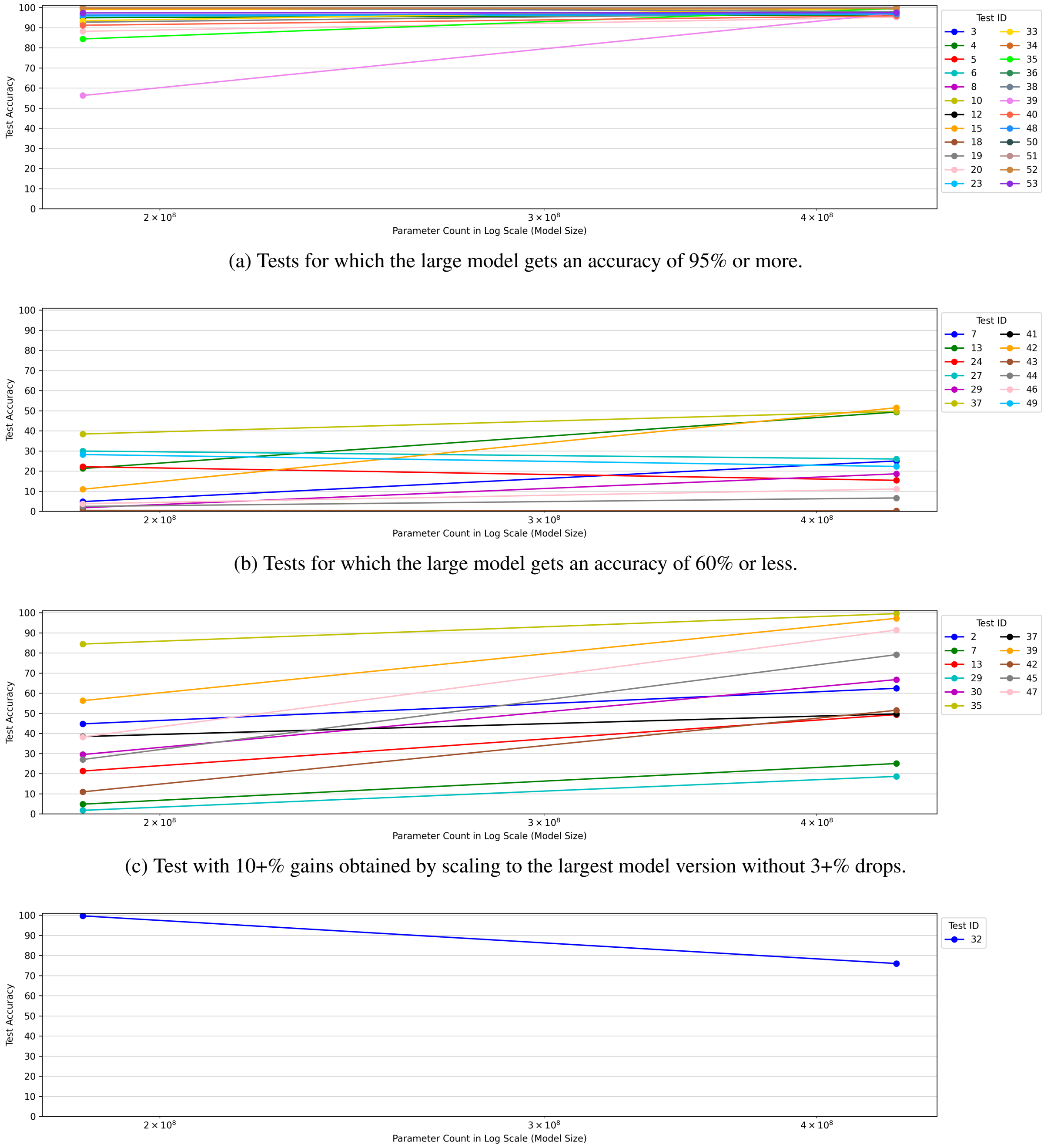
(a) Tests for which the large model gets an accuracy of 95% or more.

(b) Tests for which the large model gets an accuracy of 60% or less.

(c) Test with 10+% gains obtained by scaling to the largest model version without 3+% drops.

(d) Tests with scaling complications (10+% drops).

Figure 10: RoBERTa



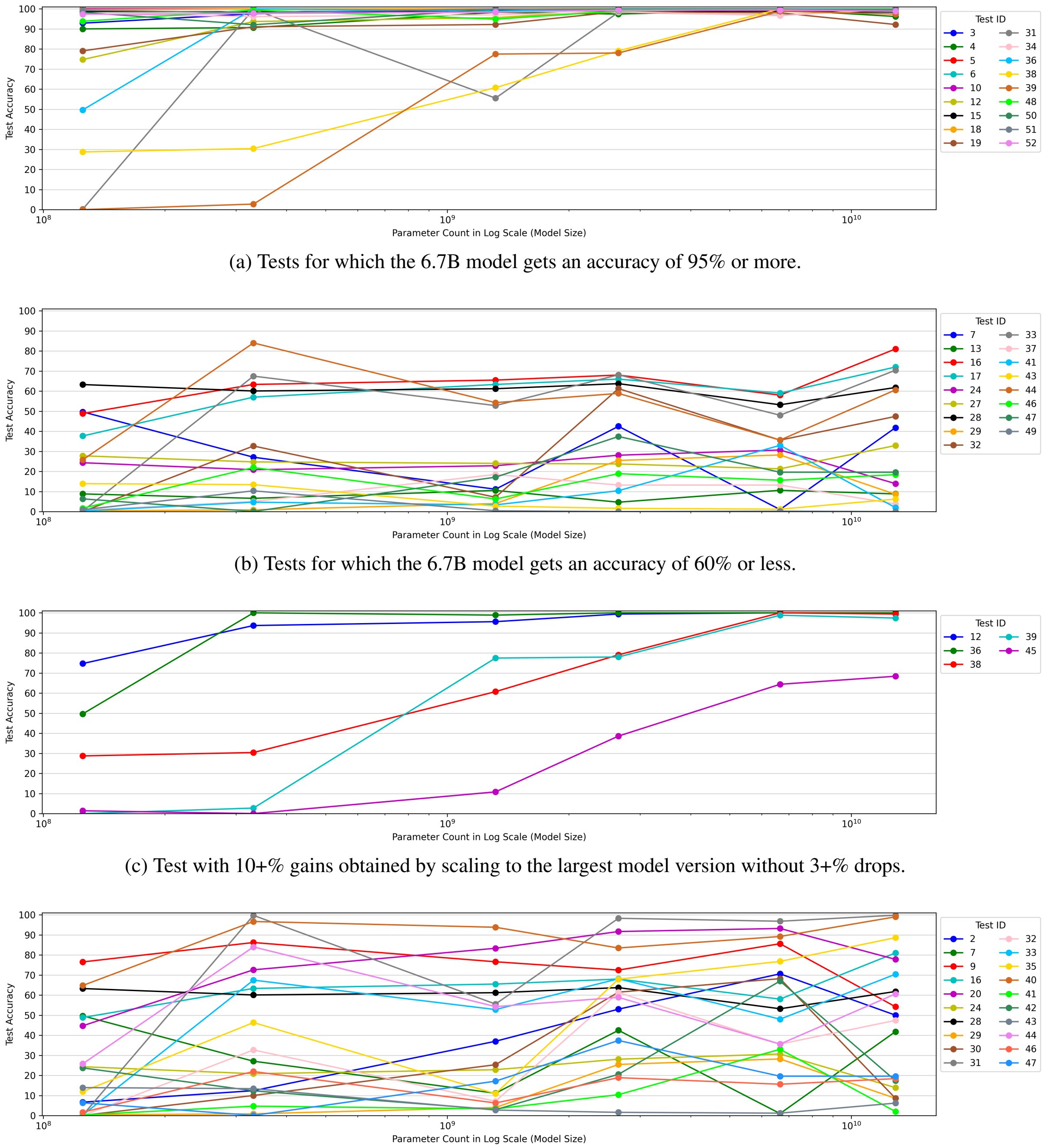
(a) Tests for which the large model gets an accuracy of 95% or more.

(b) Tests for which the large model gets an accuracy of 60% or less.

(c) Test with 10+% gains obtained by scaling to the largest model version without 3+% drops.

(d) Tests with scaling complications (10+% drops).

Figure 11: DeBERTa



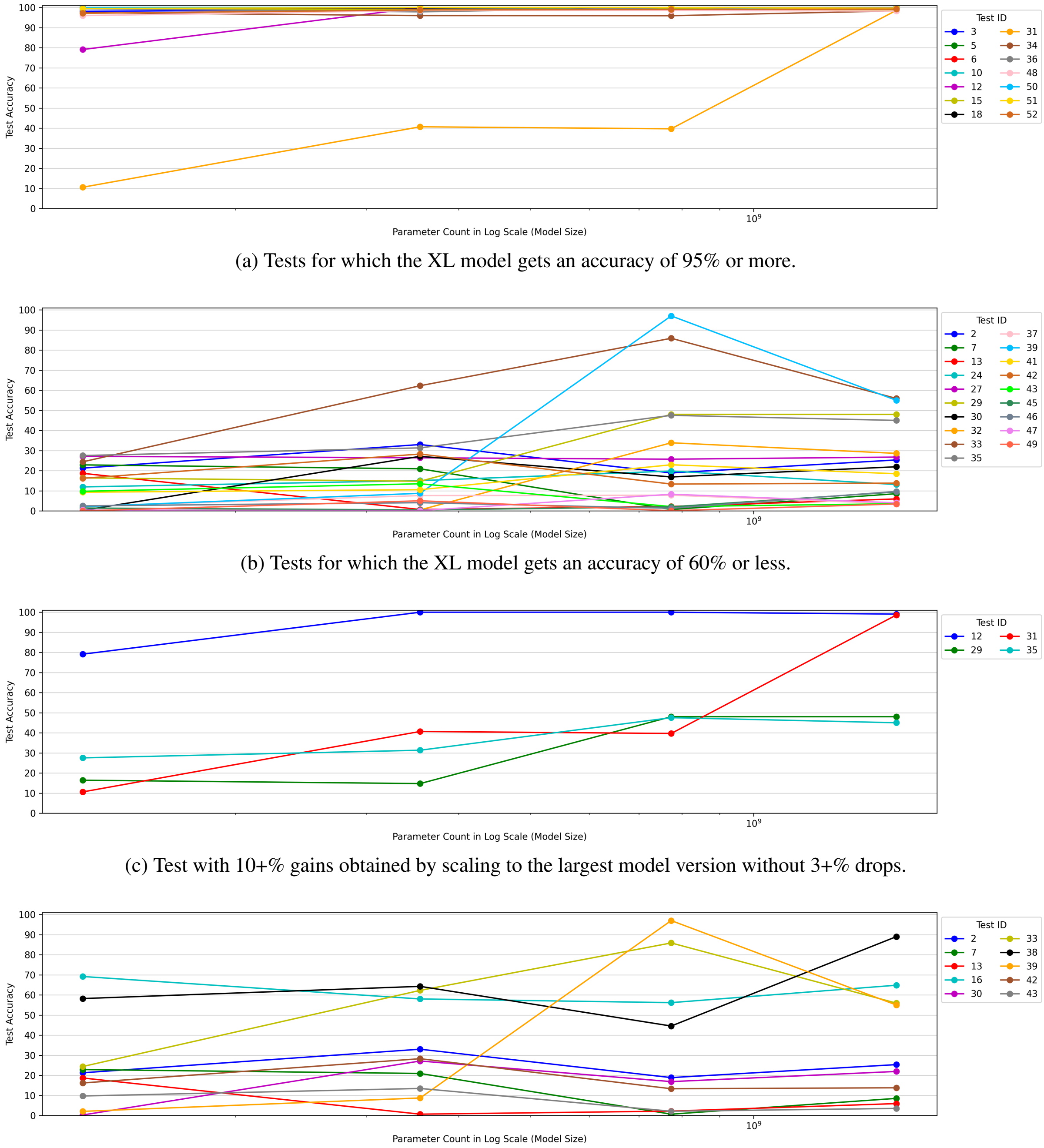
(a) Tests for which the 6.7B model gets an accuracy of 95% or more.

(b) Tests for which the 6.7B model gets an accuracy of 60% or less.

(c) Test with 10+% gains obtained by scaling to the largest model version without 3+% drops.

(d) Tests with scaling complications (10+% drops).

Figure 12: OPT



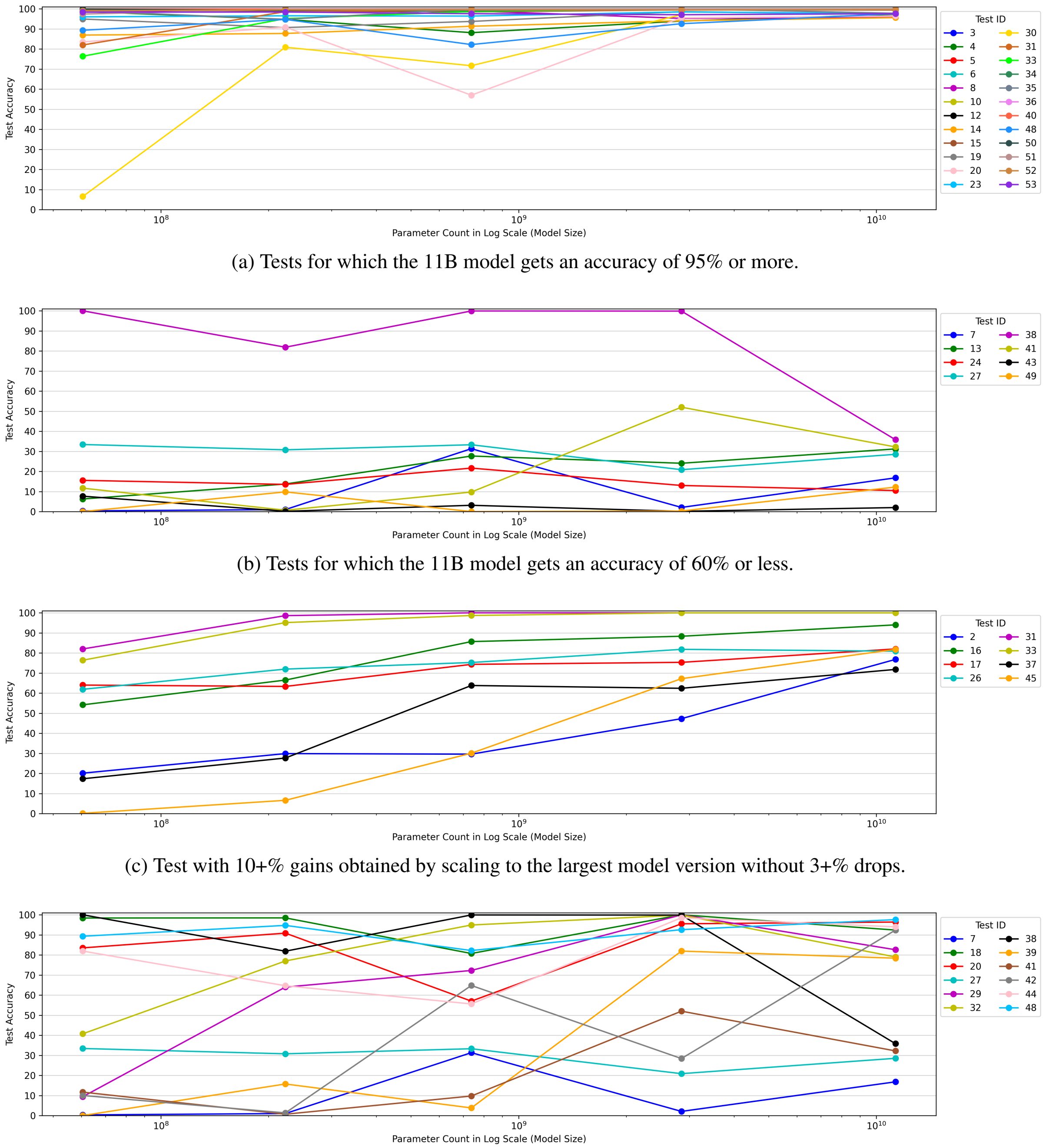
(a) Tests for which the XL model gets an accuracy of 95% or more.

(b) Tests for which the XL model gets an accuracy of 60% or less.

(c) Test with 10+% gains obtained by scaling to the largest model version without 3+% drops.

(d) Tests with scaling complications (10+% drops).

Figure 13: GPT-2



(a) Tests for which the 11B model gets an accuracy of 95% or more.

(b) Tests for which the 11B model gets an accuracy of 60% or less.

(c) Test with 10+% gains obtained by scaling to the largest model version without 3+% drops.

(d) Tests with scaling complications (10+% drops).

Figure 14: T5

Task	CoT Prompt
CondaQA	**Passage:** The end of the long-held animosity between Moscow and Beijing was marked by the visit to China by Soviet General Secretary Mikhail Gorbachev in 1989. After the 1991 demise of the Soviet Union, China's relations with Russia and the former states of the Soviet Union became more amicable as the conflicting ideologies of the two vast nations no longer stood in the way. A new round of bilateral agreements was signed during reciprocal head of state visits. As in the early 1950s with the Soviet Union, Russia has again become an important source of military technology for China, as well as for raw materials and trade. Friendly relations with Russia have been an important advantage for China, complementing its strong ties with the U.S. **Question:** Can China rely on both the US and Russia as supportive allies? **Give the rationale before answering.** If a country has either friendly relations or strong ties with another country, one can expect the other country is their ally. So the answer is *YES.* ### ... ### **Passage:** Trams have operated continuously in Melbourne since 1885 (the horse tram line in Fairfield opened in 1884, but was at best an irregular service). Since then they have become a distinctive part of Melbourne's character and feature in tourism and travel advertising. Melbourne's cable tram system opened in 1885, and expanded to one of the largest in the world, with of double track. The first electric tram line opened in 1889, but closed only a few years later in 1896. In 1906 electric tram systems were opened in St Kilda and Essendon, marking the start of continuous operation of Melbourne's electric trams. **Question:** If I wanted to take a horse tram in 1884, could I look up the next tram on a schedule? **Give the rationale before answering.**
BoolQ	**Passage:** Love Island Australia is an Australian dating reality show based on the British series Love Island. The series is presented by Sophie Monk and narrated by Eoghan McDermott. The series began airing on 9Go! and 9Now on 27 May 2018. The final aired on 5 July 2018, with Grant Crapp and Tayla Damir winning and sharing the $50,000 prize money. Eden Dally and Erin Barnett finished as runners up. **Question:** is there a prize for love island australia **Give the rationale before answering.** There is a $50000 prize money. So the answer is *True.* ### … ### **Passage:** Conscription in the United States, commonly known as the draft, has been employed by the federal government of the United States in five conflicts: the American Revolution, the American Civil War, World War I, World War II, and the Cold War (including both the Korean War and the Vietnam War). The third incarnation of the draft came into being in 1940 through the Selective Training and Service Act. It was the country's first peacetime draft. From 1940 until 1973, during both peacetime and periods of conflict, men were drafted to fill vacancies in the United States Armed Forces that could not be filled through voluntary means. The draft came to an end when the United States Armed Forces moved to an all-volunteer military force. However, the Selective Service System remains in place as a contingency plan; all male civilians between the ages of 18 and 25 are required to register so that a draft can be readily resumed if needed. United States Federal Law also provides for the compulsory conscription of men between the ages of 17 and 45 and certain women for militia service pursuant to Article I, Section 8 of the United States Constitution and 10 U.S. Code § 246. **Question:** was there a draft in the revolutionary war **Give the rationale before answering.**
PERSPECTRUM	**Perspective:** Interfering with a species for cosmetic reasons is poor practice **Claim:** The breeding of white tigers in captivity should be banned. Does the perspective support or undermine the claim? Answer with: supports, undermines or not a valid perspective. **Give the rationale before answering.** Breeding tigers in captivity is interfering with the species. So the answer is *supports.* ### … ### **Perspective:** I would just be so upset that that guy cheated and stole that moment away from me. **Claim:** should cheaters be given a second chance. Does the perspective support or undermine the claim? Answer with: supports, undermines or not a valid perspective. **Give the rationale before answering.**

Table 10: Example CoT prompts used for contrast sets

Task	CoT Prompt
ROPES	**Passage:** Hemophilia is the name of a group of hereditary diseases that affect the body's ability to control blood clotting. Hemophilia is caused by a lack of clotting factors in the blood. Clotting factors are normally released by platelets. Since people with hemophilia cannot produce clots, any cut can put a person at risk of bleeding to death. The risk of internal bleeding is also increased in hemophilia, especially into muscles and joints. This disease affected the royal families of Europe. Greg and Bobby are coworkers who are currently eating lunch together on their break. They haven't known each other very long and are having a casual conversation when Bobby mentions that he is a hemophiliac. Greg, unsure of what exactly that means, looks confused and taken aback. Bobby, noticing the bewilderment on Greg's face, takes a moment to explain the details of his condition to Greg.**Question:** Does Greg or Bobby have a lower platelet count? **Give the rationale before answering.** Bobby is hemophiliac which would mean that he lacks clotting factors normally released by platelets. So the answer is *Bobby.* ### ... ### **Passage:** The theory of evolution by natural selection was proposed at about the same time by both Charles Darwin and Alfred Russel Wallace, shown in Figure below , and was set out in detail in Darwin's 1859 book On the Origin of Species . Natural selection is a process that causes heritable traits that are helpful for survival and reproduction to become more common, and harmful traits, or traits that are not helpful or advantageous for survival to become more rare in a population of organisms. This occurs because organisms with advantageous traits are more "fit" to survive in a particular environment and have "adapted" to the conditions of that environment. These individuals will have greater reproductive success than organisms less fit for survival in the environment. This will lead to an increase in the number of organisms with the advantageous trait(s) over time. Over many generations, adaptations occur through a combination of successive, small, random changes in traits, and natural selection of those variants best-suited for their environment. Natural selection is one of the cornerstones of modern biology. **Question:** The tree frogs range covers 2 different habitats, One has a lot of trees and the other a lot of tall green grass with few if any trees. Tree frogs can either be green or brown. Which color tree frog has an advantage living in trees? **Give the rationale before answering.**
DROP	**Passage:** Švitrigaila was losing his influence in the Slavic principalities and could no longer resist Poland and Sigismund. On 4 September 1437 he attempted to reconcile with Poland: he would rule the lands that still backed him , and after his death these territories would pass to the King of Poland. However, under strong protest from Sigismund, the Polish Senate declined to ratify the treaty . In 1438 Švitrigaila withdrew to Moldavia. The reign of Sigismund Kęstutaitis was brief — he was assassinated in 1440. Švitrigaila returned from exile in 1442 and ruled Lutsk until his death a decade later. Jogaila's son Casimir IV Jagiellon, born in 1426, received approval as a hereditary hospodar from Lithuania's ruling families in 1440. This event is seen by the historians Jerzy Lukowski and Hubert Zawadzki as marking the end of the succession dispute. **Question:** Who was ruler first, Sigismund of Casimir IV Jagiellon? **Give the rationale before answering.** Casimir IV Jagiellon became the ruler only after the assination of Sigismund Kęstutaitis. So the answer is *Sigismund Kęstutaitis.* ### ... ### **Passage:** The origins of al-Qaeda can be traced back to the Soviet war in Afghanistan . The United States, the United Kingdom, Saudi Arabia, Pakistan, and the People's Republic of China supported the Islamist Afghan mujahadeen guerillas against the military forces of the Soviet Union and the Democratic Republic of Afghanistan. A small number of "Afghan Arab" volunteers joined the fight against the Soviets, including Osama bin Laden, but there is no evidence they received any external assistance. In May 1996 the group World Islamic Front for Jihad Against Jews and Crusaders , sponsored by bin Laden , started forming a large base of operations in Afghanistan, where the Islamist extremist regime of the Taliban had seized power earlier in the year. In February 1998, Osama bin Laden signed a fatwā, as head of al-Qaeda, declaring war on the West and Israel. Earlier in August 1996, Bin Laden declared jihad against the United States. Later in May 1998 al-Qaeda released a video declaring war on the U.S. and the West. On 7 August 1998, al-Qaeda struck the U.S. embassies in Kenya and Tanzania, killing 224 people, including 12 Americans. In retaliation, U.S. President Bill Clinton launched Operation Infinite Reach, a bombing campaign in Sudan and Afghanistan against targets the U.S. asserted were associated with WIFJAJC, although others have questioned whether a pharmaceutical plant in Sudan was used as a chemical warfare facility. The plant produced much of the region's antimalarial drugs and around 50% of Sudan's pharmaceutical needs. The strikes failed to kill any leaders of WIFJAJC or the Taliban. Next came the 2000 millennium attack plots, which included an attempted bombing of Los Angeles International Airport. On 12 October 2000, the USS Cole bombing occurred near the port of Yemen, and 17 U.S. Navy sailors were killed. **Question:** How many years was it between when Bin Laden declared a war against the United States and when he became leader of al-Qaeda? **Give the rationale before answering.**

Table 11: Example CoT prompts used for contrast sets

Task	CoT Prompt
IMDb	**Review:** "Garde à Vue has to be seen a number of times in order to understand the sub-plots it contains. If you’re not used to french wordy films, based upon conversation and battle of wits rather than on action, don’t even try to watch it. You’ll only obtain boredom to death, and reassured opinion that french movies are not for you.<br /><br />Garde à Vue is a wordy film, essentially based upon dialogs (written by Audiard by the way)and it cruelly cuts the veil of appearances.<br /><br />Why does Maître Martineau (Serrault) prefer to be unduly accused of being a child murderer rather than telling the truth ? Because at the time of the murder he was with a 18 years old girl with which he has a 8-years sexual relation. His wife knows it, she’s jealous of it and he prefers to be executed (in 1980 in France, there was still death penalty)rather than unveiling the sole ""pure and innocent"" aspect of his pitiful life." **What is the sentiment of this review:** Positive or Negative? **Give the rationale before answering.** The reviewer describes the movie to be complex, one where the audience needs to pay close attention to the dialogues and not just visuals to understand the nuanced story. So the answer is *Positive.* ### . . . ### **Review:** May 2004, Wonderland is fairly new in the UK. Brilliant film of a brutal true story. If you know LA from the early 80’s, you will appreciate how well it is captured. The use of the elements which make up its gritty cinematic style is original, amplifying the experience and bringing the viewer very close to actually being there. The use of a disjointed ’Pulp Fiction’ style time line allows exploration of the uncertainty concerning what really happened, while the direction and performances of the cast command attention, especially Val Kilmer as John Holmes; an Oscar for sure if I were handing them out. **What is the sentiment of this review:** Positive or Negative? **Give the rationale before answering.**
MATRES	**Passage:** Dr. Barnett Slepian was *killed* in his kitchen by a sniper’s bullet last fall . Investigators said Friday they *found* a rifle buried near his home in the Buffalo suburb of Amherst. **Question:** When did the event *killed* happen in relation to the event *found*: before, after, simultaneously, or is it vague? **Give the rationale before answering.** The investigation into the killing, during which the finding took place, was started only after the killing. So the answer is *before.* ### . . . ### **Passage:** A book of memoirs and photographs from the climb by Mr Lowe , which he worked on with Dr Lewis-Jones , is due to be published in May . He *said* : " Lowe was a brilliant , kind fellow who never sought the limelight ... and 60 years on from Everest his achievements deserve wider recognition . " He was *involved* in two of the most important explorations of the 20th Century ... yet remained a humble , happy man right to the end ... an inspirational lesson to us all . " **Question:** When did the event *said* happen in relation to the event *involved*: before, after, simultaneously, or is it vague? **Give the rationale before answering.**

Table 12: Example CoT prompts used for contrast sets

Task	CoT Prompt
MCTACO	**Passage:** He succeeds James A. Taylor, who stepped down as chairman, president and chief executive in March for health reasons. How often are chief executives typically replaced? Is this the answer: every 10 weeks? **Give the rationale before answering.** Most chief executives keep their positions for years. So the answer is *no.* ### ... ### **Passage:** Roberta Adams skipped the thick how-to guide on child-custody forms and sat down at a computer at the Lamoreaux Justice Center in Orange on Wednesday. When did Roberta arrive at the Lamoreaux Justice Center? Is this the answer: 10:45 PM? **Give the rationale before answering.**
Quoref	**Passage:** In the year 1978, Gracie Bowen, a 15-year-old tomboy who lives in South Orange, New Jersey, is crazy about soccer, as are her three brothers and their former soccer star father. Although Gracie wants to join her brothers and neighbor Kyle in the nightly practices her father runs, she is discouraged by everyone except her older brother, Johnny. Johnny, Gracie and Kyle attend Columbia High School, where Johnny is the captain and star player for the varsity soccer team. After missing a shot at the end of a game, the despondent Johnny drives off with a friend's car and dies in a traffic accident. Struggling with grief, Gracie decides that she wants to replace her brother on the team. Her father does not believe that girls should play soccer, telling her she is neither tough nor talented enough. Her mother is a nurse who lacks the competitive drive of the rest of her family and fears for Gracie's safety. Her mother later tells Gracie that she would have liked to become a surgeon, but that option had not been available to her as a woman. Rejected and depressed, Gracie begins to rebel; she stops doing her schoolwork, is caught cheating on an exam, and experiments with wild and self-destructive behavior. She is finally caught by her father almost having sex with a guy she met near the docks after telling her friend, "I want to do something that I've never done before." This serves as a wake-up call for her parents, particularly her father. He quits his job to work with her on her soccer training. **Question:** What happens to the sibling that supports Gracie's interest in soccer? **Give the rationale before answering.** Garcie's older brother Johnny who was supportive of her interest in soccer died in a car accident. So the answer is *dies in a traffic accident.* ### ... ### **Passage:** Bath once had an important manufacturing sector, particularly in crane manufacture, furniture manufacture, printing, brass foundries, quarries, dye works and Plasticine manufacture, as well as many mills. Significant Bath companies included Stothert & Pitt, Bath Cabinet Makers and Bath & Portland Stone. Nowadays, manufacturing is in decline, but the city boasts strong software, publishing and service-oriented industries, being home to companies such as Future plc and London & Country mortgage brokers. The city's attraction to tourists has also led to a significant number of jobs in tourism-related industries. Important economic sectors in Bath include education and health (30,000 jobs), retail, tourism and leisure (14,000 jobs) and business and professional services (10,000 jobs). Major employers are the National Health Service, the city's two universities, and the Bath and North East Somerset Council, as well as the Ministry of Defence although a number of MOD offices formerly in Bath have recently moved to Bristol. Growing employment sectors include information and communication technologies and creative and cultural industries where Bath is one of the recognised national centres for publishing, with the magazine and digital publisher Future plc employing around 650 people. Others include Buro Happold (400) and IPL Information Processing Limited (250). The city boasts over 400 retail shops, half of which are run by independent specialist retailers, and around 100 restaurants and cafes primarily supported by tourism. **Question:** half of what are run by independent specialists? **Give the rationale before answering.**

Table 13: Example CoT prompts used for contrast sets

Task	Prompt
CondaQA	**Passage:** The end of the long-held animosity between Moscow and Beijing was marked by the visit to China by Soviet General Secretary Mikhail Gorbachev in 1989. After the 1991 demise of the Soviet Union, China's relations with Russia and the former states of the Soviet Union became more amicable as the conflicting ideologies of the two vast nations no longer stood in the way. A new round of bilateral agreements was signed during reciprocal head of state visits. As in the early 1950s with the Soviet Union, Russia has again become an important source of military technology for China, as well as for raw materials and trade. Friendly relations with Russia have been an important advantage for China, complementing its strong ties with the U.S. **Question:** Can China rely on both the US and Russia as supportive allies? **The answer is** *YES.* ### ... ### **Passage:** Trams have operated continuously in Melbourne since 1885 (the horse tram line in Fairfield opened in 1884, but was at best an irregular service). Since then they have become a distinctive part of Melbourne's character and feature in tourism and travel advertising. Melbourne's cable tram system opened in 1885, and expanded to one of the largest in the world, with of double track. The first electric tram line opened in 1889, but closed only a few years later in 1896. In 1906 electric tram systems were opened in St Kilda and Essendon, marking the start of continuous operation of Melbourne's electric trams. **Question:** If I wanted to take a horse tram in 1884, could I look up the next tram on a schedule?
BoolQ	**Passage:** Love Island Australia is an Australian dating reality show based on the British series Love Island. The series is presented by Sophie Monk and narrated by Eoghan McDermott. The series began airing on 9Go! and 9Now on 27 May 2018. The final aired on 5 July 2018, with Grant Crapp and Tayla Damir winning and sharing the $50,000 prize money. Eden Dally and Erin Barnett finished as runners up. **Question:** is there a prize for love island australia **The answer is** *True.* ### … ### **Passage:** Conscription in the United States, commonly known as the draft, has been employed by the federal government of the United States in five conflicts: the American Revolution, the American Civil War, World War I, World War II, and the Cold War (including both the Korean War and the Vietnam War). The third incarnation of the draft came into being in 1940 through the Selective Training and Service Act. It was the country's first peacetime draft. From 1940 until 1973, during both peacetime and periods of conflict, men were drafted to fill vacancies in the United States Armed Forces that could not be filled through voluntary means. The draft came to an end when the United States Armed Forces moved to an all-volunteer military force. However, the Selective Service System remains in place as a contingency plan; all male civilians between the ages of 18 and 25 are required to register so that a draft can be readily resumed if needed. United States Federal Law also provides for the compulsory conscription of men between the ages of 17 and 45 and certain women for militia service pursuant to Article I, Section 8 of the United States Constitution and 10 U.S. Code § 246. **Question:** was there a draft in the revolutionary war
PERSPECTRUM	Interfering with a species for cosmetic reasons is poor practice The breeding of white tigers in captivity should be banned. Does the perspective support or undermine the claim? Answer with: supports, undermines or not a valid perspective. **The answer is** *supports.* ### … ### I would just be so upset that that guy cheated and stole that moment away from me. should cheaters be given a second chance. Does the perspective support or undermine the claim? Answer with: supports, undermines or not a valid perspective.

Table 14: Example prompts used for contrast sets without CoT

Task	Prompt
ROPES	**Passage:** Hemophilia is the name of a group of hereditary diseases that affect the body's ability to control blood clotting. Hemophilia is caused by a lack of clotting factors in the blood. Clotting factors are normally released by platelets. Since people with hemophilia cannot produce clots, any cut can put a person at risk of bleeding to death. The risk of internal bleeding is also increased in hemophilia, especially into muscles and joints. This disease affected the royal families of Europe. Greg and Bobby are coworkers who are currently eating lunch together on their break. They haven't known each other very long and are having a casual conversation when Bobby mentions that he is a hemophiliac. Greg, unsure of what exactly that means, looks confused and taken aback. Bobby, noticing the bewilderment on Greg's face, takes a moment to explain the details of his condition to Greg. **Question:** Does Greg or Bobby have a lower platelet count? **The answer is** *Bobby.* ### ... ### **Passage:** The theory of evolution by natural selection was proposed at about the same time by both Charles Darwin and Alfred Russel Wallace, shown in Figure below , and was set out in detail in Darwin's 1859 book On the Origin of Species . Natural selection is a process that causes heritable traits that are helpful for survival and reproduction to become more common, and harmful traits, or traits that are not helpful or advantageous for survival to become more rare in a population of organisms. This occurs because organisms with advantageous traits are more "fit" to survive in a particular environment and have "adapted" to the conditions of that environment. These individuals will have greater reproductive success than organisms less fit for survival in the environment. This will lead to an increase in the number of organisms with the advantageous trait(s) over time. Over many generations, adaptations occur through a combination of successive, small, random changes in traits, and natural selection of those variants best-suited for their environment. Natural selection is one of the cornerstones of modern biology. **Question:** The tree frogs range covers 2 different habitats, One has a lot of trees and the other a lot of tall green grass with few if any trees. Tree frogs can either be green or brown. Which color tree frog has an advantage living in trees?
DROP	**Passage:** Švitrigaila was losing his influence in the Slavic principalities and could no longer resist Poland and Sigismund. On 4 September 1437 he attempted to reconcile with Poland: he would rule the lands that still backed him , and after his death these territories would pass to the King of Poland. However, under strong protest from Sigismund, the Polish Senate declined to ratify the treaty . In 1438 Švitrigaila withdrew to Moldavia. The reign of Sigismund Kęstutaitis was brief — he was assassinated in 1440. Švitrigaila returned from exile in 1442 and ruled Lutsk until his death a decade later. Jogaila's son Casimir IV Jagiellon, born in 1426, received approval as a hereditary hospodar from Lithuania's ruling families in 1440. This event is seen by the historians Jerzy Lukowski and Hubert Zawadzki as marking the end of the succession dispute. **Question:** Who was ruler first, Sigismund of Casimir IV Jagiellon? **The answer is** *Sigismund Kęstutaitis.* ### ... ### **Passage:** The origins of al-Qaeda can be traced back to the Soviet war in Afghanistan . The United States, the United Kingdom, Saudi Arabia, Pakistan, and the People's Republic of China supported the Islamist Afghan mujahadeen guerillas against the military forces of the Soviet Union and the Democratic Republic of Afghanistan. A small number of "Afghan Arab" volunteers joined the fight against the Soviets, including Osama bin Laden, but there is no evidence they received any external assistance. In May 1996 the group World Islamic Front for Jihad Against Jews and Crusaders , sponsored by bin Laden , started forming a large base of operations in Afghanistan, where the Islamist extremist regime of the Taliban had seized power earlier in the year. In February 1998, Osama bin Laden signed a fatwā, as head of al-Qaeda, declaring war on the West and Israel. Earlier in August 1996, Bin Laden declared jihad against the United States. Later in May 1998 al-Qaeda released a video declaring war on the U.S. and the West. On 7 August 1998, al-Qaeda struck the U.S. embassies in Kenya and Tanzania, killing 224 people, including 12 Americans. In retaliation, U.S. President Bill Clinton launched Operation Infinite Reach, a bombing campaign in Sudan and Afghanistan against targets the U.S. asserted were associated with WIFJAJC, although others have questioned whether a pharmaceutical plant in Sudan was used as a chemical warfare facility. The plant produced much of the region's antimalarial drugs and around 50% of Sudan's pharmaceutical needs. The strikes failed to kill any leaders of WIFJAJC or the Taliban. Next came the 2000 millennium attack plots, which included an attempted bombing of Los Angeles International Airport. On 12 October 2000, the USS Cole bombing occurred near the port of Yemen, and 17 U.S. Navy sailors were killed. **Question:** How many years was it between when Bin Laden declared a war against the United States and when he became leader of al-Qaeda?

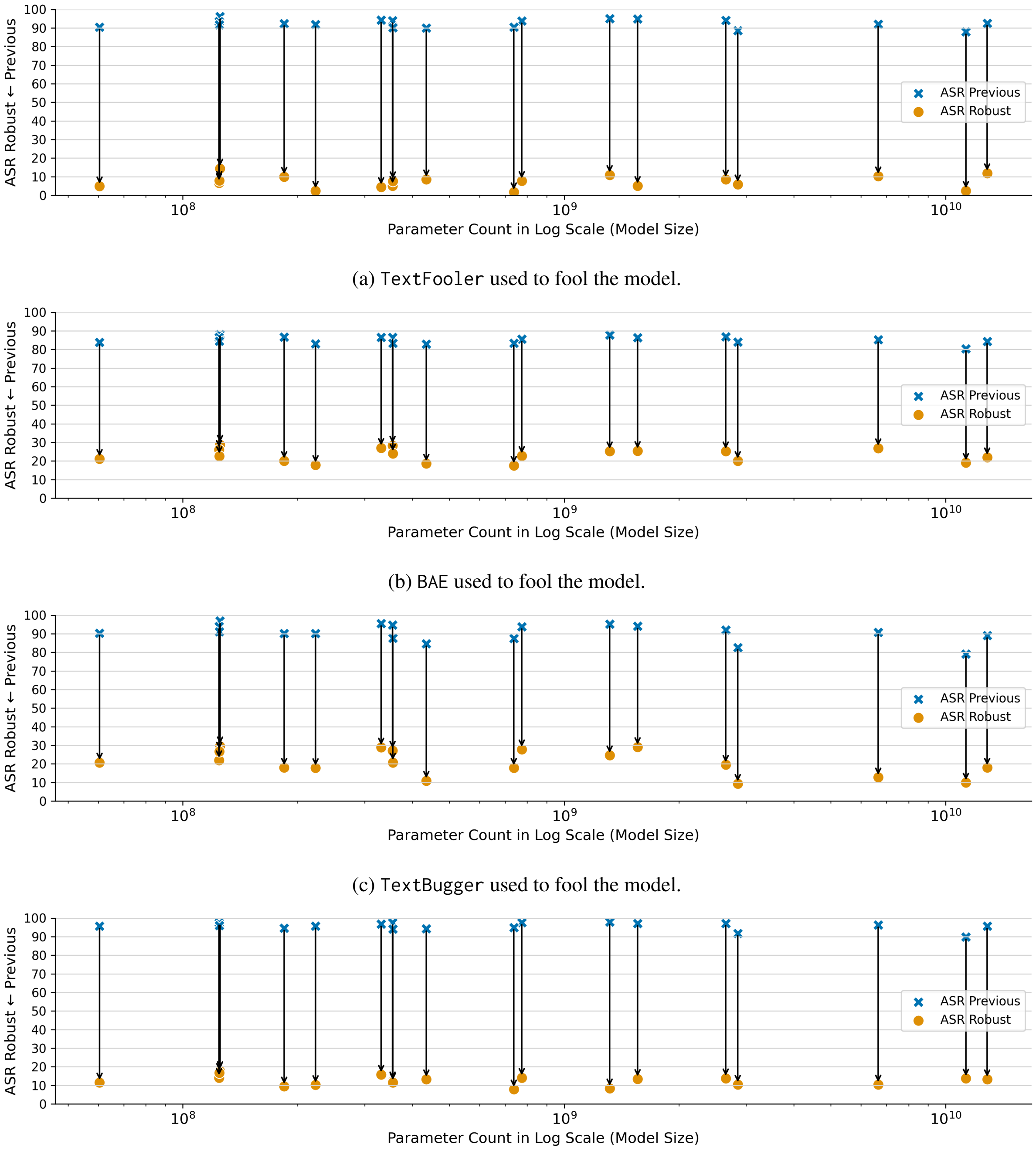
Table 15: Example prompts used for contrast sets without CoT

Task	Prompt
IMDb	**Review:** "Garde à Vue has to be seen a number of times in order to understand the sub-plots it contains. If you're not used to french wordy films, based upon conversation and battle of wits rather than on action, don't even try to watch it. You'll only obtain boredom to death, and reassured opinion that french movies are not for you.<br /><br />Garde à Vue is a wordy film, essentially based upon dialogs (written by Audiard by the way)and it cruelly cuts the veil of appearances.<br /><br />Why does Maître Martineau (Serrault) prefer to be unduly accused of being a child murderer rather than telling the truth ? Because at the time of the murder he was with a 18 years old girl with which he has a 8-years sexual relation. His wife knows it, she's jealous of it and he prefers to be executed (in 1980 in France, there was still death penalty)rather than unveiling the sole ""pure and innocent"" aspect of his pitiful life." **What is the sentiment of this review:** Positive or Negative? **The answer is** *Positive.* ### . . . ### **Review:** May 2004, Wonderland is fairly new in the UK. Brilliant film of a brutal true story. If you know LA from the early 80's, you will appreciate how well it is captured. The use of the elements which make up its gritty cinematic style is original, amplifying the experience and bringing the viewer very close to actually being there. The use of a disjointed 'Pulp Fiction' style time line allows exploration of the uncertainty concerning what really happened, while the direction and performances of the cast command attention, especially Val Kilmer as John Holmes; an Oscar for sure if I were handing them out. **What is the sentiment of this review:** Positive or Negative?
MATRES	**Passage:** Dr. Barnett Slepian was *killed* in his kitchen by a sniper's bullet last fall . Investigators said Friday they *found* a rifle buried near his home in the Buffalo suburb of Amherst. **Question:** When did the event *killed* happen in relation to the event *found*: before, after, simultaneously, or is it vague? **The answer is** *before.* ### . . . ### **Passage:** A book of memoirs and photographs from the climb by Mr Lowe , which he worked on with Dr Lewis-Jones , is due to be published in May . He *said* : " Lowe was a brilliant , kind fellow who never sought the limelight ... and 60 years on from Everest his achievements deserve wider recognition . " He was *involved* in two of the most important explorations of the 20th Century ... yet remained a humble , happy man right to the end ... an inspirational lesson to us all . " **Question:** When did the event *said* happen in relation to the event *involved*: before, after, simultaneously, or is it vague?

Table 16: Example prompts used for contrast sets without CoT

Task	Prompt
MCTACO	**Passage:** He succeeds James A. Taylor, who stepped down as chairman, president and chief executive in March for health reasons. How often are chief executives typically replaced? Is this the answer: every 10 weeks? **The answer is** *no.* ### . . . ### **Passage:** Roberta Adams skipped the thick how-to guide on child-custody forms and sat down at a computer at the Lamoreaux Justice Center in Orange on Wednesday. When did Roberta arrive at the Lamoreaux Justice Center? Is this the answer: 10:45 PM?
Quoref	**Passage:** In the year 1978, Gracie Bowen, a 15-year-old tomboy who lives in South Orange, New Jersey, is crazy about soccer, as are her three brothers and their former soccer star father. Although Gracie wants to join her brothers and neighbor Kyle in the nightly practices her father runs, she is discouraged by everyone except her older brother, Johnny. Johnny, Gracie and Kyle attend Columbia High School, where Johnny is the captain and star player for the varsity soccer team. After missing a shot at the end of a game, the despondent Johnny drives off with a friend's car and dies in a traffic accident. Struggling with grief, Gracie decides that she wants to replace her brother on the team. Her father does not believe that girls should play soccer, telling her she is neither tough nor talented enough. Her mother is a nurse who lacks the competitive drive of the rest of her family and fears for Gracie's safety. Her mother later tells Gracie that she would have liked to become a surgeon, but that option had not been available to her as a woman. Rejected and depressed, Gracie begins to rebel; she stops doing her schoolwork, is caught cheating on an exam, and experiments with wild and self-destructive behavior. She is finally caught by her father almost having sex with a guy she met near the docks after telling her friend, "I want to do something that I've never done before." This serves as a wake-up call for her parents, particularly her father. He quits his job to work with her on her soccer training. **Question:** What happens to the sibling that supports Gracie's interest in soccer? **The answer is** Garcie's older brother Johnny who was supportive of her interest in soccer died in a car accident. So the answer is *dies in a traffic accident.* ### . . . ### **Passage:** Bath once had an important manufacturing sector, particularly in crane manufacture, furniture manufacture, printing, brass foundries, quarries, dye works and Plasticine manufacture, as well as many mills. Significant Bath companies included Stothert & Pitt, Bath Cabinet Makers and Bath & Portland Stone. Nowadays, manufacturing is in decline, but the city boasts strong software, publishing and service-oriented industries, being home to companies such as Future plc and London & Country mortgage brokers. The city's attraction to tourists has also led to a significant number of jobs in tourism-related industries. Important economic sectors in Bath include education and health (30,000 jobs), retail, tourism and leisure (14,000 jobs) and business and professional services (10,000 jobs). Major employers are the National Health Service, the city's two universities, and the Bath and North East Somerset Council, as well as the Ministry of Defence although a number of MOD offices formerly in Bath have recently moved to Bristol. Growing employment sectors include information and communication technologies and creative and cultural industries where Bath is one of the recognised national centres for publishing, with the magazine and digital publisher Future plc employing around 650 people. Others include Buro Happold (400) and IPL Information Processing Limited (250). The city boasts over 400 retail shops, half of which are run by independent specialist retailers, and around 100 restaurants and cafes primarily supported by tourism. **Question:** half of what are run by independent specialists?

Table 17: Example prompts used for contrast sets without CoT



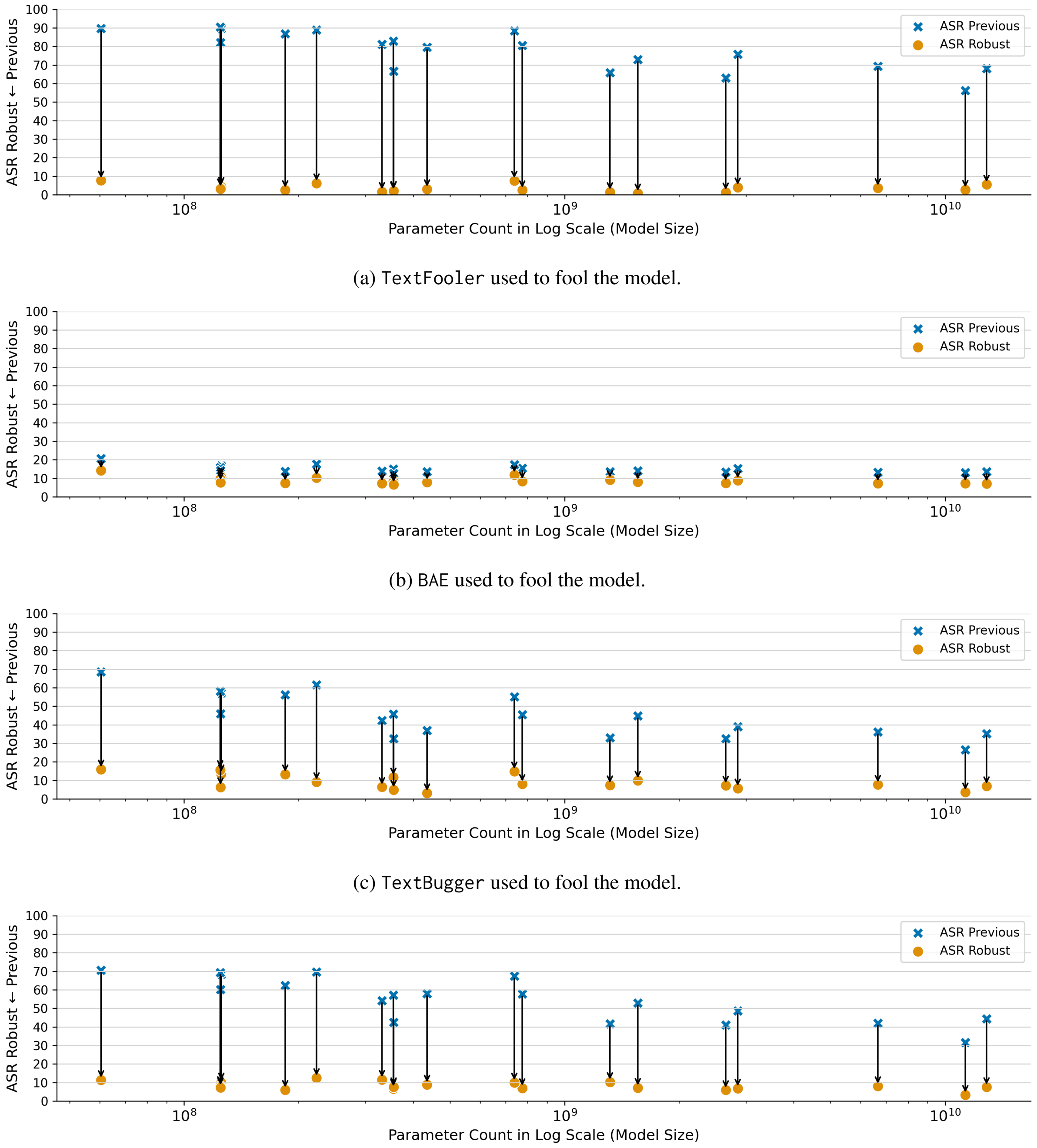
(a) `TextFooler` used to fool the model.

(b) `BAE` used to fool the model.

(c) `TextBugger` used to fool the model.

(d) `PWWS` used to fool the model.

Figure 15: The change in the attack success rate (ASR) as measured in prior work (1) vs. our robust modification (2) in the **MNLI** experimental setup. The attack type is unknown because `TextFooler` is used to train the defense.



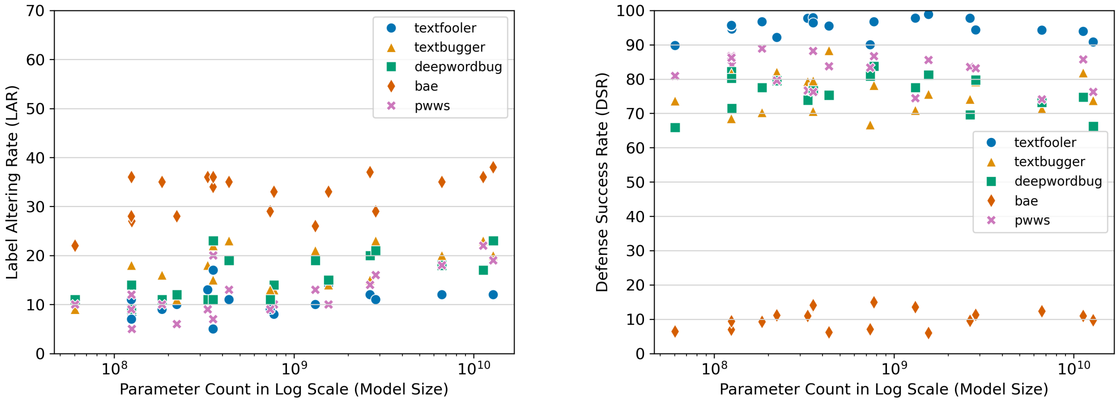
(a) `TextFooler` used to fool the model.

(b) `BAE` used to fool the model.

(c) `TextBugger` used to fool the model.

(d) `PWWS` used to fool the model.

Figure 16: The change in the attack success rate (ASR) as measured in prior work (1) vs. our robust modification (2) in the **AGNews** experimental setup. The attack type is unknown because `TextFooler` is used to train the defense.



(a) LAR

(b) DSR

Figure 17: Label altering rate (LAR; 4) and defense success rate (DSR; 5) obtained in the **AGNews** setting. Higher values of these measurements contribute to worse effectiveness of attacks, i.e., lower attack success rate.

original	**perturbed**	**model**	**model_pred**	**gpt4_pred**	**label**
Oil exports flow as strike woes ease A general strike in Nigeria, which has raised fears over oil supply from the world #39;s seventh-largest exporter, will likely end its first phase on Thursday quot;all going well quot;, union leaders said.	Oil exports flow as strike woes ease A massive strike in Nigeria, which has raised fears over oil supply from the world #39;s seventh-largest exporter, will likely end its first phase on Thursday quot;all going well quot;, union leaders said.	facebook/opt-13b	Business	Business	World
Veritas Keeps Reaching into Its Wallet By acquiring KVault, which makes e-mail-archiving software, it aims to erode EMC #39;s lead and rebuild investors #39; confidence.	Veritas Keeps Reaching into Its Wallet By acquiring ups, which makes e-mail-archiving software, it aims to erode EMC #39;s lead and rebuild investors #39; confidence.	t5-3b	Business	Business	Sci/Tech
Thailand Shows No Easy War Against Wildlife Crime (Reuters) Reuters - With an AK-47 assault rifle slung over\his shoulder, Sompong Prajobjan roamed one of Thailand’s lush\national parks for more than a decade.	Thailand Shows No Easy War Against mass Crime (Reuters) Reuters - With an AK-47 assault rifle slung over\his shoulder, Sompong Prajobjan roamed one of Thailand’s lush\national parks for more than a decade.	facebook/opt-6.7b	World	World	Sci/Tech
Google stock falls as share lockups expire SAN FRANCISCO, - Shares of Google Inc. fell as much as 6.5 percent Tuesday, as selling restrictions were lifted on 39 million shares held by employees and early investors in the newly public Web search company.	Google stock falls as share lockups expire la FRANCISCO, - Shares of Google Inc. fell as much as 6.5 percent Tuesday, as selling restrictions were lifted on 39 million shares held by employees and early investors in the newly public Web search service.	gpt2-large	Sci/Tech	Business	Business
VZ Wireless Slams National 411 Directory WASHINGTON – Verizon Wireless, the nation #39;s largest wireless carrier, clashed with other cellular carriers on Tuesday, telling a US Senate committee that a proposal for a national wireless telephone directory is a quot;terrible idea quot; and that the proposal	network Wireless Slams National 411 Directory WASHINGTON – Verizon Wireless, the nation #39;s largest wireless carrier, clashed with other cellular carriers on Tuesday, telling a US Senate committee that a proposal for a national wireless telephone directory is a quot;terrible idea quot; and that the proposal	facebook/opt-125m	Sci/Tech	Sci/Tech	Business
Douglas, Fraser, and blue await you Jim McLeod has a great day job, but a seasonal sideline is his #39; #39;tree #39; #39; calling. Throughout the year, he #39;s president and owner of a software company called InfoCode Corp.	Douglas, Fraser, and blue await you Jim McLeod has a regular day job, but a seasonal sideline is his #39; #39;special #39; #39; calling. Throughout the year, he #39;s president and owner of a software firm called InfoCode Corp.	gpt2-medium	Sci/Tech	Business	Business
Pa. Golfer Cleared of Not Yelling ’Fore’ (AP) AP - A golfer plunked in the face by an errant ball was unable to convince a jury that the man who hit him was negligent for failing to yell "Fore!"	Pa. Golfer Cleared of Not Yelling ’golf (AP) pa - A golfer plunked in the face by an errant ball was able to convince a jury that the man who assaulted him was negligent for failing to yell "golf!"	roberta-base	World	Sports	Sports
Intel lauds milestone in shrinking chips Contradicting fears that the semiconductor industry #39;s pace of development is slowing, Intel Corp has announced that it has achieved a milestone in shrinking the size of transistors that will power its next-generation chips.	Intel lauds milestone in growing chips Contradicting fears that the semiconductor industry #39;s pace of development is slowing, Intel Corp has announced that it has achieved a milestone in shrinking the size of transistors that will power its next-generation chips.	facebook/opt-1.3b	Business	Sci/Tech	Sci/Tech
Earthquake Rocks Indonesia’s Bali, One Dead BALI, Indonesia (Reuters) - A powerful earthquake rocked Indonesia’s premier tourist island of Bali Wednesday, killing one person, injuring at least two and triggering some panic, officials said.	Earthquake Rocks Indonesia’s Bali, One Dead bali, Indonesia (Reuters) - A powerful earthquake rocked Indonesia’s premier tourist island of Bali Wednesday, killing one person, injuring at least two and triggering some panic, officials said.	t5-large	Sci/Tech	World	World
Surviving Biotech’s Downturns Charly Travers offers advice on withstanding the volatility of the biotech sector.	Surviving Biotech’s Downturns Charly s offers advice on withstanding the volatility of the biotech sector.	t5-base	Sci/Tech	Business	Business

Table 18: Examples of AGNews attacked by bae

original	perturbed	model	model_pred	gpt4_pred	label
Jordan prince loses succession Jordan #39;s Prince Hamzah says he is conceding to the wish of King Abdullah II to strip him of his crown as heir to the throne. quot;I obey the command of my elder brother out of my loyalty, love	Jordan princ loses successioz JBordan #q39;s Pricne Hamzah asys he is conceidng to the wish of Kng Abdullha II to strip him of his crwn as hier to the thorne. qwot;I obZey the ommand of my elder brother out of my loyalKy, love	facebook/opt-13b	Sports	World	World
Trust Digital Gets CEO, Cash Influx Trust Digital Inc., a McLean software company, is getting a new chief executive and \$3.1 million in new investments as it tries to expand its business making security software for wireless devices. &lt;FONT face="verdana,MS Sans Serif,arial,helvetica" size="-2"\ color="#666666"&gt;&lt;B&gt;-The Washington Post&lt;/B&gt;&lt;/FONT&gt;	Trust Digiutal GetP COE, Cash Influx Trust Dgiital Inc., a McLean software cmopany, is getting a new chief executive and \$3.1 million in new investments as it ries to expand its business making security software for wireless devices. &lt;FONT face="verdana,MS Sans Serif,arial,helvetica" size="-2"\ color="#666666"&gt;&lt;B&gt;-The Washington Post&lt;/B&gt;&lt;/FONT&gt;	t5-3b	Business	Business	Sci/Tech
PeopleSoft sweetens employee compensation Business software maker PeopleSoft Friday said it was boosting compensation packages for all employees except its chief executive in a move that would raise the	PeopleSoft sweetens employee ompensation Busienss oftware maker PeoplSeoft Fridya maid it was Zboosting comensation pakcages for all employees except its cihef eGecutive in a Aove that wound raise the	facebook/opt-6.7b	Sci/Tech	Business	Business
Second Acts Former House speaker Thomas M. Finneran is the new president of the Massachusetts Biotechnology Council, a trade group that counts more than 400 members, including Genzyme Corp. and Biogen Idec Inc., the two largest biotechnology companies in the state. Its previous president left under pressure earlier this year, and some members say they chose Finneran, who quit his legislative post ...	Second Acts Former House speaker Thmoas M. Fineran is the new presqident of the Massachusemts Biotechnolgy Council, a trade group that counts more than 400 members, including Genzyme Corp. and Biogen Idec InOc., the two largest biotechnology cmpanies in the stabe. Its pDrevious rpesident left under pressure earlMer this year, and some members say they chose Finneran, who quit his legislative post ...	gpt2-large	World	Business	Business
Putin Says Russia Working on New Nuclear Systems (Reuters) Reuters - Russia is working on new nuclear missile\systems that other powers do not have in order to protect\itself against future security challenges, President Vladimir\Putin said Wednesday.	Putin Says Russia Working on New Nuclear Systems (Reuers) Reuters - Russia is working on new nuclera missile\systems that other powers do not have in order to protect\itself against future security challenges, Poresident Vladimir\Putin said Wednvesday.	facebook/opt-125m	Sci/Tech	World	World
Challenger disappoints with writedown The Kerry Packer-backed Challenger Financial Services Group has reported its first net loss since incorporating, impacted by a massive writedown of goodwill.	Challenger disappoints with writedown The KFrry Packer-babked ChallengeBr Financial eSrvices Gnoup has reporetd its first net lxss since incorporating, impacted by a massive writedown of goodwill.	gpt2-medium	Sci/Tech	Business	Business
Hewlett-Packard buys Synstar Palo Alto-based Hewlett-Packard Co. has bought IT services company Synstar plc, of Bracknell, England for about \$293.3 million. Synstar has some 1,500 customers across Europe, selling IT support for various computer platforms.	Hewlett-Packard ubys Synstar Palo Alto-based Hewlett-Packard Co. has bought IT services company Synstar plc, of Bracknell, England for about \$293.3 million. Synstar has some 1,500 customers across Europe, selling IT support for various computer platforms.	roberta-base	Sci/Tech	Business	Business
Alaska #39;s summer tourism pegged at 1.4 million visitors The number of summer visitors to Alaska rose from the year before, prompting the president of the Alaska Travel Industry Association to say tourism appeared to be back on track since leveling off after the 2001 terrorist attacks.	Alaska #39;s summer tourims pegged at 1.4 million visitors The number of summer visitors to Alaska rose from the year before, prompting the president of the Alaska TraPel IndAstry Association to say tourism appeared to be back on track since leveling off after the 2001 terrorist attacks.	facebook/opt-1.3b	Sci/Tech	World	Business
Two top Nokia executives resign HELSINKI: Nokia said the respected head of its networks unit had resigned and another top networks official left in the second major departure of top management in two weeks at the world #39;s largest mobile phone maker.	Two top Nokia executives resign HELSINKI: Nokia said the respected head of its networks unit had resigned and another top networks official left in the second major dparture of top management in two weeks at the world #39;s largest mobile phone aker.	t5-large	Sci/Tech	Business	Business
ATA: Customers Won #39;t Be Affected By Bankruptcy INDIANAPOLIS – ATA says it will honor all tickets and maintain its full schedule, after filing for Chapter 11 bankruptcy Tuesday.	ATA: Cusomers Won #39;t Be Affected By Bankruptcy INDIANAPOLIS – ATEA says it will honor all tickets and maintain its full scheduel, after filing for hCapter k1 bankruptcy Tuesday.	t5-base	Sports	Business	Business

Table 19: Examples of AGNews attacked by deepwordbug

original	**perturbed**	**model**	**model_pred**	**gpt4_pred**	**label**
Sun looks for props with new server, storage hardware Sun Microsystems will hold its quarterly product launch this week, unleashing a raft of new hardware offerings spanning servers to storage.	Sun search for property with new server, store hardware Sun Microsystems will admit its quarterly product launch this week, unleashing a passel of newfangled hardware offerings spanning servers to memory.	facebook/opt-13b	Business	Sci/Tech	Sci/Tech
Caymas to open with security gateways Security start-up Caymas Systems launches Monday with products to protect the flow of corporate data.	Caymas to clear with surety gateways protection start-up Caymas organisation plunge Monday with merchandise to protect the menstruate of corporal data.	t5-3b	Business	Business	Sci/Tech
U.S. Asks Laos to Check Massacre Report (AP) AP - The State Department said Monday it is taking seriously allegations that Laotian military forces may have massacred children of the country's Hmong ethnic minority.	U.entropy. Asks Laos to learn slaughter study (AP) AP - The commonwealth Department said Monday it is consider seriously allegations that Laotian military force-out may have massacred tyke of the country's Hmong pagan minority.	facebook/opt-6.7b	Sci/Tech	World	World
EU Move on Cyprus Eases Way for Turkey Deal BRUSSELS (Reuters) - The European Union and Turkey inched toward a historic agreement on starting membership talks on Friday as EU leaders softened their demands on the crucial sticking point of Cyprus.	EU Move on Cyprus Eases Way for Turkey Deal BRUSSELS (Reuters) - The European Union and Turkey inched toward a historic agreement on starting membership talks on Friday as europium leaders softened their demands on the crucial sticking point of Cyprus.	gpt2-large	Business	World	World
Smith setback for Windies West Indies have been forced to make a second change to their Champions Trophy squad because of injury. Dwayne Smith is suffering from a shoulder problem and has been replaced by Ryan Hinds.	metalworker blow for Windies Occident indie have been thrust to puddle a 2nd commute to their fighter prize team because of trauma. Dwayne metalworker is brook from a berm job and has been replaced by Ryan hind.	facebook/opt-125m	Sci/Tech	Sports	Sports
Hewlett-Packard buys Synstar Palo Alto-based Hewlett-Packard Co. has bought IT services company Synstar plc, of Bracknell, England for about \$293.3 million. Synstar has some 1,500 customers across Europe, selling IT support for various computer platforms.	Hewlett-Packard buys Synstar Palo Alto-based Hewlett-Packard Co. has corrupt IT services society Synstar plc, of Bracknell, England for about \$293.3 million. Synstar has some 1,500 customers across Europe, selling IT support for various computer platforms.	gpt2-medium	Sci/Tech	Business	Business
Liu brings China 4th gold in weightlifting at Athens Games ATHENS, Aug. 19 (Xinhuanet) – Chinese Hercules Liu Chunhong Thursday lifted three world records on her way to winning the women #39;s 69kg gold medal at the Athens Olympics, the fourth of the power sport competition for China.	Liu brings chinaware quaternary amber in weightlifting at Athens Games ATHENS, Aug. 19 (Xinhuanet) – Chinese Hercules Liu Chunhong Thursday lifted 3 domain records on her way to winning the womanhood #39;s 69kg Au medal at the Athens Olympics, the quarter of the tycoon athletics competition for chinaware.	roberta-base	Sci/Tech	Sports	Sports
Gibbs doubts Skins were tipping plays If the Cleveland Browns knew what plays the Washington Redskins were going to run before the ball was snapped Sunday, a review of the game tape 24 hours later revealed scant evidence of it.	Gibbs dubiety tegument were tap gaming If the Cleveland embrown knew what gambling the WA Injun were going to operate before the testis was snap Sunday, a inspection of the crippled videotape 24 hours belated revealed light evidence of it.	facebook/opt-1.3b	Business	Sports	Sports
England's Lawyers Try to Get Photos Thrown Out Lawyers for Pfc. Lynndie R. England sought Wednesday to throw out evidence at the heart of the Abu Ghraib prison scandal – the now-infamous photos showing her smiling and pointing at naked Iraqi detainees.	England's Lawyers taste to stimulate Photos project KO'd Lawyers for perfluorocarbon. Lynndie R. England sought Wed to throw out evidence at the kernel of the Abu Ghraib prison scandal – the now-infamous photos showing her smiling and pointing at naked Iraqi detainees.	t5-large	Sci/Tech	World	World
MLB: NY Yankees 7, Minnesota 6 (12 inn.) Hideki Matsui drove in Derek Jeter with a 12th-inning sacrifice fly Wednesday night, giving the New York Yankees a dramatic, 7-6 win over Minnesota.	MLB: NY Yankees heptad, Minnesota 6 (12 inn.) Hideki Matsui get in Derek Jeter with a 12th-inning sacrifice fly Wed Nox, collapse the New York Yankees a dramatic, 7-6 profits over Minnesota.	t5-base	World	Sports	Sports

Table 20: Examples of AGNews attacked by pwws

original	perturbed	model	model_pred	gpt4_pred	label
Panama flooding kills nine people At least nine people - seven of them children - have died in flooding in the capital of Panama. The authorities say at least 13 people are still missing after heavy rainfall caused rivers to break their banks.	Panama flooding murder nine people At least ninth citizens - seventh of them kid - have die in floods in the capital of Panama. De authority sy at laest 13 people are again lacking after he avy rainall caused stream to rupture their banque.	facebook/opt-13b	Business	World	World
Hit TV series 24 goes from small screen to smaller screen (AFP) AFP - The hit US television show quot;24 quot; is going from the small screen to the smaller after 20th Century Fox and Vodaphone struck a groundbreaking deal to distribute the drama on mobile telephones.	Hit TV series 24 goes from small screen to smaller screen (AFP) AFP - The hit US television show quot;24 quot; is going from the small screen to the smaller after 20th Century Fox and Vodaphone struck a groundbreaking deal to distribute the drama on moible telephones.	t5-3b	World	Sci/Tech	Sci/Tech
Cali Cartel Boss Sent to U.S. on Drug Charges BOGOTA, Colombia (Reuters) - The former boss of the Cali drug cartel, who once controlled most of the world's cocaine trade, was sent to the United States on Friday to face trafficking and money laundering charges.	Cali Car tel Boss Sent to wu.S. on Drug Charges BOGOTA, Colombia (Reuters) - The former boss of the Ca li drug cartel, who once controlled most of the w orld's cocaine trade, was sent to the United States on Friday to ace trafficking and money laundering charges.	facebook/opt-6.7b	Business	World	World
BOND REPORT CHICAGO (CBS.MW) - Treasurys remained solidly lower Wednesday in the wake of election results that had President Bush ahead of Democratic challenger John Kerry.	BONDS APPRISE CHICAGO (CAS.TURBINES) - Treasurys remained solidly lwoer Wednesday in the wake of election results that had President Bush ahead of Democratic challenger John Kerry.	gpt2-large	World	Business	Business
Iran shuts reformist websites WEBSITES CLOSE to Iran #39;s leading reformist party have been blocked by religious hardliners in the police bureau of public morals.	Iran shuts reformist sites WEBSITES CLOSE to Iran #39;s leading reformist party have been blocked by religious hardliners in the police bureau of public morals.	facebook/opt-125m	World	World	Sci/Tech
Two top Nokia executives resign HELSINKI: Nokia said the respected head of its networks unit had resigned and another top networks official left in the second major departure of top management in two weeks at the world #39;s largest mobile phone maker.	Two top Nokia executives resign HELSINKI: Nokia said the respected head of its networks unity had resigned and another top networks official left in the second major departure of top management in two weeks at the monde #39;s largest mobile phone maker.	gpt2-medium	Sci/Tech	Business	Business
Two top Nokia executives resign HELSINKI: Nokia said the respected head of its networks unit had resigned and another top networks official left in the second major departure of top management in two weeks at the world #39;s largest mobile phone maker.	Two top Nokia executives resign HELSINKI: Nokia said the respected head of its networks drives had resigned and another top networks official left in the second major departure of supremo management in two weeks at the world #39;s largest mobile phone maker.	roberta-base	Sci/Tech	Business	Business
Scandal won #39;t go away ATHENS – It was telling yesterday that the majority of the dozens of journalists who asked questions and attended a news conference into a Greek doping scandal were mostly Canadian. Question after question came from Canadians. We were all there, I think, ...	Scandal won #39;t go away ATHENS – It was telling yesterday that the majority of the dozens of journalists who asked questions and attended a news conference into a Greek fortification scandal were principally Population. Questions after ques tion cmae from Canadian. We were all there, I think, ...	facebook/opt-1.3b	World	Sports	Sports
Plane of NASCAR Team Hendrick Missing (AP) AP - A plane carrying members of the Hendrick Motorsports organization was missing Sunday after losing contact with the Federal Aviation Administration on its way to a NASCAR race, and a search was underway for the aircraft.	Plane of NASCAR Team Hendrick Missing (AP) AP - A plane carrying members of the Hendrick Motorsports organization was missing Sunday after losing contact with the Federal Aviation Administration on its way to a NASAR race, and a search was underway for the aircraft.	t5-large	Sci/Tech	Sports	Sports
Paula #39;s going for gold PAULA RADCLIFFE has decided she WILL run in tonight #39;s 10,000m race at the Athens Olympics. Today #39;s dramatic decision comes just days after Britain #39;s star long-distance runner was left weeping at the roadside after pulling up in the Olympic marathon.	Paula #3;s going for golden PAOLA RADCLIFFE has deicded she WILL r un in to night #39;s 1O,000m race at the Athens Olympics. Today #39;s dramatic decision comes just days after Britain #39;s star long-distance runner was left weeping at the brooch after pulling up in the Olympic marathon.	t5-base	Sci/Tech	Sports	Sports

Table 21: Examples of AGNews attacked by Textbugger

original	**perturbed**	**model**	**model_pred**	**gpt4_pred**	**label**
Bush enforces new US diplomacy (AFP) AFP - President George W. Bush has rewritten US foreign policy during four years at the White House, with the war on terror now taking priority and doubt cast on some traditional alliances.	Bush enforces roman AMERICANS extraterrestrials (AFP) AFP - President George W. Buch enjoys rewritten US spacemen policy during four annum at the Blanco Household, with the war on fear now assuming supremacy and duda colada on some usual partner.	facebook/opt-13b	Sci/Tech	World	World
Dependent species risk extinction The global extinction crisis is worse than thought, because thousands of quot;affiliated quot; species also at risk do not figure in calculations.	Subordinated species risk extinction The global extinction crisis is worse than thought, for thousands of quot;associated quot; species also at menace fait not figure in calculations.	t5-3b	World	Sci/Tech	Sci/Tech
Heineken Profit Dips But Repeats Outlook AMSTERDAM (Reuters) - Dutch brewer Heineken posted a 4.5 percent fall in core profit for the first half on Wednesday, at the low end of expectations as a weak dollar and sluggish markets hurt business.	Heineken Profit Dips But Repeated Predictions COPENHAGEN (Newsday) - Bassi blackbird Heineken seconded a 4.5 percent dip in keys winnings for the first half on Sunday, at the tenuous termination of aspirations as a insufficient money and apathetic marketplace defaced business.	facebook/opt-6.7b	Sports	Business	Business
MLB, Va. Officials Meet Chicago White Sox owner Jerry Reinsdorf led a team of negotiators from Major League Baseball in a three-hour meeting Wednesday with the leaders of the Virginia Baseball Stadium Authority.	MLB, Does. Officials Meet Boston Pai Sox owner Jerry Reinsdorf boosted a computers of negotiators from Major Society Hardball in a three-hour assembly Wednesday with the fuhrer of the Ginny Mitt Stadium Authority.	gpt2-large	Business	Sports	Sports
Check Boeing sops, Airbustells US Check Boeing sops, Airbustells &lt;b&gt;...&lt;/b&gt; European aircraft maker Airbus on Thursday criticised a US move to take a fight about subsidies to the World Trade Organisation (WTO), saying it showed its rivals unwillingness to address its own subsidies.	Check Boeing sops, Airbustells V Controls Boeing sops, Airbustells &lt;b&gt;...&lt;/b&gt; European airforce setters Airliner on Jue critic a V resettled to bears a faces about financed to the Globo Negotiation Arranged (WTO), tells it found its hopefuls unwilling to tackled its distinctive awards.	facebook/opt-125m	World	Business	Business
Stocks Open Lower; Inflation Data Weighs NEW YORK (Reuters) - U.S. stocks opened lower on Tuesday after a government report showing a much larger-than-expected rise in U.S. producer prices in October raised inflation concerns.	Arsenals Open Lower; Blowing Data Peso NOUVEAU BRONX (Reuters) - wu.ies. stocks opener least on Tuesday after a government communique depicting a much larger-than-expected escalating in oder.ies. farmers expenditure in Janeiro referred inflation implicated.	gpt2-medium	Sports	Business	Business
Stocks Open Lower; Inflation Data Weighs NEW YORK (Reuters) - U.S. stocks opened lower on Tuesday after a government report showing a much larger-than-expected rise in U.S. producer prices in October raised inflation concerns.	Stockpiles Open Lower; Inflation Data Weighs NEW YORK (Reuters) - ni.r. sharing unblocked lower on Sonntag after a government communique showcases a much larger-than-expected rise in hu.p. ranchers prices in October raised inflation feared.	roberta-base	World	Business	Business
Dreams of perfect market are fine, as long as they don't come true In these times of financial wrongdoing and subsequent systemic changes, it's only natural to wonder what a perfect investment world would look like.	Reve of perfect market are awesome, as longer as they don't entered realistic Under these times of monetary malfunction and subsequent systemic modifications, it's only natural to wonder what a perfect capital world would visualise like.	facebook/opt-1.3b	Sci/Tech	Business	Business
Liu brings China 4th gold in weightlifting at Athens Games ATHENS, Aug. 19 (Xinhuanet) – Chinese Hercules Liu Chunhong Thursday lifted three world records on her way to winning the women #39;s 69kg gold medal at the Athens Olympics, the fourth of the power sport competition for China.	Liu establishes China 4th kim in gymnastics at Athens Games GRECO, Jun. 19 (Xinhuanet) – Chinese Hercules Liang Chunhong Hoy lifted three world records on her arteries to attaining the feminine #39;s 69kg gold decoration at the Poseidon Olympics, the fourth of the electricity sport hostilities for China.	t5-large	World	Sports	Sports
Packers lose Flanagan for the season Green Bay Packers Pro Bowl center Mike Flanagan will undergo surgery on his left knee and miss the rest of the season. Coach Mike Sherman made the announcement after practice Friday, meaning for the second	Slaughtering waste Mcgrath for the season Environmentalist Golfo Packers Pro Goblet clinics Geraldo Conner hope suffer surgeons on his leave hips and fails the rest of the seasons. Buses Michaela Sherman realised the publicity after reality Hoy, signify for the second	t5-base	World	Sports	Sports

Table 22: Examples of AGNews attacked by TextFooler

original	**perturbed**	**model**	**model_pred**	**gpt4_pred**	**label**
Premise: Lifetime Extension of SCR De-NOx Catalysts Using SCR-Tech’s High Efficiency Ultrasonic Regeneration Process *Hypothesis:* There is a 10 year extension of SCR De-NOx catalysts.	*Premise:* ongoing Extension of SCR De-NOx Catalysts Using SCR-Tech’s High Efficiency Ultrasonic Regeneration Process *Hypothesis:* There is a 10 year extension of SCR De-NOx catalysts.	gpt2	Neutral	Neutral	Contradiction
Premise: Vrenna looked it and smiled. *Hypothesis:* Vreanna wore a pleased expression when she saw it.	*Premise:* Vrenna looked it and shuddered. *Hypothesis:* Vreanna wore a pleased expression when she saw it.	microsoft/deberta-v3-base	Contradiction	Contradiction	Entailment
Premise: There is a good restaurant in the village, in addition to a well-stocked mini-market for self-catering visitors. *Hypothesis:* The village has nowhere to dine.	*Premise:* There is a good restaurant in the village, in addition to a well-stocked mini-market for self-catering visitors. *Hypothesis:* another village has nowhere to dine.	t5-small	Neutral	Contradiction	Contradiction
Premise: Several of the organizations had professional and administrative staffs that provided analytical capabilities and facilitated their members’ participation in the organization’s activities. *Hypothesis:* Organizations didn’t care about members’ participation.	*Premise:* Several of the organizations had professional and administrative staffs that provided analytical capabilities and facilitated their members’ participation in the organization’s activities. *Hypothesis:* others didn’t care about members’ participation.	facebook/opt-350m	Neutral	Contradiction	Contradiction
Premise: He appropriated for the State much of the personal fortunes of the princes, but found it harder to curtail the power of land-owners who had extensive contacts with the more conservative elements in his Congress Party. *Hypothesis:* He was able to take much of the princes’ individual fortunes for the State, but it was more difficult to wrest power from land owners in contact with the conservative elements of the Congress Party.	*Premise:* He appropriated for the country much of the personal fortunes of the princes, but found it harder to curtail the power of owners who had extensive contacts with the more conservative elements in his congress country. *Hypothesis:* He was able to take much of the princes’ individual fortunes for the State, but it was more difficult to wrest power from land owners in contact with the conservative elements of the Congress Party.	t5-small	Neutral	Entailment	Entailment
Premise: But when he was persuaded by divers means to help us, he gave up after one week, declaring it beyond his powers. *Hypothesis:* He decided it was too difficult because he was distracted by other topics.	*Premise:* But when he was persuaded by divers means to help us, he gave up after one week, declaring it beyond his powers. *Hypothesis:* He decided it was too easy because he was distracted by other topics.	microsoft/deberta-v3-large	Contradiction	Contradiction	Neutral
Premise: Most of the dances are suggestive of ancient courtship rituals, with the man being forceful and arrogant, the woman shyly flirtatious. *Hypothesis:* Majority of the dances are influenced by hip hop.	*Premise:* many of the dances are suggestive of ancient courtship rituals, with the man being forceful and arrogant, the woman shyly flirtatious. *Hypothesis:* Majority of the dances are influenced by hip hop.	gpt2	Neutral	Contradiction	Contradiction
Premise: exactly and when i’m sitting here on the sofa cross-stitching and all of a sudden somebody a man’s got their hand on my door knob it’s like uh like oh no and so i don’t i don’t like that and i guess the only way to prevent it would be just to pass a city ordinance to prevent that or *Hypothesis:* I don’t ever sit on my sofa and do cross-stitch.	*Premise:* exactly and when get sitting here on the sofa cross-stitching and all of a sudden somebody a man’s got their hand on my door knob it’s like uh like oh no and so i don’t i don’t like that and i guess the only way to prevent it would be just to pass a city ordinance to prevent that or *Hypothesis:* I don’t ever sit on my sofa and do cross-stitch.	facebook/opt-2.7b	Neutral	Contradiction	Contradiction
Premise: Do not talk. *Hypothesis:* Don’t speak until they all leave.	*Premise:* Do not talk. *Hypothesis:* please speak until they all leave.	gpt2-medium	Contradiction	Contradiction	Neutral
Premise: The University of Nevada-Las Vegas boasts a student population over 23,000 (though, like most of the people in Las Vegas, they are commuters). *Hypothesis:* Most of the students of The University of Nevada are commuters.	*Premise:* The University of Nevada-Las Vegas boasts a student population over 23,000 (though, like most of the people in Las Vegas, they are foreigners). *Hypothesis:* Most of the students of The University of Nevada are commuters.	facebook/opt-6.7b	Neutral	Neutral	Entailment

Table 23: Examples of MNLI attacked by bae

original	perturbed	model	model_pred	gpt4_pred	label
Premise: The draft treaty was Tommy's bait. *Hypothesis:* Tommy took the bait of the treaty.	*Premise:* The draft treaty was Tommy's bait. *Hypothesis:* Tommy took the bjit of the treaty.	facebook/opt-13b	Contradiction	Contradiction	Neutral
Premise: The anthropologist Napoleon Chagnon has shown that Yanomamo men who have killed other men have more wives and more offspring than average guys. *Hypothesis:* Yanomamo men who kill other men have better chances at getting more wives.	*Premise:* The anthropologist Napoleon Chagnon has shown that YEnomamo men who have killed other men have more wies and more offspring than average guys. *Hypothesis:* Yanomamo men who kill other men have bettler chances at gzetting more wives.	t5-3b	Neutral	Entailment	Entailment
Premise: The providers worked with the newly created Legal Assistance to the Disadvantaged Committee of the Minnesota State Bar Association (MSBA) to create the Minnesota Legal Services Coalition State Support Center and the position of Director of Volunteer Legal Services, now the Access to Justice Director at the Minnesota State Bar Association. *Hypothesis:* The Access to Justice Director was formerly called the Director of Volunteer Legal Services.	*Premise:* The providers worked with the newly created Legal Assistance to the Disadvantaged Committee of the Minnesota State Bar Association (MSBA) to create the Minnesota Legal Services Coalition State Support Center and the position of Director of Voluteer Legal Services, now the Access to Justice Director at the Minnesota State Bar Association. *Hypothesis:* The Access to Justice Director was formerly called the Director of Volunteer Legal Services.	facebook/opt-6.7b	Neutral	Entailment	Entailment
Premise: Tax purists would argue that the value of the homemakers' hard work–and the intrafamily benefits they presumably receive in return for it–should, in fact, be treated as income and taxed, just like the wages paid to outside service providers such as baby sitters and housekeepers. *Hypothesis:* To tax purists, the value of the homemakers' hard work should not be taxed.	*Premise:* Tax purists would argue that the value of the Khomemakers' hard owrk–and the intrafamily benefits they presumably receive in return for it–shZuld, in fact, be treated as income and txed, just like the wages paid to outside service providers such as baby sitters and housekeepers. *Hypothesis:* o Sax uprists, the Hvalue of the homemiakers' hard worO should not be taxeOd.	gpt2-large	Neutral	Contradiction	Contradiction
Premise: Also, the tobacco executives who told Congress they didn't consider nicotine addictive might now be prosecuted for fraud and perjury. *Hypothesis:* Some tobacco executives told Congress nicotine is not addictive.	*Premise:* Also, the tobacco executives who told Congress they didn't consider nicotine addictive might now be prosecuted for fraud and perjury. *Hypothesis:* SoEme tobacco executives told Congress nicotine is not addictiAve.	facebook/opt-125m	Contradiction	Entailment	Entailment
Premise: 4 million, or about 8 percent of total expenditures for the two programs). *Hypothesis:* The figure of 4 million is likely to rise in the coming years.	*Premise:* 4 mlilion, or about 8 percent of total expenditures for the two programs). *Hypothesis:* The figure of 4 million is likelo to rsie in the coming yeavs.	gpt2-medium	Contradiction	Neutral	Neutral
Premise: Although claims data provide the most accurate information about health care use, ensuring adequate follow-up for purposes of obtaining information from patient self-report is important because many people do not report alcohol-related events to insurance compa-nies. *Hypothesis:* Patients naturally always report to insurance companies when health problems may be a direct result of alcohol.	*Premise:* Although claims data provide the most accurate information about health care use, ensuring adequate follow-up for purposes of obtaining information from patient self-report is important because many people do not report alohol-related events to insurance compa-nies. *Hypothesis:* Patients naturally always report to insurance companies when health problems may be a direct result of alcohol.	roberta-base	Neutral	Contradiction	Contradiction
Premise: Today the strait is busy with commercial shipping, ferries, and fishing boats, and its wooded shores are lined with pretty fishing villages, old Ottoman mansions, and the villas of Istanbul's wealthier citizens. *Hypothesis:* Today, the strait is empty after a huge sand storm killed everyone there.	*Premise:* Today the strati is busy with commercial shipping, ferries, and fishing boats, and its wooded shores are lined with pretty fishing villages, old Ottoman mansions, and the villas of Istanbul's wealthier citizens. *Hypothesis:* Today, the strait is mpty after a huge sand storm kliled everyone there.	facebook/opt-1.3b	Neutral	Contradiction	Contradiction
Premise: I have a situation. *Hypothesis:* Everything is fine and I have nothing on my mind.	*Premise:* I have a situation. *Hypothesis:* EvCrything is fien and I have notying on my minkd.	t5-large	Neutral	Contradiction	Contradiction
Premise: She had thrown away her cloak and tied her hair back into a topknot to keep it out of the way. *Hypothesis:* She shaved her head.	*Premise:* She had thrown away her cloak and zied her hair back into a topknot to keep it out of the way. *Hypothesis:* She shavfd her head.	t5-base	Entailment	Contradiction	Contradiction

Table 24: Examples of MNLI attacked by deepwordbug

original	perturbed	model	model_pred	gpt4_pred	label
Premise: This is an excerpt from the voice-over credo read in the opening credits for the new UPN series Star Pitiful Helpless Giant , starring former Secretary of State George Shultz. *Hypothesis:* Star Pitiful Helpless Giant is a show on WGN.	*Premise:* This is an excerpt from the voice-over credo read in the initiative acknowledgment for the unexampled UPN series Star Pitiful Helpless Giant , starring former Secretary of State George Shultz. *Hypothesis:* Star Pitiful Helpless Giant is a show on WGN.	facebook/opt-13b	Neutral	Neutral	Contradiction
Premise: The game of billiards is also hot. *Hypothesis:* People hate playing billiards.	*Premise:* The game of billiards is too blistering. *Hypothesis:* People hate playing billiards.	t5-3b	Neutral	Neutral	Contradiction
Premise: Several of the organizations had professional and administrative staffs that provided analytical capabilities and facilitated their members' participation in the organization's activities. *Hypothesis:* Organizations didn't care about members' participation.	*Premise:* Several of the organizations had professional and administrative staffs that provided analytical capabilities and alleviate their members' participation in the organization's activities. *Hypothesis:* Organizations didn't like about members' participation.	facebook/opt-6.7b	Neutral	Contradiction	Contradiction
Premise: Tax purists would argue that the value of the homemakers' hard work–and the intrafamily benefits they presumably receive in return for it–should, in fact, be treated as income and taxed, just like the wages paid to outside service providers such as baby sitters and housekeepers. *Hypothesis:* To tax purists, the value of the homemakers' hard work should not be taxed.	*Premise:* Tax purists would argue that the value of the homemakers' hard work–and the intrafamily benefits they presumably receive in return for it–should, in fact, be treated as income and task, just like the wages paid to outside service providers such as baby sitters and housekeepers. *Hypothesis:* To tax purists, the value of the homemakers' strong workplace should not be taxed.	gpt2-large	Neutral	Contradiction	Contradiction
Premise: Well, we've just got to get down to it, that's all. *Hypothesis:* We should take a break from this.	*Premise:* Well, we've just got to dumbfound down to it, that's all. *Hypothesis:* We should take a break from this.	facebook/opt-125m	Neutral	Neutral	Contradiction
Premise: exactly and when i'm sitting here on the sofa cross-stitching and all of a sudden somebody a man's got their hand on my door knob it's like uh like oh no and so i don't i don't like that and i guess the only way to prevent it would be just to pass a city ordinance to prevent that or *Hypothesis:* I don't ever sit on my sofa and do cross-stitch.	*Premise:* exactly and when i'm sitting here on the sofa cross-stitching and all of a sudden somebody a man's got their hand on my door knob it's like uh like oh no and so i don't i don't like that and i guess the only way to prevent it would be just to pass a city ordinance to prevent that or *Hypothesis:* I don't always sit on my sofa and do cross-stitch.	gpt2-medium	Neutral	Neutral	Contradiction
Premise: Such multicolored reef dwellers as the parrotfish and French angelfish, along with weirdly shaped coral, crawfish, or turtles hiding in crevices, can be yours for the viewing in these clear waters where visibility of 30 m (100 ft) is common. *Hypothesis:* Since the reef has been bleached and rendered lifeless by global warming, it's no longer possible to see sea life in the cloudy water.	*Premise:* Such multicolored reef dwellers as the parrotfish and French angelfish, along with weirdly shaped coral, crawfish, or turtles hiding in crevices, can be yours for the viewing in these unclouded waters where visibility of 30 m (100 ft) is uncouth. *Hypothesis:* Since the Rand has been bleached and rendered lifeless by global warming, it's no long potential to see sea sprightliness in the cloudy H2O.	roberta-base	Neutral	Contradiction	Contradiction
Premise: Homes or businesses not located on one of these roads must place a mail receptacle along the route traveled. *Hypothesis:* The other roads are far too rural to provide mail service to.	*Premise:* Homes or businesses not located on unity of these roads moldiness place a mail receptacle along the route traveled. *Hypothesis:* The other roads are ALIR too rural to allow mail service to.	facebook/opt-1.3b	Contradiction	Contradiction	Neutral
Premise: and that you're very much right but the jury may or may not see it that way so you get a little anticipate you know anxious there and go well you know *Hypothesis:* Even if you're correct, I think the jury would pick up on that.	*Premise:* and that you're very much right but the jury may or may not escort it that way so you get a little anticipate you know anxious there and go well you know *Hypothesis:* Even if you're correct, I think the jury would pick up on that.	t5-large	Neutral	Neutral	Contradiction
Premise: The technical how-tos for these three strategies will be summarized later in this paper. *Hypothesis:* There are seven strategies discussed in the paper.	*Premise:* The technical how-tos for these leash strategies will be summarized later in this paper. *Hypothesis:* There are seven strategies discussed in the paper.	t5-base	Neutral	Neutral	Contradiction

Table 25: Examples of MNLI attacked by pwws

original	perturbed	model	model_pred	gpt4_pred	label
Premise: It is the official solution, Liq. *Hypothesis:* This is officially the solution.	*Premise:* Him is the functionary solution, Liq. *Hypothesis:* This is officially the solution.	facebook/opt-350m	Neutral	Neutral	Entailment
Premise: Calcutta seems to be the only other production center having any pretensions to artistic creativity at all, but ironically you're actually more likely to see the works of Satyajit Ray or Mrinal Sen shown in Europe or North America than in India itself. *Hypothesis:* Most of Mrinal Sen's work can be found in European collections.	*Premise:* Calcutta seems to be the only other production center having any pretensions to artistic creativity at all, but ironically you're actually more likely to see the works of Satyajit Ray or Mrinal Sen shown in Europe or North America than in India itself. *Hypothesis:* Msot of Mrinal Sen's work can be found in European collections.	roberta-large	Entailment	Neutral	Neutral
Premise: It is at the moment of maximum audience susceptibility that we hear, for the first time, that the woman was fired not because of her gender but because of her sexual preference. *Hypothesis:* The woman was fired not for her sexual preference but purely on the basis of gender.	*Premise:* It is at the moment of maximum audience susceptibility that we hear, for the first time, that the woman was fired not because of her gneder but because of her sex preference. *Hypothesis:* The woman was triggered not for her sexual preference but purely on the basis of sexual.	t5-11b	Entailment	Contradiction	Contradiction
Premise: It is the official solution, Liq. *Hypothesis:* This is officially the solution.	*Premise:* It is the official solution, Liq. *Hypothesis:* This is officially the solution.	gpt2-large	Contradiction	Entailment	Entailment
Premise: Like Arabs and Jews, Diamond warns, Koreans and Japanese are joined by blood yet locked in traditional enmity. *Hypothesis:* Koreans and Japanese have no tension between them.	*Premise:* Like Arabs and Jews, Diamond warns, Norwegians and Japanese are joined by blood still lock ed in traditional enmity. *Hypothesis:* Koreans and Japanese have no ension between them.	t5-3b	Neutral	Neutral	Contradiction
Premise: Placido Domingo's appearance on the package, compellingly photographed in costume as the ancient King of Crete, (Anthony Tommasini, the New York Times) is the main selling point for this new recording of one of Mozart's more obscure operas–a fact that does not make critics happy. *Hypothesis:* The attracting feature of the new Mozart recording is Placido Domingo's appearance.	*Premise:* Nunez Domingo's appearane on the package, compellingly photographed in costume as the ancient King of Crete, (Anthony Tommasini, the New York Times) is the main selling point for this new recording of one of Mozart's more obscure operas–a fact that does not make critics happy. *Hypothesis:* The attracting feature of the new Mozart recording is Placido Domingo's appearance.	facebook/opt-1.3b	Contradiction	Entailment	Entailment
Premise: Those Creole men and women you'll see dancing it properly have been moving their hips and knees that way since childhood. *Hypothesis:* Creole dances are learned from childhood.	*Premise:* Those Creole men and women you'll see dancing it properly have been moving their hips and knees that way since childhood. *Hypothesis:* Creole dances are learned from childhood.	facebook/opt-1.3b	Neutral	Entailment	Entailment
Premise: Exhibitions are often held in the splendid entrance hall. *Hypothesis:* The exhibitions in the entrance hall are usually the most exciting.	*Premise:* Exhibitions are often hled in the splendid entrance hall. *Hypothesis:* The exhibitions in the entrance hall are usually the most exciting.	gpt2-xl	Entailment	Neutral	Neutral
Premise: The Edinburgh International Festival (held annually since 1947) is acknowledged as one of the world's most important arts festivals. *Hypothesis:* The festival showcases exhibitions from all seven continents.	*Premise:* The Edinburgh International Festival (held annually since 1947) is acknowledged as one of the world's most important arts festivals. *Hypothesis:* Nova festival showcases exhibitions from all seven continents.	gpt2	Contradiction	Neutral	Neutral
Premise: Calcutta seems to be the only other production center having any pretensions to artistic creativity at all, but ironically you're actually more likely to see the works of Satyajit Ray or Mrinal Sen shown in Europe or North America than in India itself. *Hypothesis:* Most of Mrinal Sen's work can be found in European collections.	*Premise:* Calcutta seems to be the only other production center having any pretensions to artistry creativity at all, but ironically you're actually more likely to see the works of Satyajit Ray or Mrinal Sen shown in Europeans or North America than in India itself. *Hypothesis:* Most of Mrinal Sen's work can be found in European collections.	gpt2-large	Entailment	Neutral	Neutral

Table 26: Examples of MNLI attacked by Textbugger

original	**perturbed**	**model**	**model_pred**	**gpt4_pred**	**label**
Premise: That is exactly what our head coupon issuer Alan Greenspan did in 1987–and what I believe he would do again. *Hypothesis:* This is what Greenspan did in 1987 and what I think he will do again, much to the detriment of the economy.	*Premise:* That is exactly what our head coupon issuer Alan Greenspan did in 1987–and what I believe he would do again. *Hypothesis:* Hong is what Greenspan didnt in 1987 and what I inkling he yearn do again, much to the afflict of the economically.	gpt2	Contradiction	Contradiction	Neutral
Premise: GAO recommends that the Secretary of Defense revise policy and guidance *Hypothesis:* GAO recommends that the Secretary of Defense keep policy and guidance the same	*Premise:* GAO recommends that the Secretary of Defense revise policy and guidance *Hypothesis:* BRED insinuated that the Department of Defends conservation polices and hints the same	microsoft/deberta-v3-base	Neutral	Neutral	Contradiction
Premise: But when he was persuaded by divers means to help us, he gave up after one week, declaring it beyond his powers. *Hypothesis:* He decided it was too difficult because he was distracted by other topics.	*Premise:* But when he was persuaded by divers means to help us, he gave up after one week, declaring it beyond his powers. *Hypothesis:* He decided it was too rigid for he was distracted by other issuing.	t5-small	Contradiction	Neutral	Neutral
Premise: Today the strait is busy with commercial shipping, ferries, and fishing boats, and its wooded shores are lined with pretty fishing villages, old Ottoman mansions, and the villas of Istanbul's wealthier citizens. *Hypothesis:* Today, the strait is empty after a huge sand storm killed everyone there.	*Premise:* Today the strait is busy with commercial shipping, ferries, and fishing boats, and its wooded shores are lined with pretty fishing villages, old Ottoman mansions, and the villas of Istanbul's wealthier citizens. *Hypothesis:* Sundays, the strait is empty after a huge sand storm killed everyone there.	facebook/opt-350m	Neutral	Contradiction	Contradiction
Premise: get something from from the Guess Who or *Hypothesis:* Get something from the Guess Who,	*Premise:* get something from from the Guess Who or *Hypothesis:* Get something from the Bet Who,	t5-small	Contradiction	Contradiction	Entailment
Premise: However, co-requesters cannot approve additional co-requesters or restrict the timing of the release of the product after it is issued. *Hypothesis:* They cannot restrict timing of the release of the product.	*Premise:* However, co-requesters cannot approve additional co-requesters or restrict the timing of the release of the product after it is issued. *Hypothesis:* They cannot coerce timing of the discards of the product.	microsoft/deberta-v3-large	Contradiction	Neutral	Entailment
Premise: Tax purists would argue that the value of the homemakers' hard work–and the intrafamily benefits they presumably receive in return for it–should, in fact, be treated as income and taxed, just like the wages paid to outside service providers such as baby sitters and housekeepers. *Hypothesis:* To tax purists, the value of the homemakers' hard work should not be taxed.	*Premise:* Tax purists would argue that the value of the homemakers' hard work–and the intrafamily benefits they presumably receive in return for it–should, in fact, be treated as income and taxed, just like the wages paid to outside service providers such as baby sitters and housekeepers. *Hypothesis:* To tax purists, the appraised of the homemakers' hard work should not are imposition.	gpt2	Neutral	Contradiction	Contradiction
Premise: Cop Bud White (Crowe) and Ed Exley (Pearce) almost mix it up (59 seconds) : *Hypothesis:* Bud White and Ed Exley almost mix it up.	*Premise:* Cop Bud White (Crowe) and Ed Exley (Pearce) almost mix it up (59 seconds) : *Hypothesis:* Bud White and Ed Exley almost melting it up.	facebook/opt-2.7b	Neutral	Contradiction	Entailment
Premise: According to this plan, areas that were predominantly Arab the Gaza Strip, the central part of the country, the northwest corner, and the West Bank were to remain under Arab control as Palestine, while the southern Negev Des?Yrt and the northern coastal strip would form the new State of Israel. *Hypothesis:* We want to give Palestine and Israel a two-state solution that benefits both of them.	*Premise:* According to this plan, areas that were predominantly Arab the Gaza Strip, the central part of the country, the northwest corner, and the West Bank were to remain under Arab control as Palestine, while the southern Negev Des?Yrt and the northern coastal strip would form the new State of Israel. *Hypothesis:* We going to give Palestine and Israel a two-state solution that prerogatives both of them.	gpt2-medium	Entailment	Neutral	Neutral
Premise: Search out the House of Dionysos and the House of the Trident with their simple floor patterns, and the House of Dolphins and the House of Masks for more elaborate examples, including Dionysos riding a panther, on the floor of the House of Masks. *Hypothesis:* The House of Dolphins and the House of Masks are more elaborate than the House of Dionysos and the House of the Trident.	*Premise:* Search out the House of Dionysos and the House of the Trident with their simple floor patterns, and the House of Dolphins and the House of Masks for more elaborate examples, including Dionysos riding a panther, on the floor of the House of Masks. *Hypothesis:* The House of Dolphins and the House of Masks are more devising than the House of Dionysos and the House of the Trident.	facebook/opt-6.7b	Neutral	Neutral	Entailment

Table 27: Examples of MNLI attacked by TextFooler