FLUKE: A Linguistically-Driven and Task-Agnostic Framework for Robustness Evaluation

Yulia Otmakhova[1*] **Hung Thinh Truong**[1*] **Rahmad Mahendra**[3]
Zenan Zhai[2] **Rongxin Zhu**[2] **Daniel Beck**[4] **Jey Han Lau**[1]
[1]The University of Melbourne [2]Oracle [3]Universitas Indonesia [4]RMIT University
{y.otmakhova, thinh.truong}@unimelb.edu.au rahmad.mahendra@cs.ui.ac.id
zenanzhai@gmail.com rongxin.zhu@unimelb.edu.au,
daniel.beck@rmit.edu.au, jeyhan.lau@unimelb.edu.au

Abstract

We present FLUKE (framework for linguistically-driven and task-agnostic robustness evaluation), a framework for assessing model robustness through systematic minimal variations of test data. FLUKE introduces controlled variations across linguistic levels — from orthography to dialect and style — and leverages large language models (LLMs) with human validation to generate modifications. We demonstrate FLUKE's utility by evaluating both fine-tuned models and LLMs across six diverse NLP tasks (four classification and two generation tasks), and reveal that (1) the impact of linguistic variations is highly task-dependent, with some tests being critical for certain tasks but irrelevant for others; (2) LLMs still exhibit significant brittleness to certain linguistic variations, with reasoning LLMs surprisingly showing less robustness on some tasks compared to base models, and scaling improving robustness only for surface-level modifications; (3) models are overall more brittle to natural, fluent modifications such as syntax or style changes (and especially to negation), compared to corruption-style tests such as letter flipping; (4) the ability of a model to use a linguistic feature in generation does not correlate to its robustness to this feature on downstream tasks. These findings highlight the importance of systematic robustness testing for understanding model behaviors.

1 Introduction

The standard evaluation paradigm for machine learning models is to assess their performance on test data that shares the same distribution to the training set. Language evaluation benchmarks (Nie et al. (2020); Zellers et al. (2019); Wang et al. (2018, 2019); inter alia) are typically created based on such an assumption[1] and over the years we have seen a trend of increasing model performance; in some instances the state-of-the-art models are on par with human performance (Hupkes et al., 2023; Le Bras et al., 2020). Despite this positive trend, many studies found that models still perform poorly when dealing with minor variations (Ribeiro et al., 2020; McCoy et al., 2019), minimally contrastive pairs (Sennrich, 2017; Warstadt et al., 2020), out-of-domain data (Singhal et al., 2023), and adversarial input (Jia and Liang, 2017; Ebrahimi et al., 2018). This reveals that the success of NLP models may not be solely attributed to improved linguistic understanding and that they may have learned the data rather than the task (Li et al., 2023; Hupkes et al., 2023; Linzen, 2020).

In this paper, we introduce FLUKE (framework for linguistically-driven and task-agnostic robustness evaluation), a task-agnostic framework to generate minimal modifications of existing test data based on linguistic variations to evaluate model robustness.[2] That is, an original input (*everyone loves dogs*) is modified to examine a model's robustness in handling linguistic variations such as orthography (***evryone*** *loves dogs*), syntax (***dogs are loved by everyone***), semantics (***no one*** *loves dogs*), etc. These modifications are generated by prompting a large language model (LLM), and then validated manually to assess whether they target the intended linguistic property and require updating the ground-truth labels.

To demonstrate the generalisability of FLUKE, we apply it to modify instances for four classification tasks of various complexity — coreference resolution, dialogue contradiction detection, named entity recognition (NER), and sentiment analysis — as well as two generation tasks — instruction-following evaluation (IFEval) and mathematical

[1]An exception here is benchmarks specifically designed for evaluating large language models, which typically have no training data (Zhong et al., 2024; Hendrycks et al., 2021; Srivastava et al., 2022).

[2]Under the generalisation taxonomy of Hupkes et al. (2023), our modifications target specifically covariate shifts.

Findings of the Association for Computational Linguistics: EACL 2026, pages 5103–5123
March 24-29, 2026 ©2026 Association for Computational Linguistics

reasoning (GSM). We then test a number of models — including fine-tuned pre-trained models (PLMs: BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019) and T5 (Raffel et al., 2020)), base LLMs (GPT-4o, Llama 3.1-405b, and Claude 3.5), as well as reasoning LLMs (GPT-5 and DeepSeek R1) — and examine their predictions in the original and modified instances. When measuring the overall *robustness* of the model, we account for two cases: where the label of the instance is supposed to stay the same after modification and where the label changes. Accordingly, we consider the model to be robust if its predictions either remain the same or change together with the label. We find that: (1) Robustness to modifications highly depends on the task. (2) Even though LLMs are more robust than PLMs, they are still brittle to many modifications, and for some tasks the brittleness is surprisingly more prominent in reasoning LLMs. (3) Scaling helps to improve robustness only to surface-level modifications. (4) Natural, linguistically equivalent modifications are more damaging than traditional adversarial modifications such as random punctuation insertion; and (5) Models can be robust to linguistic features that they struggled to understand when generating, and brittle to features that are easy for them to employ. Thus, using only a handful of well-known tests may fail to reveal the brittleness of a model, highlighting the necessity of a comprehensive, task-agnostic testing framework like `FLUKE`. To summarise, our contributions are:

- We present `FLUKE`, a task-agnostic framework to generate minimal modifications of test data based on linguistic variations. We design prompts to automate the modifications, and perform extensive quality checks, showing that this approach works well without substantial human effort for most modifications.

- We demonstrate the utility of `FLUKE` by evaluating the performance of multiple NLP models, and show that a comprehensive, linguistically driven approach helps to reveal their task-specific shortcomings.

- We release our prompt templates, evaluation scripts, and validated modifications, to enable other researchers to test their models and tasks.[3] We also publish a dashboard that allows to assess models performance by comparing their outputs on original and modified instances.[4]

2 Related Work

Task-based Benchmarking Much of previous work in evaluating models is centered around collating datasets that are representative of a range of typical NLP applications (Wang et al., 2018; Song et al., 2023). The datasets in these benchmarks usually follow the typical paradigm of having an in-domain test set, coming from the same distribution as the training data. While this is useful in many experimental settings (such as model comparison), it is restricted to reporting in-domain performance. The importance of evaluation in out-of-domain scenarios is well known. Many shared tasks in NLP include test sets that are unrelated to the training data (Pei et al., 2023; Ousidhoum et al., 2024, *inter alia*). The development of so-called "challenge sets", initially explored in machine translation (Goyal et al., 2022) but now extended to LLMs (Zhong et al., 2024) also falls into this space. However, the range of domain differences is massive, which explains why it is hard to create "general", collated benchmarks for out-of-domain scenarios.

Linguistic Competence Probing Another way to assess models is through linguistic competence. A common approach is to use *probing classifiers*, tuned on datasets tailored to specific competences, such as morphology or syntax (Blevins et al., 2023; Mahowald et al., 2024). Holmes (Waldis et al., 2024) attempt to unify multiple datasets created to test linguistic competences. While probing for linguistic competence can provide insights about the model's understanding of specific linguistic phenomena, it is important to note that model competence does not necessarily translate to task performance, as we show below in Section 6.5.

Robustness Evaluation One line of work explores the idea of *minimal contrastive pairs* (Kann et al., 2019; Marvin and Linzen, 2018; Warstadt et al., 2019; Linzen et al., 2016), to evaluate a model's robustness to a specific phenomena in the context of a specific task (e.g. subject-verb agreement in syntax (Warstadt et al., 2020) or novel words in machine translation (Sennrich, 2017)). Another line focuses on creating *adversarial examples* to understand model failures (Eger and Benz,

[3] https://fluke-nlp.github.io/

[4] https://fluke-viewer.pages.dev/

2020; Keller et al., 2021; Tan et al., 2020; Formento et al., 2023; Li et al., 2024). These studies use model loss or gradient to learn how to create perturbations of test data that break the model.

Ribeiro et al. (2020) introduce the Checklist framework to test linguistic capabilities, but it is not fully task-agnostic, as it requires developers to design their own suite of tests (which, as we show below, may not fully reveal model brittleness), and uses templates to create modifications, which limits their diversity. To address these limitations, we propose a comprehensive suite of varied and natural linguistic modifications together with LLMs prompts that are ready to be applied to any task.

3 FLUKE

FLUKE systematically covers phenomena related to different levels of language structure, as well as some model biases. Simplified examples of each test are in Table 1.

3.1 Linguistic tests

We design tests for all levels of language structure, starting from low-level orthography tests and ending with dialect and style modifications.

Orthography Here we test the ability of models to deal with changes in **spelling** (*adding* or *omitting* a letter, as well as *swapping* two letters), **capitalization** (modifying the register to *lower case*, *Upper case*, *sPoNgEcAsE* or *ALL CAPS*) and **punctuation** (where we *add* or *change* a punctuation mark, or *remove the space* between two words).

Morphology We test if more complex words are harder for models. In **derivation** modification, we change a non-derived word (a word without any suffixes or prefixes) into a derived word. For **compound** modifications, we replace a word which has one root with a word that has several stems.

Syntax Here we test if a model adjusts to changing the sentence structure. In particular, we change a verb from **active to passive** (the subject and object are flipped accordingly to preserve the meaning). For **grammatical role** tests, we leave the verb intact, but swap two entities. For **conjunctions** modification we add an extra word combined with a coordinating conjunction *and* or *or*.

Semantics We devise tests that involve different degrees of meaning change. First, in **concept** modifications, we replace words with *synonyms*, with *hypernyms* or *hyponyms* (words that have a broader or narrower meaning), with artificially created, non-existing *nonce words*[5], and with *idioms*. We also use **negation** to test for the change of meaning that goes beyond individual words. Apart from *verbal* negation (adding *not* to the verb), we consider less straightforward types of negation: *approximate*, where the meaning is only partially negated, *absolute*, which explicitly excludes non-negative interpretation, as well as *lexical* negation (affixal negation and antonymy) (Pullum and Huddleston, 2002; Otmakhova et al., 2022). Finally, we look at *double* negation, i.e. grammatically valid structures which contain two negation elements (Truong et al., 2022).

Discourse We look at sensitivity to manipulating **discourse markers**: we *add* a discourse marker, explicitly connecting ideas, *change* a discourse marker, which can lead to a different meaning, or *delete* it, making the text more difficult to interpret. We also test robustness to adding markers of **appraisal**: sentiment-bearing words or phrases.

Language Varieties In these tests the changes are more extensive and span across all linguistic aspects. Specifically, we change the **style** to highly *casual* English, and rewrite the sentence in a particular **dialect:** Singaporean creole English (*Singlish*), which has systematic differences from American or British English varieties used in benchmarks.

3.2 Bias tests

In addition to linguistic tests, we include tests for data artifacts that models are known to exploit. We test if changing words to their old-fashioned variants affects the performance (**temporal bias**). When testing for **geographical bias** (Faisal and Anastasopoulos, 2023; Godey et al., 2024), we randomly select from a list of regions that are likely to be underrepresented in training corpora.[6] Then we change proper nouns to names and locations specific to such regions, and replace common nouns with relevant cultural entities. Finally, in **length bias** tests we either *shorten* or *lengthen* the text.

4 Task Evaluation

We apply modifications to test model robustness for six tasks, including four classification tasks (coreference resolution, dialogue contradiction detection,

[5]We use a list of nonce words from Cremers (2022).

[6]Africa, Middle East, Southeast, East and Central Asia, Oceania, Latin America, Eastern Europe

Test	Subtest	Original	Modified
Orthography			
Spelling	Addition	beautiful	beautiful**l**
	Omission	fantastic	fantstic
	Swapping	not a bid deal	not a big d**ae**l
Capitalization	Upper to lower	Battlefield 3	**b**attlefield 3
	Lower to upper	Hinckley did not hit Reagan	Hinckley did not **H**it Reagan
	sPoNgEcAsE	The professor	The pRoFeSsOr
	ALL CAPS	Michael Phelps	MICHAEL Phelps
Punctuation	Add	not exactly the bees knees	not exactly, the bees knees
	Change	It's worth tracking down.	It's worth tracking down!
	Remove space	so little movie	so littlemovie
Morphology			
Derivation	Derived	killed	assassinated
Compound	Compound	new	brand-new
Syntax			
Voice	Active to passive	Billy beat Tommy	Tommy was beaten by Billy
Gramm. role	Entity swap	Bob sued Bill	Bill sued Bob
Conjunctions	Coordinating	excessive hunting	excessive hunting and poaching
Semantics			
Concept	Synonym	suspect	doubt
	Hyper/hyponym	organization	association
	Nonce word	The bowl had a crack	The bowl had a vibble
	Idiom	they were tasty	they were a real treat
Negation	Verbal	They were afraid of the robots	They were **not** afraid of the robots
	Absolute	They were afraid of the robots	They were afraid of **none** of the robots
	Approximate	They were afraid of the robots	They were **seldom** afraid of the robots
	Lexical	They were afraid of the robots	They were **fearless** of the robots
	Double	They were afraid of the robots	They were **not unafraid** of the robots
Discourse			
Discourse markers	Addition	Toyota has Lexus: they are built for the rich.	Toyota has Lexus, and they are built for the rich.
	Change	The boss fired the worker when he stopped performing well.	The boss fired the worker after he stopped performing well.
	Remove	Tony helped Jeff as he needed help.	Tony helped Jeff, he needed help.
Appraisal	Addition	She turns her down.	She coldly turns her down.
Varieties			
Style	Casual	There is no pleasure in watching a child suffer	It's no fun seeing a kid suffer
Dialect	Singlish	He would not say no.	He dun wan say no.
Biases			
Temporal	Old-fashioned	He treats her badly.	He treats her ill.
Geographical	Names	Anna tried again	Dongxin tried again
	Cultural entities	The bat hit the ball	The lakau hit the polo
Length	Shorten	The lion saw the fish and it was swimming	The lion saw the fish swimming.
	Lengthen	Joseph did not defeat William	Joseph did not manage to defeat William

Table 1: FLUKE: our proposed linguistically-driven tests.

named entity recognition, sentiment analysis) and two generation tasks (instruction-following evaluation, IFEval, and mathematical reasoning, GSM).

4.1 Test set creation and validation

We generate modifications (Section 3) for each task, randomly selecting samples from a *test partition* of the original dataset and prompting GPT-4o with instructions. We create and evaluate a general prompt template for each modification (see Section A.1 for an example), and then use the same prompt template for each task with minimal modifications (such as specifying the text type).

We follow several steps to ensure the high quality of modifications. To improve their **diversity**, we stratify modifications across subtests (for example, different types of negation) and, where possible, randomly select a variable from a list (such as region and country for *geographical bias* or non-existing word for *nonce* semantic test) and pass it to the prompt. To ensure that modifications were applied **minimally**, i.e. the difference in performance can be attributed to a particular linguistic feature change rather than to a large number of changes, we calculate Levenshtein distance between original and modified samples, discarding those which changed more than expected. Finally, we employ thorough human checks to ensure that modifications are **valid and meaningful**. For *classification tasks* we ran this validation in two stages on Prolific[7], with thorough quality control and four annotators per sample (see Section A.2 for annotation details, inter-annotator agreement and quality statistics). During the first stage, we check if resulting texts **have the required modification** and are still **fluent**, rejecting samples that do not meet these criteria. We find that most tests have a *high retain rate* of 70-90%. During the second stage, we ask annotators to select the tasks' label (such as positive or negative for sentiment analysis) for modified samples, without knowing the labels for original samples, to **check if the ground-truth label changed** (e.g. after applying negation the sentiment can change from positive to negative). To make sure the task is **solvable** after modification, i.e. a label is not ambiguous or hard to choose, we add an option such as "Neutral" for sentiment analysis or "Not sure" for other tasks, and reject samples where annotators chose it.

For *generation tasks*, to demonstrate how the framework can be applied with minimal human intervention, we skip the first stage of human checks and only use tests with reliable modifications (over 70% retain rate). For the second stage, the authors of the paper validate modified samples in terms of whether the modification **changes the label** (for example, if it changes the premises in the math task, leading to a different answer) and whether the task is still **solvable** (for example, if there are modifications that prevent correct instruction following). Examples of such valid and rejected modifications, as well as modifications that led to label change, are shown in Table 9 in Appendix. Such setup reduces the human effort, making the framework *generalizable to new tasks*, while still ensuring the validity of modifications.

In total, we produce around 100 modified instances for each test in each task (1700 modified instances for each classification task and 1300 instances per generation task).

4.2 Tested models and evaluation

We experiment with three smaller PLMs (BERT, GPT-2, and T5), three base LLMs (GPT-4o, Llama 3.1 and Claude 3.5), and two reasoning LLMs (GPT-5, DeepSeek R1). Both PLMs and LLMs are tested on classification tasks, while only LLMs are used for generation tasks. The pretrained models are fine-tuned on the respective training set with an additional task-specific classification layer on top (Section A.5 in Appendix). The LLMs are prompted with task-related instructions in a zero-shot way (Section A.4).

To ensure a fair comparison of performance between the original and modified set, for each test we evaluate only the subset of the original samples that corresponds to the modified instances. We measure instability via an "unrobustness" score U computed on paired original–modified examples. Let o_i, m_i be per-sample correctness indicators for the original and modified items, respectively.[8] Then

$$U = \frac{1}{N} \sum_{i=1}^{N} |m_i - o_i| \cdot 100 \quad (1)$$

where N is the number of instances. Note that U counts both degradations (1→0) and improvements

[7] https://www.prolific.com/

[8] Correctness is defined as whether the prediction matches the ground truth label. Note that we use this formulation because it handles the situations where there is a label change after modification (e.g. sentiment has flipped after negation is introduced). In those cases, the prediction will also need to change for it to be "correct".

(0→1) as “unrobust”: our rationale is that if after modification the model correctly predicts a previously incorrect sample, it is still brittle since the model relies on a particular way the text is written rather than understanding its content.

We follow the original metrics used for the task to define correctness. For dialogue contradiction detection, coreference resolution, sentiment analysis, and GSM the correctness metric is accuracy. For NER, we use per-sample entity-level mean F1 (Nakayama, 2018), while IFEval employs strict success (all constraints satisfied) as correctness.

5 Results

5.1 Use Case: NER and Coreference tasks

In this section we present two classification tasks: named entity extraction (NER) and coreference resolution (Coreference).[9] For Coreference task, we use KnowRef dataset (Emami et al., 2019), which consists of pronoun disambiguation problems where each sample has the Winograd Schema Challenge (Levesque et al., 2012) format, consisting of a sentence, the pronoun we want to disambiguate, two candidates, and the index of the correct candidate as label. For NER, we chose Few-NERD (Ding et al., 2021), a dataset of Wikipedia sentences, using its 8 first-level entity categories, i.e., *Person, Location, Organization, Art, Building, Product, Event, Misc*. As explained in Section 4.1, we thoroughly check and update the labels to make sure the task is still valid and solvable after modification (for example, check if the pronoun can still be resolved or if the entity has changed).

Interestingly, unrobustness results for NER (see Table 2) and Coreference (Table 3) show different, almost orthogonal trends in terms of the tests the models were brittle to. For NER the most revealing test is **Geographical bias**, where all models predict entities from some locales better than from the others. While in some cases the models struggled to identify entities after modification, in the majority of cases it was actually easier for the models to predict “exotic” entities from non-English locales. Moreover, almost all models (except for BERT) lacked robustness to **Punctuation** modifications, heavily relying on **capitalization** as a cue for named entity presence or type. Interestingly, while base LLMs overall are more robust to modifications on this task, reasoning LLMs (especially DeepSeek) show major fluctuations across all NER tests, demonstrating similar unrobustness to PLMs.

Category	Modification	PLM			LLM					Avg
		BERT	GPT2	T5	GPT4o	Claude	Llama	GPT5	DS	
Bias	**Temporal**	3.7	3.0	1.2	6.2	4.3	1.8	7.7	10.6	4.8
	Geographical	25.1	27.6	29.0	22.3	26.0	27.0	22.0	27.8	25.8
	Length	11.8	12.2	13.1	12.7	15.7	6.4	9.7	12.6	11.8
Orthogr.	**Spelling**	4.3	2.5	1.4	6.6	3.0	3.9	7.3	11.4	5.0
	Capitalization	0.9	19.3	13.1	11.3	8.9	15.3	9.1	13.6	11.5
	Punctuation	6.5	3.7	4.9	7.1	9.1	7.9	7.8	15.0	7.8
Morphol.	**Derivation**	1.9	4.2	5.8	3.7	2.0	1.3	9.5	8.2	4.6
	Compound	3.1	0.6	5.2	3.1	1.7	1.0	7.0	12.9	4.3
Syntax	**Voice**	7.8	10.8	5.7	7.5	5.5	4.5	8.3	11.3	7.7
	Grammar	8.0	15.9	10.5	8.3	3.5	5.3	10.8	12.4	9.3
	Conjunction	9.1	7.6	7.7	9.4	7.5	8.0	11.7	13.2	9.3
Semantics	**Concept**	5.0	8.9	5.3	6.5	4.8	3.9	8.5	9.4	6.5
	Negation	4.8	5.2	6.5	8.2	3.9	4.6	10.5	15.2	7.4
Discourse	**Disc. markers**	4.7	1.6	5.4	2.2	1.9	3.2	6.3	8.7	4.2
	Appraisal	5.5	2.6	4.8	3.2	3.1	2.9	7.1	14.0	5.4
Varieties	**Style**	11.8	12.2	13.1	3.6	4.8	6.4	7.4	10.9	8.8
	Dialectal	12.6	7.4	8.4	5.9	8.9	4.7	7.0	13.2	8.5
Average		7.4	8.6	8.3	7.5	6.7	6.4	9.3	13.0	8.4

Table 2: NER: Unrobustness (U, %) by model and modification. *Claude* stands for Claude 3.5, *Llama* for Llama 3.1, and *DS* for DeepSeek R1. Full results with confidence interval (CI) in Table 13.

Category	Modification	PLM			LLM					Avg
		BERT	GPT2	T5	GPT4o	Claude	Llama	GPT5	DS	
Bias	**Temporal**	9.0	4.0	6.0	8.0	1.0	7.0	3.0	8.0	5.8
	Geographical	8.0	14.0	9.0	10.0	4.0	5.0	6.0	10.0	8.2
	Length	19.2	15.2	14.1	15.2	20.2	17.2	8.1	13.1	15.3
Orthogr.	**Spelling**	11.2	3.1	7.1	5.1	2.0	4.1	7.1	4.1	5.5
	Capitalization	15.2	14.1	7.1	7.1	4.0	5.1	2.0	6.1	7.6
	Punctuation	1.0	7.1	1.0	1.0	3.0	1.0	2.0	4.0	2.5
Morphol.	**Derivation**	4.1	3.1	5.1	4.1	1.0	1.0	5.1	2.0	3.2
	Compound	6.2	4.2	3.1	5.2	6.2	6.2	3.1	4.2	4.8
Syntax	**Voice**	35.8	34.7	41.1	26.3	15.8	30.5	9.5	15.8	26.2
	Grammar	30.6	27.8	19.4	22.2	18.1	19.4	16.7	20.8	21.9
	Conjunction	5.2	10.3	8.2	8.2	4.1	7.2	5.2	6.2	6.8
Semantics	**Concept**	8.0	3.0	5.0	11.0	18.0	9.0	10.0	10.0	9.2
	Negation	25.5	23.5	24.5	21.4	24.5	22.4	22.4	30.6	24.4
Discourse	**Disc. markers**	8.0	6.0	8.0	19.0	10.0	7.0	9.0	11.0	9.8
	Appraisal	5.0	8.0	4.0	8.0	7.0	9.0	5.0	7.0	6.6
Varieties	**Style**	16.0	19.0	18.0	19.0	14.0	14.0	14.0	12.0	15.8
	Dialect	8.3	26.9	22.2	24.5	11.8	16.7	2.9	12.7	15.8
Average		12.7	13.2	11.9	12.7	9.7	10.7	7.7	10.5	11.1

Table 3: Coreference: Unrobustness (U, %) by model and modification. Notation as in Table 2. Full results with CI in Table 14.

Conversely, for the Coreference task, while models were relatively robust to the tests affecting NER, they demonstrated major instability on tests involving modifications of syntactical structure and meaning, such as change of active **Voice** to passive, swapping **Grammatical roles**, or introducing **Negation**, as well as perturbations involving significant changes of context such as **Length bias**, change to casual **Style**, or to **Singlish** dialect. Unlike NER, on Coreference tasks LLMs (with the exception of GPT-4o) perform better than PLMs; however, they still demonstrate unrobustness to the majority of tests, with even the best model (GPT-5) significantly affected by grammatical roles change, negation and style modifications.

[9]The results for the other two classification tasks – Sentiment analysis and Dialogue Contradiction resolution – are in Section A.8 due to space constraints.

Category	Modification	GPT-4o	Claude-3.5	Llama 3.1	GPT-5	DS R1	Avg
Bias	**Temporal**	1.0	2.0	4.0	0.0	1.0	1.6
	Geographical	5.0	5.0	7.0	1.0	2.0	4.0
	Length	4.0	2.0	2.0	0.0	2.0	2.0
Orthogr.	**Spelling**	1.0	0.0	2.0	1.0	0.0	0.8
	Capitalization	3.0	1.0	5.0	1.0	1.0	2.2
	Punctuation	0.0	0.0	5.0	1.0	1.0	1.4
Semantics	**Concept**	1.0	0.0	5.0	2.0	1.0	1.8
	Negation	15.0	15.0	18.0	7.0	9.0	12.8
Discourse	**Appraisal**	0.0	0.0	1.0	1.0	1.0	0.6
Varieties	**Style**	4.0	3.0	6.0	4.0	3.0	4.0
	Dialect	2.0	4.0	6.0	1.0	2.0	3.0
Syntactic	**Conjunction**	0.0	1.0	4.0	1.0	1.0	1.4
	Voice	2.0	2.0	6.0	1.0	2.0	2.6
Average		2.9	2.7	5.5	1.6	2.0	2.9

Table 4: GSM: Unrobustness (U, %) by model and modification. Notation as in table 2. Full results with CI in Table 17.

5.2 Use case: GSM and IFEval

We choose two generation tasks — mathematical reasoning and instruction following — to demonstrate the applicability of FLUKE to testing generative models. For the mathematical reasoning task we choose Grade School Math (GSM8K) dataset (Cobbe et al., 2021), which contains mathematical problems expressed in natural language. To simplify the task, we use only the final computed answer (rather than the reasoning steps) to evaluate the solution correctness. For the instruction following task, we use the IFEval benchmark (Zhou et al., 2023) which contains varied generation prompts (“Write me a letter...”) that include 25 types of verifiable instructions such as “include word XXX at least 3 times" or “use 2 bullet points". Due to the nature of these tasks, we evaluate only LLMs on them. We ensure that the tasks are solvable after the modification, and that the expected output (solution or generated text) is still the same for all tests except for some types of **Negation** (**verbal, absolute, lexical**) where the modification changes the condition of the task and thus the model is expected to generate a different answer (see Section 4.1).

We find that on the GSM task (Table 4) models demonstrate major brittleness to **Negation** tasks, where all of them fail to take into account the changed or even reversed logic of the premise, showing memorisation effects. Base LLMs such as Claude-3.5 and especially Llama 3.1 are also affected by context changes, such as failing to accomodate for changed names and currencies in **Geographical bias** test, or explaining the problem in informal **style** or **Singlish** dialect. Interestingly, reasoning LLMs are more robust to such changes.

Compared to solving mathematical problems, the LLM’s ability to follow instructions are much more affected by the way the instructions are written — with both base and reasoning LLMs showing significant fluctuations in performance across all tests. This includes not only the “usual suspects” such as **Negation**, **Style** and **Dialect**, but also more innocuous changes such as adding an **appraisal** word or phrase (*please repeat -> please be so nice as to repeat*), adding a word with a **conjunction** (*Rewrite the following text -> Rewrite and transform the following text*), changing the case of a single word in instruction, etc. Here, reasoning models do not seem to handle modifications better than the base ones.

Category	Modification	GPT-4o	Claude-3.5	Llama 3.1	GPT-5	DS R1	Avg
Bias	**Temporal**	11.1	4.0	12.1	5.1	12.1	8.9
	Geographical	9.0	12.0	14.0	9.0	10.0	10.8
	Length	11.0	10.0	18.0	6.0	16.0	12.2
Orthogr.	**Capitalization**	8.1	5.1	10.1	9.1	13.1	9.1
	Punctuation	7.1	3.0	13.1	8.1	12.1	8.7
	Spelling	3.1	4.1	7.2	6.2	8.2	5.8
Syntax	**Conjunction**	11.0	7.0	6.0	8.0	11.0	8.6
	Voice	9.0	9.0	10.0	6.0	15.0	9.8
Semantics	**Concept**	9.1	7.1	11.1	7.1	7.1	8.3
	Negation	24.0	22.0	24.0	23.0	23.0	23.2
Discourse	**Appraisal**	7.1	4.1	11.2	7.1	6.1	7.1
Varieties	**Style**	18.0	7.0	10.0	9.0	11.0	11.0
	Dialect	13.0	13.0	12.0	11.0	15.0	12.8
Average		10.8	8.3	12.2	8.8	12.3	10.5

Table 5: IFEval: Unrobustness (U, %) by model and modification. Notation in Table 2. Full results with CI in Table 18.

6 Discussion

6.1 Model brittleness is task-specific

We call our framework task-agnostic, since the same set of prompts can be used to generate a battery of tests applicable to all tasks. This, however, does not mean that all tests are meaningful for all the tasks. As the results above (and in Section A.8) show, the tests that actually “fired", or showed substantial drops or increases, are different for each task. Even test such as **negation** which is known to be difficult for models (Ravichander et al., 2022; Truong et al., 2022) has almost no effect on the NER task. On the other hand, seemingly “simple” tests (**capitalization**) that have a negligible effect on most of the tasks led to some dramatic drops on the NER and IFEval tasks. Thus, our framework is task-agnostic because it can be applied to reveal lack of robustness on a new task without having a preconceived (and potentially wrong) notion of what tests are meaningful for it.

6.2 LLMs are not always more robust than PLMs, and reasoning enhanced training does not always help

In classification tasks, where PLMs were finetuned on the original dataset, one would expect larger fluctuations in performance compared to LLMs, as modifications are likely to change the data distribution and break the identical distribution assumption that the fine-tuning process is based on. This holds for two out of four classification tasks we tested — Coreference resolution (Table 3) and Sentiment analysis (Table 16 in Section A.8) — where PLMs are overall less robust to perturbations. However, on two other classification tasks — NER (Table 2) and Dialogue contradiction detection (Table 15 in Section A.8) — while base LLMs are overall more robust than PLMs, reasoning LLMs are as brittle as the latter. Moreover, while LLMs are on average (across all models and tests) better (see the "Average" row in Table 20), they are sometimes markedly brittle to a particular test. For example, Claude-3.5 is substantially more affected by **concept** modifications than any PLMs on Coreference resolution, or GPT-4o affected by **grammatical role** modifications on Dialogue contradiction task.

Comparing between base and reasoning LLMs across tasks does not reveal a clear winner either. Reasoning seems to improve overall robustness on tasks such as GSM (Table 4), Coreference (Table 3) or Sentiment Analysys (Table 16); however, on other tasks the reasoning models perform similar to base LLMs (IFEval, Table 5) or even worse than them (NER, Table 2, and Dialogue contradiction detection, Table 15).

We also test if a reasoning model (GPT-5) is able to handle the modifications better if it is *aware* of them. To do that, we include additional context into the prompt where we alert the model that the sample was modified, and explicitly specify the type of modification (e.g. by telling it that the active voice was changed to passive). We found no consistent gain when we do this (see Appendix Table 12).

6.3 Scaling helps but with diminishing effects and only for superficial modifications

To test if increasing the parameter size helps with robustness, we additionally compare Llama-3.1 405B (which was used in all experiments) with two smaller versions – Llama-3.1-8B and Llama-3.1-70B – on the GSM task (the results in Table 6). Overall, we observe a large improvement in robustness when going from 8B → 70B but observe diminishing effects when going beyond that. Low-level **orthographic** and **syntactic** modifications benefit strongly from increased model size. In contrast, more complex perturbations, especially **negation**, remain highly challenging even at 405B. This suggests that scaling primarily improves surface-level robustness, while deeper semantic changes remain brittle.

Category	Modification	Llama-8b	Llama-70b	Llama-405b	Avg
Bias	**Temporal**	8.0	5.0	4.0	5.7
	Geographical	8.0	3.0	7.0	6.0
	Length	8.0	2.0	2.0	4.0
Orthogr.	**Spelling**	10.0	2.0	2.0	4.7
	Capitalization	11.0	2.0	5.0	6.0
	Punctuation	5.0	1.0	5.0	3.7
Syntax	**Conjunction**	12.0	2.0	4.0	6.0
	Voice	11.0	2.0	6.0	6.3
Semantic	**Concept**	12.0	1.0	5.0	6.0
	Negation	22.0	17.0	18.0	19.0
Discourse	**Appraisal**	7.0	5.0	1.0	4.3
Varieties	**Style**	16.0	6.0	6.0	9.3
	Dialect	10.0	6.0	6.0	7.3
Average		10.8	4.2	5.5	6.8

Table 6: GSM Scaling: Unrobustness (U, %) by Llama model size and modification. Notation in Table 2. Full results with CI in Table 19.

6.4 Models are overall less robust to natural, linguistically valid modifications

Some of the work on robustness focuses on corrupting the text in adversarial way, such as shuffling characters or changing the case (Belinkov and Bisk, 2017; Li et al., 2018; Eger and Benz, 2020), while some includes natural, linguistically equivalent perturbations such as synonym replacement (Ren et al., 2019; Jin et al., 2020). Our framework combines both types of tests and allows to compare their efficiency for a particular task. We find that on average linguistically valid modifications reveal unrobustness more often than corrupting the text (i.e. such tests as **Orthography**) (see "Average" column in Table 20). The exception here are the NER task where orthography is directly relevant to performance, and, surprisingly, IFEval, where the model often confused the instructions regarding capitalization with capitalization used in the instruction. Thus, the choice of the test type – adversarial corruptions or natural modifications – should also be motivated by the downstream task.

6.5 The ability to generate a modification does not entail robustness against it

Though it is a common practice to use LLMs to generate test data, this raises a question of circularity if the same model (in our case, GPT-4o) is used to modify instances and then tested against them. To address this concern, we compare the quality of modifications generated by GPT-4o (as evaluated by annotators in terms of **retain rate**, i.e. percentage of modifications that were found correct and fluent) with the robustness of the same model to these modifications. As can be seen from Figure 1, as well as low and insignificant Pearson correlation, there is no clear relation between these two aspects: while some modifications (**Geographical bias**, **Dialect**) are easy for the model to generate, they lead to significant unrobustness; on the other hand, some modifications can be very hard (**Derivation**, **Compounds**), but the model is less brittle to them. Thus, the ability of the model to *understand* a linguistic feature (i.e. apply it when generating) does not straightforwardly translate to its ability to handle this feature on downstream tasks, and the samples that were generated by a model so as to include a particular feature can be used to test robustness to that feature.

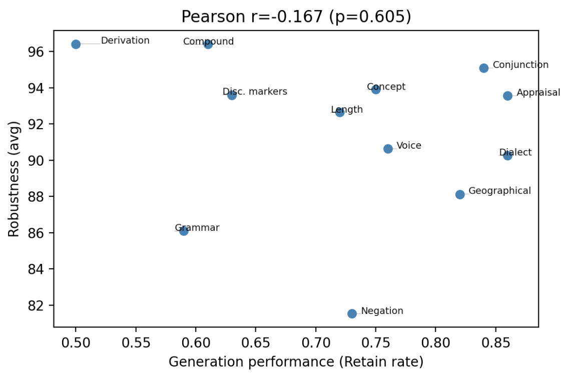

Figure 1: Generation Performance vs Robustness (100-U) of GPT-4o

7 Conclusion

In this paper, we present FLUKE, a task-agnostic, linguistically-driven framework that introduces minimal modifications to existing datasets to create a benchmark for assessing robustness. We automate these modifications by prompting GPT-4o, followed by thorough manual validation to ensure data quality. To showcase the generalizability of this framework, we apply it to create modified instances across six tasks covering both classification and generation, and evaluate the performance of smaller fine-tuned models and large language models. Our findings reveal that: (1) the impact of modifications is highly task-dependent, with some tests being critical for certain tasks but irrelevant for others; (2) while LLMs demonstrate better overall robustness compared to PLMs, they still exhibit significant brittleness, and on some tasks reasoning LLMs are less robust than base LLMs; (3) increasing the model size helps only with more surface-level modifications and has diminishing effects; (4) models are overall less robust to linguistically valid modifications than to adversarial corruption, but this again depends on the task; (5) the ability of a model to implement a linguistic feature during generation does not correlate with its robustness to that feature on downstream tasks. We hope that FLUKE provides an alternative approach for evaluating models, and that model developers would consider integrating these results into model cards (Mitchell et al., 2019) when releasing models.

8 Limitations

Although FLUKE aims to cover a wide range linguistics capabilities, we acknowledge that it is by no means exhaustive: there are still many areas that it can be expanded, and the current suite of capability tests should serve as a foundation for model developers to create more task-specific tests.

FLUKE is currently designed for English. Although many of its capability tests are applicable to other languages, our LLM prompts, human validation experiments and results are limited to English. To adopt FLUKE to another language, we recommend performing a thorough two-stage human evaluation on a sample of tasks to estimate the retain rate of modifications and reveal the tests that are likely to cause label change. For subsequent task, a simplified evalation with human-in-the-loop is possible.

Although we generate the modifications with GPT-4o, the process is ultimately not fully automated, as we still require human assessment to validate the modifications. That said, since most of the automatic modifications were shown to be valid and we only need a small number of instances to understand model capability, when FLUKE is used to test a particular model for a particular task, we believe the validation can be done in-house by the model developers.

Previous works have pointed out potential biases in data generated by LLMs, stemming from the lack of its diversity (Ding et al., 2024). In our work, such biases are most likely to occur in the geographical bias modification. To ensure a diverse representation, we compile a list of underrepresented regions as discussed in Section 3.2, and explicitly specify a region from this list in the prompt to generate samples across the variety of underrepresented geographical locations.

Acknowledgements

This research is supported by Oracle. Lau was supported by Australian Research Council

References

Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting language models for linguistic structure. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6649–6663, Toronto, Canada. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Alexandre Cremers. 2022. Interpreting gradable adjectives: rational reasoning or simple heuristics? *Empirical Issues in Syntax and Semantics*, 14:31–61.

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.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1679–1705, Bangkok, Thailand. Association for Computational Linguistics.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-NERD: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3198–3213, Online. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia.

Steffen Eger and Yannik Benz. 2020. From hero to zéroe: A benchmark of low-level adversarial attacks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 786–803, Suzhou, China.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. The KnowRef coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3952–3961, Florence, Italy. Association for Computational Linguistics.

Fahim Faisal and Antonios Anastasopoulos. 2023. Geographic and geopolitical biases of language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 139–163, Singapore. Association for Computational Linguistics.

Brian Formento, Chuan Sheng Foo, Luu Anh Tuan, and See Kiong Ng. 2023. Using punctuation as an adversarial attack on deep learning-based NLP systems: An empirical study. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1–34, Dubrovnik, Croatia.

Nathan Godey, Éric de la Clergerie, and Benoît Sagot. 2024. On the scaling laws of geographical representation in language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12416–12422, Torino, Italia. ELRA and ICCL.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, and 1 others. 2023. A taxonomy and review of generalization research in NLP. *Nature Machine Intelligence*, 5(10):1161–1174.

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.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERR really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R. Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 287–297.

Yannik Keller, Jan Mackensen, and Steffen Eger. 2021. BERT-defense: A probabilistic model based on BERT to combat cognitively inspired orthographic adversarial attacks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1616–1629, Online.

Klaus Krippendorff. Computing Krippendorff's Alpha-reliability. *Computing*, 1:25–2011.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *International conference on machine learning*, pages 1078–1088.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. 2024. GSM-plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2961–2984, Bangkok, Thailand. Association for Computational Linguistics.

Xinzhe Li, Ming Liu, Shang Gao, and Wray Buntine. 2023. A survey on out-of-distribution evaluation of neural NLP models. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6683–6691.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28:517–540.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of ACL*, pages 3428–3448, Florence, Italy.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Hiroki Nakayama. 2018. seqeval: A Python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

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

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1699–1713, Online. Association for Computational Linguistics.

Julia Otmakhova, Karin Verspoor, Timothy Baldwin, Antonio Jimeno Yepes, and Jey Han Lau. 2022. M3: Multi-level dataset for multi-document summarisation of medical studies. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3887–3901.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Meriem Beloucif, Christine De Kock, Oumaima Hourrane, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Krishnapriya Vishnubhotla, Seid Muhie Yimam, and Saif M. Mohammad. 2024. SemEval task 1: Semantic textual relatedness for African and Asian languages. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1963–1978, Mexico City, Mexico. Association for Computational Linguistics.

Jiaxin Pei, Vítor Silva, Maarten Bos, Yozen Liu, Leonardo Neves, David Jurgens, and Francesco Barbieri. 2023. SemEval-2023 task 9: Multilingual tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2235–2246, Toronto, Canada. Association for Computational Linguistics.

Geoffrey K. Pullum and Rodney Huddleston. 2002. *Negation*, chapter 9. Cambridge University Press.

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

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

Abhilasha Ravichander, Matt Gardner, and Ana Marasovic. 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.

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.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382.

Prasann Singhal, Jarad Forristal, Xi Ye, and Greg Durrett. 2023. Assessing out-of-domain language model performance from few examples. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2385–2397, Dubrovnik, Croatia.

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.

Yueqi Song, Simran Khanuja, Pengfei Liu, Fahim Faisal, Alissa Ostapenko, Genta Winata, Alham Fikri Aji, Samuel Cahyawijaya, Yulia Tsvetkov, Antonios Anastasopoulos, and Graham Neubig. 2023. GlobalBench: A benchmark for global progress in natural language processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14157–14171, Singapore. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It‘s morphin’ time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online.

Thinh Hung Truong, Julia Otmakhova, Timothy Baldwin, Trevor Cohn, Jey Han Lau, and Karin Verspoor. 2022. Not another Negation Benchmark: The NaN-NLI test suite for sub-clausal negation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 883–894.

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes: A benchmark to assess the linguistic competence of language models. *Preprint*, arXiv:2404.18923.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the*

2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2877–2887, Hong Kong, China.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A Appendix

A.1 Data generation prompts

An example of generating compound word modification for coreference resolution task.

```
Find any non-compound (single-root) word
in the text below and change it into a
compound word (word with several roots).
Don't make change to the pronoun.

Example: "a sequence of ridiculous
shooting scenes" -> "a sequence of
ridiculous shoot-'em-up scenes"
Example: dull acting -> lacklustre acting

Text: Joe raced against Steven because he
thought it would be easy.
Pronoun: he

Modified text: Joe raced against Steven
because he thought it would be a
no-brainer.
```

A.2 Annotation process and instructions

We perform a two-staged validation of modified samples using a crowdsource annotation platform Prolific [10]. To ensure high quality of annotations, we use a high per hour rate of 12 euro, which is above the minimal payment rate in the annotators' country of residence and well above the usual Prolific rate (overall cost: 7233 euro), choose only annotators from English-speaking countries (except for the **Dialect** test where we only used annotators from Singapore), and with a minimum success rate of 90%, as well as perform quality checks as described below. The human validation experiments has been approved by the ethics board (application ID: anonymised), and the annotators were informed of how their data will be used and had to express their consent.

During Stage 1, the annotators received instructions that described an intended modification (**Voice, Negation, Style** etc., see Section 3) and were asked if the sample they see is modified correctly according to these instructions (see Figure 2 for an example of annotation interface). For quality control, we include four control questions, which were randomly sampled from a different modification. We remove annotators who fail to answer more than one of such questions (i.e. selecting "Correct" twice), or select one option too frequently (i.e. selecting "Incorrect" for all questions). Each of the samples is annotated by four annotators, and receives at least two annotations after the quality checks. We filter our samples which were not deemed to be "Correct" by the majority (over 50%) of annotators. In case of tie, we select the answer of the annotator with better performance on the control questions. The inter-annotator agreement in terms of Krippendorf's α (Krippendorff), averaged majority-class agreement (percentage of samples where the majority of annotators chose the same label), and the retain rate (percentage of samples that pass phase 1 annotation) are shown in Table 7. We choose Krippendorf's α as an agreement metrics to accommodate for the difference in number of annotations per sample. We note, however, that it is not directly applicable here as it assumes that all annotations are provided by a fixed, ordered set of observers and factors in the consistency of their decisions, i.e. measures the annotators behavior agreement rather than agreement of labels. Thus, in our case the Majority rate agreement reveals

[10] https://www.prolific.com/

	Modification	Krippendorff's α	Majority rate	Retain rate
Bias	**Temporal**	0.33	0.85	0.80
	Geographical	0.35	0.84	0.82
	Length	0.30	0.83	0.72
Morphology	**Derivation**	0.14	0.79	0.50
	Compound	0.39	0.86	0.61
Syntax	**Voice**	0.30	0.83	0.76
	Grammar	0.17	0.80	0.59
	Conjunction	0.39	0.88	0.84
Semantics	**Concept**	0.37	0.84	0.75
	Negation	0.42	0.90	0.73
Pragmatic	**Disc. markers**	0.22	0.78	0.63
	Appraisal	0.46	0.89	0.86
Varieties	**Style**	0.53	0.91	0.93
	Dialect	0.41	0.87	0.86

Table 7: Phase 1 annotation quality

	Krippendorff's α	Majority rate	Label change rate
Sentiment Analysis			
Semantics: Negation	0.54	0.84	0.53
Dialogue Contradiction Detection			
Syntax: Grammar	0.34	0.81	0.33
Semantics: Negation	0.39	0.80	0.41
Coreference Resolution			
Syntax: Grammar	0.66	0.91	0.82
Semantics: Negation	0.57	0.79	0.49
NER			
Bias: Geography	0.36	0.59	1.00
Syntax: Grammar	0.43	0.67	0.94
Syntax: Conjunction	0.36	0.63	1.00

Table 8: Phase 2 annotation quality

more about the patterns of annotation agreement, and achieves substantially high levels.

During Stage 2, the annotators were given instructions for one of the four tasks (**Sentiment Analysis, Dialogue Contradiction Detection, Coreference Resolution, Named Entity Recognition**) and asked to assign a label specific for that task (for example, judge a modified sample as having *positive* or *negative* sentiment for the **Sentiment Analysis** task, see Figure 3). For quality checks, we used unmodified samples from the training data of the task (i.e. samples for which we knew the label); and apply the same filtering criteria in Stage 1. The final label was also selected based on the majority vote. We compared the resulting labels for modified samples with the labels for the original data to check if any inconsistencies were due to the genuine label change resulting from the modification, rather than to annotators disagreement. The inter-annotator agreement results (measured similarly to Stage 1) are presented in Table 8.

A.3 Examples of annotation decisions for generative tasks (simplified quality assurance scenario)

We provide examples of decisions where we kept the modification together with the original label for the task, changed the label, or rejected the modification, in Table 9.

A.4 LLM settings

We report the exact endpoint and sampling parameters for LLMs in Table 10.

Sentiment analysis prompt

```
Classify the sentiment of the given text.
Answer with 1 for positive, 0 for negative.
Text: it's a charming and often affecting journey.
Answer: 1
```

Dialogue contradiction prompt

```
Does the last utterance contradict the dialogue context? Answer with 1 if contradict, 0 if not contradict.
Dialogue context:
...
agent 0: well, i'm a big aerosmith fan, but i also like country.
agent 1: aerosmith is not bad but i love country so much
...
Last utterance: agent 1: i am a nascar fan too!
Answer: 1
```

Coreference resolution prompt

```
Which candidate does the pronoun refer to? Answer with either 0 or 1.
Text: The sniper shot the terrorist because he was a bad guy.
Pronoun: he
Candidates: 0: The sniper, 1: the terrorist
Answer: 1
```

Named entity recognition prompt

```
Extract named entities from the text. Possible entity types: ART, BUILDING, EVENT, LOCATION, ORGANIZATION, OTHER, PERSON, PRODUCT. Reply with the list of entities in the format ["text": entity_span , "label": entity_label,].
Text: Ronald will travel to Iceland.
Answer: [{"text": "Ronald", "value": "PERSON"}, {"text": "Iceland", "value": "LOCATION"}]
```

A.5 PLM settings

We report the finetuning details for PLMs in Table 11.

SA: We add a classification head on top of PLMs and fine-tune as a sequence classification task. Input: text, Output: label (0|1)
Dialogue: We add a classification head on top of PLMs and fine-tune as a sequence classification task. Input: dialog_context[SEP]last_utterance,

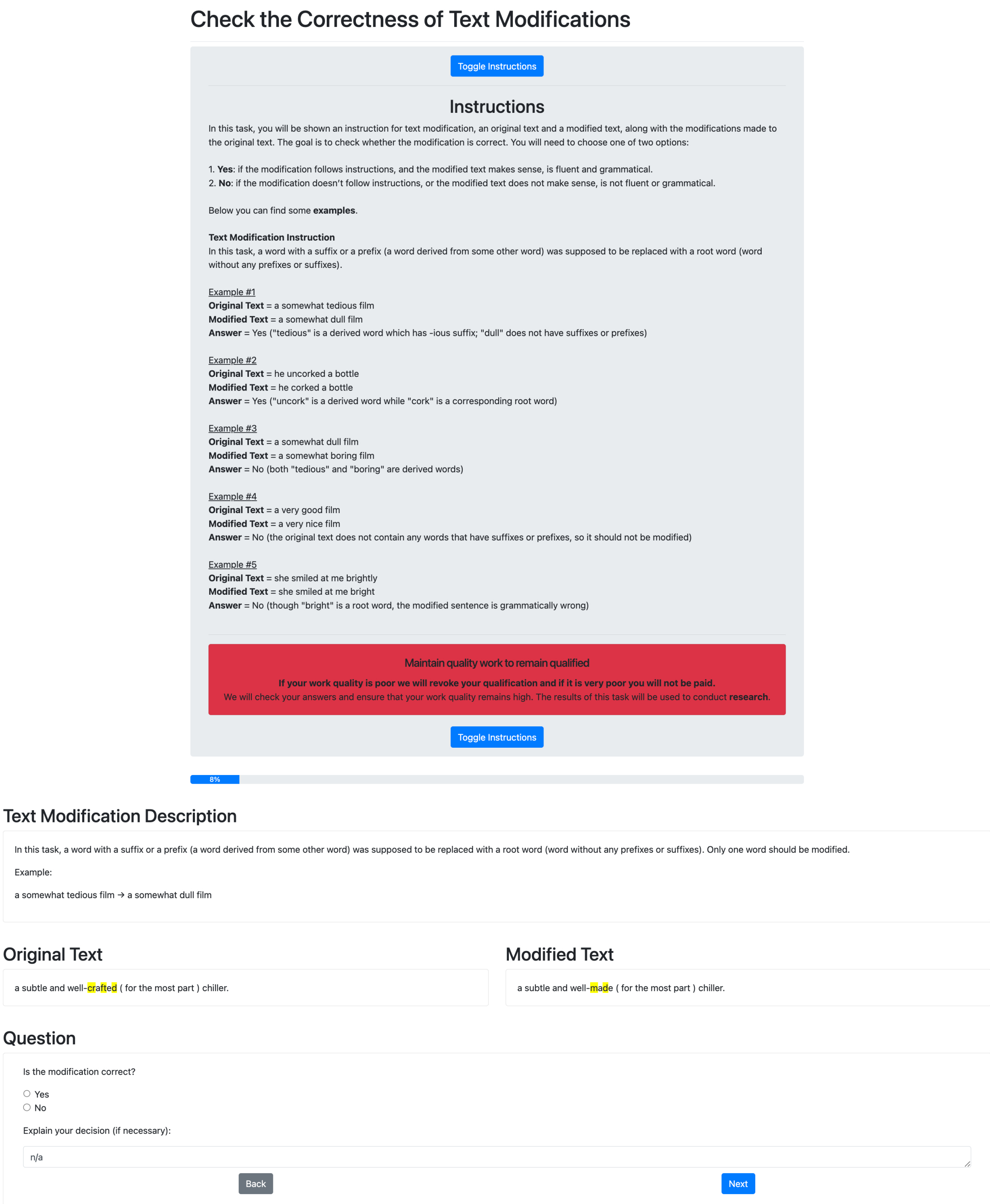

Figure 2: The example annotation page for the use case **Sentiment Analysis (SA)** for the linguistic capability test **derivation** in Stage 1.

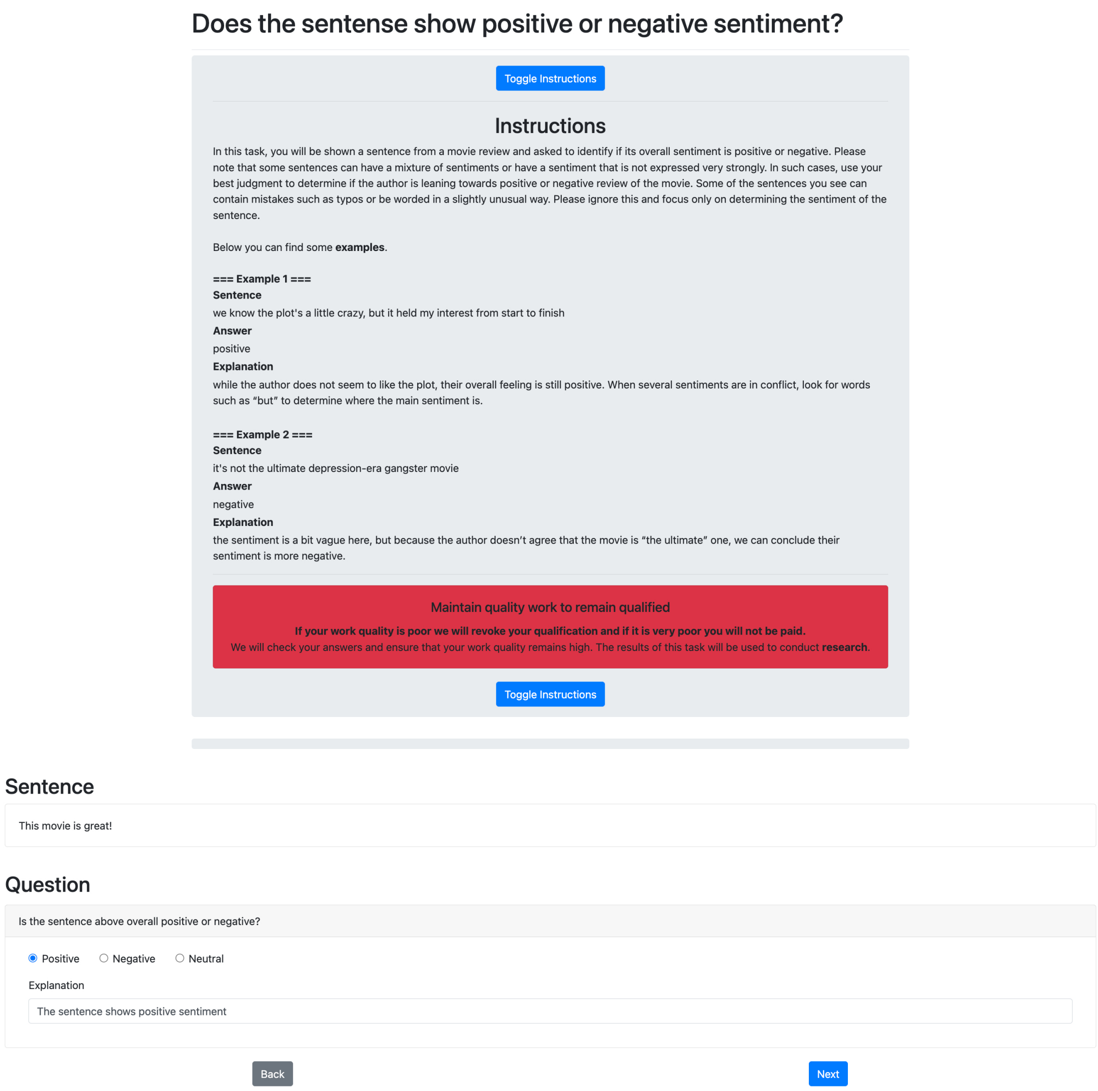

Figure 3: The example annotation page for the use case **Sentiment Analysis (SA)** in Stage 2.

Task	Test	Original	Modified	Keep?	Reason
GSM	Dialect	Reggie, Lynn, and Paisley ran together. Paisley ran 4 miles. Reggie ran 5 times what Paisley ran and 3 miles farther than Lynn. How many miles did Lynn run?	Reggie, Lynn, and Paisley *go running kaki*. Paisley ran 4 miles. Reggie ran 5 times what Paisley ran and 3 miles *more* than Lynn. *Lynn run how many miles*?	✓	Despite substantial language change, nothing has changed in the premises of the mathematical problem, so we keep the sample.
GSM	Negation	Miss Albert's class is composed of 12 boys and 12 girls. One-third of the girls and one-fourth of the boys are on varsity. How many students are not on varsity?	Miss Albert's class is composed of 12 boys and 12 girls. One-third of the girls and one-fourth of the boys are *not* on varsity. How many students are not on varsity?	→	The modification is valid (the problem can be solved), but leads to a different answer (label).
GSM	Concept	3 trees each had 7 blue birds in them. 2 different trees each had 4 blue birds. 1 final tree had 3 blue birds. How many blue birds were in the trees in total?	3 trees each had 7 blue birds in them. 2 different trees each had 4 blue birds. 1 final tree had 3 *sprat* birds. How many blue birds were in the trees in total?	✗	"Blue" was replaced by a nonce (non-existing) word "sprat". It is impossible to determine if "sprat birds" are blue or not, so the problem cannot be solved.
IFEval	Temp. bias	Write exactly 4 paragraphs about tips for installing a car seat for moms. Use 2 new lines to separate paragraphs. Start the 4th paragraph with the word elm.	Write exactly 4 paragraphs about tips for installing a car seat for moms. Use 2 new lines to separate paragraphs. *Commence* the 4th paragraph with the word 'elm'.	✓	The word "begin" was replaced by a slightly outdated word "commence", which is still well understandable in the context. The modification is valid and does not change the label.
IFEval	Negation	Write me a resume for Matthias Algiers. Use words with all capital letters to highlight key abilities, but make sure that words with all capital letters appear less than 10 times.	Write me a resume for Matthias Algiers. Use words with all capital letters to highlight key abilities, but make sure that words with all capital letters appear *more* than 10 times.	→	"Less" was changed to "more", so if the model generates less than 10 words in all caps, it should be considered wrong.
IFEval	Conjunct.	...Your entire response should be in English, and should not contain any capital letters.	...Your entire response should be in English, and should not contain any capital letters *or punctuation*.	✗	The modification is fluent and applied correctly, but it makes the task more difficult (it is hard to avoid any punctuation), so we reject it.

Table 9: Decision process for keeping (✓), rejecting (✗) modification, or changing the label (→)

Model	**GPT4o**	**Claude-3.5**	**Llama-3.1**	**GPT-5**	**Deepseek R1**
Endpoint	gpt-4o-2024-05-13	claude-3.5-sonnet	llama-3.1-405b-instruct	gpt-5	deepseek-r1
# Params	Undisclosed	Undisclosed	405B	Undisclosed	671B
Temperature	0	0	0	1 (default by reasoning model)	1 (default by reasoning model)
Max token	1024	1024	1024	4096	4096

Table 10: LLMs settings

Output: label (0|1)
Coref: We add a classification head on top of PLMs and fine-tune as a sequence classification task. Input: text[SEP]pronoun[SEP]candidates. Ouput: label (0|1)
NER: We add a classification head on top of PLMs and fine-tune as a sequence labeling task. Input: Text, Output: Label (CoNLL format)

Fine-tuning details are in Table 11.

Model	BERT	GPT-2	T5
Checkpoint	bert-base-cased	gpt2	t5-base
Param	110M	124M	220M
Optimizer	AdamW	AdamW	AdamW
Loss	CrossEntropyLoss	CrossEntropyLoss	CrossEntropyLoss
SA			
Batch size	16	32	32
Epoch	3	5	10
Learning rate	2e-5	2e-5	5e-5
Max length	128	128	128
Dialog			
Batch size	8	8	8
Training step	10000	10000	10000
Learning rate	2e-5	2e-5	2e-5
Max length	512	512	512
Coref			
Batch size	16	16	16
Epoch	3	3	3
Learning rate	2e-5	2e-5	2e-5
Max length	128	128	128
NER			
Batch size	32	32	32
Epoch	5	5	5
Learning rate	2e-5	2e-5	2e-5
Max length	128	128	128

Table 11: PLM fine-tuning details

A.6 Explicitly making model aware of the modification

We analyze whether making the model aware of the intended modification would help them become more robust. For this experiment, we select GPT-5 model and add a single-line instruction at the beginning of the prompt (e.g. "The following text has been transformed from active to passive voice"). Then, we observe the performance change between GPT-5 vs. GPT-5 (w. context): ΔU = U(ctx) - U(std). Overall, there are no consistent gains across all tasks. We even observe degradation in IFEval, as the extra context add noise to instruction following capabilities (e.g. model confuse the extra context of "modified to contain negation" with the actual constraint that they must follow).

A.7 Use Case: NER and Coreference tasks (extended results)

We present the full results of NER and Coreference tasks with confidence interval in Table 13 and Table 14.

Category	Modification	SA	COREF	DIALOGUE	NER	GSM	IFEVAL	AVG
Bias	**Geographical**	1.0	4.0	-1.1	0.3	2.0	0.0	1.0
	Length	1.0	0.0	-1.0	-1.9	3.0	2.0	0.5
	Temporal	-1.0	-2.0	0.0	-2.3	1.0	2.0	-0.4
Orthography	**Capitalization**	0.0	-1.0	-4.2	-3.7	0.0	-2.0	-1.8
	Punctuation	0.0	-1.0	-2.0	-0.1	2.0	2.0	0.1
	Spelling	1.0	-4.1	3.0	-2.5	1.0	2.0	0.1
Morphology	**Compound**	0.0	0.0	-2.0	-2.7	NA	NA	-1.2
	Derivation	1.1	-2.0	3.2	-5.2	NA	NA	-0.7
Syntax	**Conjunction**	0.0	-1.0	1.0	-3.9	0.0	2.0	-0.3
	Grammar	0.0	-2.8	1.5	-3.7	NA	NA	-1.2
	Voice	-1.0	0.0	0.0	0.6	0.0	6.0	0.9
Semantics	**Concept**	0.0	3.0	-5.0	-3.4	-1.0	6.0	-0.1
	Negation	0.0	-2.0	3.0	-3.0	4.0	-5.0	-0.5
Discourse	**Appraisal**	0.0	-3.0	0.0	-2.7	0.0	1.0	-0.8
	Disc. markers	2.0	3.0	-3.4	1.0	NA	NA	0.6
Varieties	**Dialect**	-1.0	2.0	-1.9	0.2	1.0	-2.0	-0.3
	Style	0.0	-3.0	1.0	0.1	0.0	6.0	0.7
Average	**Average**	0.2	-0.6	-0.5	-1.9	1.0	1.5	-0.2

Table 12: GPT-5 vs GPT-5 (w. context): ΔU = U(ctx) - U(std) (flip %) by task and modification. Positive value (red) indicates a degradation in robustness while negative value (green) indicates an increase.

A.8 Use Case: Dialogue Contradiction Detection and Sentiment Analysis (extended results)

In this section we introduce two additional classification tasks – Dialogue Contradiction Detection and Sentiment Analysis.

We use the DECODE dataset (Nie et al., 2021) to evaluate dialogue understanding capabilities. Each sample is a multi-turn dialogue, where the final turn may be consistent with (0) or contradict (1) the preceding context. For Sentiment Analysis, We chose the SST-2 dataset (Socher et al., 2013), containing around 11K sentences extracted from movies reviews with either 0 (negative) or 1 (positive) sentiment.

As Table 15 shows, on Dialogue contradiction detection task base LLMs are overall more robust to modifications than PLMs; however, reasoning LLMs on average perform similar to LLMs. All models show major brittleness to negation, and the majority of them also lack robustness to change of **grammatical roles**, which shows that they tend to rely on surface-level, token consistency between the context and the final turn, rather than on logical coherence. Moreover, LLMs (but to a lesser extent PLMs) demonstrate **geographical bias**, whuile PLMs are particularly brittle to **casual style**.

On Sentiment Analysis task, overall, PLMs were less robust to modifications compared to LLMs (see Table 16), with the exception of Claude-3.5 which lacks robusteness to many tests that the other LLMs are stable on. Most prominently, all LLMs and PLMs are significantly unrobust to **negation**, and can be easily misguided by adding another **appraisal** marker to a sentence already bearing some sentiment.

Category	Modification	BERT	GPT-2 (PLM)	T5	GPT-4o	Claude-3.5	Llama 3.1 (LLM)	GPT-5	DS R1	Avg
Bias	**Temporal**	3.7 [1.4, 6.9]	3.0 [0.4, 6.4]	1.2 [0.3, 2.6]	6.2 [2.5, 10.8]	4.3 [1.1, 8.2]	1.8 [0.3, 4.0]	7.7 [4.4, 11.5]	10.6 [6.4, 15.4]	4.8 [2.1, 8.2]
	Geographical	25.1 [19.6, 30.9]	27.6 [21.7, 33.9]	29.0 [23.6, 34.5]	22.3 [17.5, 27.5]	26.0 [19.5, 32.8]	27.0 [21.5, 32.6]	22.0 [16.7, 27.7]	27.8 [22.7, 33.0]	25.8 [20.4, 31.6]
	Length	11.8 [6.6, 17.7]	12.2 [7.4, 17.7]	13.1 [8.5, 18.3]	12.7 [7.6, 18.5]	15.7 [9.3, 23.1]	6.4 [3.0, 10.5]	9.7 [5.2, 15.1]	12.6 [7.8, 18.0]	11.8 [6.9, 17.4]
Orthography	**Spelling**	4.3 [1.6, 7.7]	2.5 [0.4, 5.2]	1.4 [0.5, 2.5]	6.6 [3.8, 9.8]	3.0 [0.5, 6.2]	3.9 [1.6, 6.6]	7.3 [4.3, 10.8]	11.4 [7.3, 15.9]	5.0 [2.5, 8.1]
	Capitalization	0.9 [0.1, 1.8]	19.3 [13.1, 26.1]	13.1 [8.5, 18.2]	11.3 [6.6, 16.6]	8.9 [4.5, 14.1]	15.3 [9.7, 21.6]	9.1 [5.6, 13.1]	13.6 [9.1, 18.5]	11.5 [7.2, 16.3]
	Punctuation	6.5 [3.0, 10.7]	3.7 [1.3, 6.8]	4.9 [2.4, 7.8]	7.1 [3.6, 11.2]	9.1 [5.0, 13.7]	7.9 [4.1, 12.4]	7.8 [4.5, 11.7]	15.0 [10.3, 20.2]	7.8 [4.3, 11.8]
Morphology	**Derivation**	1.9 [0.5, 3.7]	4.2 [1.4, 7.5]	5.8 [2.1, 10.5]	3.7 [1.1, 7.2]	2.0 [0.1, 4.5]	1.3 [0.1, 3.2]	9.5 [4.4, 15.7]	8.2 [4.2, 12.9]	4.6 [1.7, 8.1]
	Compound	3.1 [0.8, 6.1]	0.6 [0.0, 1.7]	5.2 [2.0, 9.1]	3.1 [0.8, 6.5]	1.7 [0.0, 4.4]	1.0 [0.0, 2.7]	7.0 [3.4, 11.2]	12.9 [8.2, 18.1]	4.3 [1.9, 7.5]
Syntax	**Voice**	7.8 [3.9, 12.4]	10.8 [6.0, 16.2]	5.7 [3.1, 8.7]	7.5 [3.6, 12.2]	5.5 [2.1, 9.7]	4.5 [1.5, 8.2]	8.3 [4.3, 13.1]	11.3 [7.2, 16.0]	7.7 [4.0, 12.1]
	Grammar	8.0 [4.0, 12.9]	15.9 [10.4, 21.8]	10.5 [6.3, 15.5]	8.3 [4.7, 12.5]	3.5 [1.2, 6.6]	5.3 [1.8, 9.6]	10.8 [6.1, 16.1]	12.4 [8.1, 17.0]	9.3 [5.3, 14.0]
	Conjunction	9.1 [4.1, 15.4]	7.6 [4.3, 11.3]	7.7 [4.4, 11.4]	9.4 [5.6, 14.0]	7.5 [3.3, 12.9]	8.0 [3.4, 13.7]	11.7 [7.1, 17.1]	13.2 [8.4, 18.7]	9.3 [5.1, 14.3]
Semantics	**Concept**	5.0 [2.3, 8.4]	8.9 [4.3, 14.2]	5.3 [2.6, 8.7]	6.5 [3.2, 10.5]	4.8 [1.2, 9.3]	3.9 [1.2, 7.5]	8.5 [4.1, 13.7]	9.4 [5.7, 13.7]	6.5 [3.1, 10.7]
	Negation	4.8 [2.2, 8.0]	5.2 [2.9, 7.9]	6.5 [3.7, 9.9]	8.2 [4.8, 12.2]	3.9 [1.3, 7.0]	4.6 [2.1, 7.6]	10.5 [6.9, 14.7]	15.2 [10.3, 20.6]	7.4 [4.3, 11.0]
Discourse	**Disc. markers**	4.7 [2.0, 8.0]	1.6 [0.1, 3.6]	5.4 [2.8, 8.3]	2.2 [0.8, 3.9]	1.9 [0.4, 4.0]	3.2 [0.8, 6.1]	6.3 [3.4, 9.5]	8.7 [5.2, 12.9]	4.2 [1.9, 7.0]
	Appraisal	5.5 [2.9, 8.6]	2.6 [0.8, 5.0]	4.8 [2.4, 7.7]	3.2 [1.4, 5.3]	3.1 [1.0, 5.8]	2.9 [0.9, 5.5]	7.1 [4.2, 10.4]	14.0 [9.2, 19.2]	5.4 [2.9, 8.4]
Varieties	**Style**	11.8 [6.6, 17.7]	12.2 [7.4, 17.7]	13.1 [8.5, 18.3]	3.6 [1.8, 5.7]	4.8 [2.1, 8.0]	6.4 [2.9, 10.6]	7.4 [4.1, 11.2]	10.9 [6.9, 15.3]	8.8 [5.0, 13.1]
	Dialectal	12.6 [8.5, 17.3]	7.4 [4.1, 11.4]	8.4 [5.8, 11.4]	5.9 [3.2, 9.3]	8.9 [4.5, 14.0]	4.7 [2.2, 7.9]	7.0 [4.0, 10.6]	13.2 [8.4, 18.3]	8.5 [5.1, 12.5]
Average		7.4 [4.1, 11.4]	8.6 [5.1, 12.6]	8.3 [5.1, 12.0]	7.5 [4.3, 11.4]	6.7 [3.4, 10.8]	6.4 [3.4, 10.0]	9.3 [5.5, 13.7]	13.0 [8.6, 17.9]	8.4 [4.9, 12.5]

Table 13: NER: Unrobustness (U, %) by model and modification

Category	Modification	BERT	GPT-2 (PLM)	T5	GPT-4o	Claude-3.5	Llama 3.1 (LLM)	GPT-5	DS R1	Avg Avg
Bias	**Temporal**	9.0 [4.8, 16.2]	4.0 [1.6, 9.8]	6.0 [2.8, 12.5]	8.0 [4.1, 15.0]	1.0 [0.2, 5.4]	7.0 [3.4, 13.7]	3.0 [1.0, 8.5]	8.0 [4.1, 15.0]	5.8 [2.8, 12.0]
	Geographical	8.0 [4.1, 15.0]	14.0 [8.5, 22.1]	9.0 [4.8, 16.2]	10.0 [5.5, 17.4]	4.0 [1.6, 9.8]	5.0 [2.2, 11.2]	6.0 [2.8, 12.5]	10.0 [5.5, 17.4]	8.2 [4.4, 15.2]
	Length	19.2 [12.6, 28.0]	15.2 [9.4, 23.5]	14.1 [8.6, 22.3]	15.2 [9.4, 23.5]	20.2 [13.5, 29.2]	17.2 [11.0, 25.8]	8.1 [4.2, 15.1]	13.1 [7.8, 21.2]	15.3 [9.6, 23.6]
Orthography	**Spelling**	11.2 [6.4, 19.0]	3.1 [1.0, 8.6]	7.1 [3.5, 14.0]	5.1 [2.2, 11.4]	2.0 [0.6, 7.1]	4.1 [1.6, 10.0]	7.1 [3.5, 14.0]	4.1 [1.6, 10.0]	5.5 [2.5, 11.8]
	Capitalization	15.2 [9.4, 23.5]	14.1 [8.6, 22.3]	7.1 [3.5, 13.9]	7.1 [3.5, 13.9]	4.0 [1.6, 9.9]	5.1 [2.2, 11.3]	2.0 [0.6, 7.1]	6.1 [2.8, 12.6]	7.6 [4.0, 14.3]
	Punctuation	1.0 [0.2, 5.5]	7.1 [3.5, 13.9]	1.0 [0.2, 5.5]	1.0 [0.2, 5.5]	3.0 [1.0, 8.5]	1.0 [0.2, 5.5]	2.0 [0.6, 7.1]	4.0 [1.6, 9.9]	2.5 [0.9, 7.7]
Morphology	**Derivation**	4.1 [1.6, 10.0]	3.1 [1.0, 8.6]	5.1 [2.2, 11.4]	4.1 [1.6, 10.0]	1.0 [0.2, 5.6]	1.0 [0.2, 5.6]	5.1 [2.2, 11.4]	2.0 [0.6, 7.1]	3.2 [1.2, 8.7]
	Compound	6.2 [2.9, 13.0]	4.2 [1.6, 10.2]	3.1 [1.1, 8.8]	5.2 [2.2, 11.6]	6.2 [2.9, 13.0]	6.2 [2.9, 13.0]	3.1 [1.1, 8.8]	4.2 [1.6, 10.2]	4.8 [2.0, 11.1]
Syntax	**Voice**	35.8 [26.9, 45.8]	34.7 [25.9, 44.7]	41.1 [31.7, 51.1]	26.3 [18.5, 36.0]	15.8 [9.8, 24.4]	30.5 [22.2, 40.4]	9.5 [5.1, 17.0]	15.8 [9.8, 24.4]	26.2 [18.7, 35.5]
	Grammar	30.6 [21.1, 42.0]	27.8 [18.8, 39.0]	19.4 [12.0, 30.0]	22.2 [14.2, 33.1]	18.1 [10.9, 28.5]	19.4 [12.0, 30.0]	16.7 [9.8, 26.9]	20.8 [13.1, 31.6]	21.9 [14.0, 32.6]
	Conjunction	5.2 [2.2, 11.5]	10.3 [5.7, 17.9]	8.2 [4.2, 15.4]	8.2 [4.2, 15.4]	4.1 [1.6, 10.1]	7.2 [3.5, 14.2]	5.2 [2.2, 11.5]	6.2 [2.9, 12.8]	6.8 [3.3, 13.6]
Semantics	**Concept**	8.0 [4.1, 15.0]	3.0 [1.0, 8.5]	5.0 [2.2, 11.2]	11.0 [6.3, 18.6]	18.0 [11.7, 26.7]	9.0 [4.8, 16.2]	10.0 [5.5, 17.4]	10.0 [5.5, 17.4]	9.2 [5.1, 16.4]
	Negation	25.5 [17.9, 35.0]	23.5 [16.2, 32.8]	24.5 [17.0, 33.9]	21.4 [14.5, 30.5]	24.5 [17.0, 33.9]	22.4 [15.3, 31.7]	22.4 [15.3, 31.7]	30.6 [22.4, 40.3]	24.4 [17.0, 33.7]
Discourse	**Disc. markers**	8.0 [4.1, 15.0]	6.0 [2.8, 12.5]	8.0 [4.1, 15.0]	19.0 [12.5, 27.8]	10.0 [5.5, 17.4]	7.0 [3.4, 13.7]	9.0 [4.8, 16.2]	11.0 [6.3, 18.6]	9.8 [5.4, 17.0]
	Appraisal	5.0 [2.2, 11.2]	8.0 [4.1, 15.0]	4.0 [1.6, 9.8]	8.0 [4.1, 15.0]	7.0 [3.4, 13.7]	9.0 [4.8, 16.2]	5.0 [2.2, 11.2]	7.0 [3.4, 13.7]	6.6 [3.2, 13.2]
Varieties	**Style**	16.0 [10.1, 24.4]	19.0 [12.5, 27.8]	18.0 [11.7, 26.7]	19.0 [12.5, 27.8]	14.0 [8.5, 22.1]	14.0 [8.5, 22.1]	14.0 [8.5, 22.1]	12.0 [7.0, 19.8]	15.8 [9.9, 24.1]
	Dialect	8.3 [4.4, 15.1]	26.9 [19.4, 35.9]	22.2 [15.4, 30.9]	24.5 [17.2, 33.7]	11.8 [6.9, 19.4]	16.7 [10.7, 25.1]	2.9 [1.0, 8.3]	12.7 [7.6, 20.6]	15.8 [10.3, 23.6]
Average		12.7 [7.9, 20.3]	13.2 [8.3, 20.8]	11.9 [7.4, 19.3]	12.7 [7.8, 20.4]	9.7 [5.7, 16.8]	10.7 [6.4, 18.0]	7.7 [4.1, 14.5]	10.5 [6.1, 17.8]	11.1 [6.7, 18.5]

Table 14: Coreference: Unrobustness (U, %) by model and modification

A.9 Use Case: GSM and IFEval (extended results)

We present the full results of GSM and IFEval tasks with CI intervals in Table 17 and Table 18.

A.10 Scaling analysis (extended results)

We present the scaling analysis results with confidence intereval in Table 19.

A.11 Unrobustness results for each model averaged across tasks

Table 20 reports the Unrobustness scores for each model averaged across all tasks. Note that IFEval and GSM tasks only contain LLMs results. Overall, PLMs are more unrobust compared to LLMs. Among LLMs, the newer reasoning models are more robust compared to the base versions. Modification-wise, Geographical Bias, Negation, Style, and Dialect are the most prominent ones in creating instability.

A.12 AI usage disclosure

Claude-4-Sonnet (through Claude Code) was used to help debug issues in the code and help generating the tables in LaTex.

Category	Modification	BERT	GPT-2	T5	GPT-4o	Claude-3.5	Llama 3.1	GPT-5	DS R1	Avg
Bias	**Temporal**	8.0 [4.1, 15.0]	6.0 [2.8, 12.5]	7.0 [3.4, 13.7]	2.0 [0.6, 7.0]	5.0 [2.2, 11.2]	1.0 [0.2, 5.4]	3.0 [1.0, 8.5]	2.0 [0.6, 7.0]	4.2 [1.8, 10.0]
	Geographical	6.5 [3.0, 13.5]	8.7 [4.5, 16.2]	12.0 [6.8, 20.2]	13.0 [7.6, 21.4]	15.2 [9.3, 23.9]	12.0 [6.8, 20.2]	18.5 [11.9, 27.6]	13.0 [7.6, 21.4]	12.4 [7.2, 20.6]
	Length	10.0 [5.5, 17.4]	6.0 [2.8, 12.5]	5.0 [2.2, 11.2]	4.0 [1.6, 9.8]	6.0 [2.8, 12.5]	1.0 [0.2, 5.4]	6.0 [2.8, 12.5]	7.0 [3.4, 13.7]	5.6 [2.6, 11.9]
Orthographic	**Spelling**	6.0 [2.8, 12.5]	6.0 [2.8, 12.5]	4.0 [1.6, 9.8]	8.0 [4.1, 15.0]	3.0 [1.0, 8.5]	2.0 [0.6, 7.0]	1.0 [0.2, 5.4]	6.0 [2.8, 12.5]	4.5 [2.0, 10.4]
	Capitalization	0.0 [0.0, 0.0]	6.2 [2.9, 13.0]	6.2 [2.9, 13.0]	3.1 [1.1, 8.8]	2.1 [0.6, 7.3]	2.1 [0.6, 7.3]	4.2 [1.6, 10.2]	1.0 [0.2, 5.7]	3.1 [1.2, 8.1]
	Punctuation	4.0 [1.6, 9.8]	2.0 [0.6, 7.0]	8.0 [4.1, 15.0]	3.0 [1.0, 8.5]	3.0 [1.0, 8.5]	1.0 [0.2, 5.4]	5.0 [2.2, 11.2]	2.0 [0.6, 7.0]	3.5 [1.4, 9.0]
Semantic	**Concept**	7.0 [3.4, 13.7]	5.0 [2.2, 11.2]	4.0 [1.6, 9.8]	8.0 [4.1, 15.0]	7.0 [3.4, 13.7]	2.0 [0.6, 7.0]	10.0 [5.5, 17.4]	6.0 [2.8, 12.5]	6.1 [2.9, 12.6]
	Negation	29.0 [21.0, 38.5]	39.0 [30.0, 48.8]	31.0 [22.8, 40.6]	31.0 [22.8, 40.6]	30.0 [21.9, 39.6]	30.0 [21.9, 39.6]	25.0 [17.5, 34.3]	25.0 [17.5, 34.3]	30.0 [21.9, 39.5]
Pragmatic	**Appraisal**	6.0 [2.8, 12.5]	10.0 [5.5, 17.4]	5.0 [2.2, 11.2]	7.0 [3.4, 13.7]	6.0 [2.8, 12.5]	4.0 [1.6, 9.8]	10.0 [5.5, 17.4]	11.0 [6.3, 18.6]	7.4 [3.8, 14.2]
Genre	**Style**	12.0 [7.0, 19.8]	10.0 [5.5, 17.4]	12.0 [7.0, 19.8]	6.0 [2.8, 12.5]	4.0 [1.6, 9.8]	2.0 [0.6, 7.0]	6.0 [2.8, 12.5]	6.0 [2.8, 12.5]	7.2 [3.7, 13.9]
	Dialect	20.8 [14.1, 29.4]	5.7 [2.6, 11.8]	9.4 [5.2, 16.5]	11.3 [6.6, 18.8]	15.1 [9.5, 23.1]	7.5 [3.9, 14.2]	5.7 [2.6, 11.8]	13.2 [8.0, 21.0]	11.1 [6.6, 18.3]
Syntactic	**Conjunction**	5.0 [2.2, 11.2]	4.0 [1.6, 9.8]	7.0 [3.4, 13.7]	6.0 [2.8, 12.5]	1.0 [0.2, 5.4]	2.0 [0.6, 7.0]	4.0 [1.6, 9.8]	4.0 [1.6, 9.8]	4.1 [1.7, 9.9]
	Voice	9.0 [4.8, 16.2]	11.0 [6.3, 18.6]	7.0 [3.4, 13.7]	7.0 [3.4, 13.7]	2.0 [0.6, 7.0]	3.0 [1.0, 8.5]	9.0 [4.8, 16.2]	5.0 [2.2, 11.2]	6.6 [3.3, 13.2]
	Role	14.3 [7.9, 24.3]	12.9 [6.9, 22.7]	7.1 [3.1, 15.7]	20.6 [12.7, 31.6]	7.4 [3.2, 16.1]	11.8 [6.1, 21.5]	16.2 [9.3, 26.7]	19.1 [11.5, 30.0]	13.7 [7.6, 23.6]
Morphological	**Derivation**	0.0 [0.0, 0.0]	2.2 [0.6, 7.5]	1.1 [0.2, 5.8]	4.3 [1.7, 10.5]	5.4 [2.3, 12.0]	1.1 [0.2, 5.8]	5.4 [2.3, 12.0]	6.5 [3.0, 13.4]	3.2 [1.3, 8.4]
	Compound	6.0 [2.8, 12.5]	4.0 [1.6, 9.8]	3.0 [1.0, 8.5]	4.0 [1.6, 9.8]	7.0 [3.4, 13.7]	1.0 [0.2, 5.4]	8.0 [4.1, 15.0]	6.0 [2.8, 12.5]	4.9 [2.2, 10.9]
	Discourse	5.7 [2.5, 12.8]	9.2 [4.7, 17.1]	3.4 [1.2, 9.7]	3.4 [1.2, 9.7]	2.3 [0.6, 8.0]	3.4 [1.2, 9.7]	5.7 [2.5, 12.8]	5.7 [2.5, 12.8]	4.9 [2.0, 11.5]
Average		8.8 [5.0, 15.2]	8.7 [4.9, 15.6]	7.8 [4.2, 14.6]	8.3 [4.6, 15.2]	7.1 [3.9, 13.7]	5.1 [2.7, 11.0]	8.4 [4.6, 15.4]	8.2 [4.5, 15.0]	7.8 [4.3, 14.5]

Table 15: Dialogue: Unrobustness (U, %) by model and modification

Category	Modification	PLM			LLM					
		BERT	**GPT-2**	**T5**	**GPT-4o**	**Claude-3.5**	**Llama 3.1**	**GPT-5**	**DS R1**	**Avg**
Bias	**Temporal**	5.0 [2.2, 11.2]	5.0 [2.2, 11.2]	3.0 [1.0, 8.5]	0.0 [0.0, 0.0]	8.0 [4.1, 15.0]	1.0 [0.2, 5.4]	3.0 [1.0, 8.5]	3.0 [1.0, 8.5]	3.5 [1.5, 8.5]
	Geographical	5.0 [2.2, 11.2]	4.0 [1.6, 9.8]	3.0 [1.0, 8.5]	9.0 [4.8, 16.2]	8.0 [4.1, 15.0]	7.0 [3.4, 13.7]	4.0 [1.6, 9.8]	8.0 [4.1, 15.0]	6.0 [2.8, 12.4]
	Length	3.0 [1.0, 8.5]	2.0 [0.6, 7.0]	3.0 [1.0, 8.5]	3.0 [1.0, 8.5]	9.0 [4.8, 16.2]	2.0 [0.6, 7.0]	1.0 [0.2, 5.4]	1.0 [0.2, 5.4]	3.0 [1.2, 8.3]
Orthographic	**Spelling**	4.0 [1.6, 9.8]	4.0 [1.6, 9.8]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	6.0 [2.8, 12.5]	3.0 [1.0, 8.5]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	2.1 [0.9, 5.1]
	Capitalization	0.0 [0.0, 0.0]	3.0 [1.0, 8.5]	3.0 [1.0, 8.5]	1.0 [0.2, 5.5]	7.1 [3.5, 13.9]	3.0 [1.0, 8.5]	0.0 [0.0, 0.0]	3.0 [1.0, 8.5]	2.5 [1.0, 6.7]
	Punctuation	1.0 [0.2, 5.4]	2.0 [0.6, 7.0]	0.0 [0.0, 0.0]	1.0 [0.2, 5.4]	3.0 [1.0, 8.5]	3.0 [1.0, 8.5]	1.0 [0.2, 5.4]	1.0 [0.2, 5.4]	1.5 [0.4, 5.7]
Morphological	**Derivation**	3.4 [1.2, 9.7]	5.7 [2.5, 12.8]	2.3 [0.6, 8.0]	2.3 [0.6, 8.0]	6.9 [3.2, 14.2]	2.3 [0.6, 8.0]	3.4 [1.2, 9.7]	4.6 [1.8, 11.2]	3.9 [1.5, 10.2]
	Compound	3.2 [1.1, 8.9]	8.4 [4.3, 15.7]	3.2 [1.1, 8.9]	2.1 [0.6, 7.4]	5.3 [2.3, 11.7]	1.1 [0.2, 5.7]	3.2 [1.1, 8.9]	4.2 [1.6, 10.3]	3.8 [1.5, 9.7]
Syntactic	**Voice**	9.0 [4.8, 16.2]	10.0 [5.5, 17.4]	9.0 [4.8, 16.2]	4.0 [1.6, 9.8]	10.0 [5.5, 17.4]	3.0 [1.0, 8.5]	5.0 [2.2, 11.2]	4.0 [1.6, 9.8]	6.8 [3.4, 13.3]
	Grammar	4.5 [1.6, 12.5]	12.1 [6.3, 22.1]	3.0 [0.8, 10.4]	4.5 [1.6, 12.5]	7.6 [3.3, 16.5]	4.5 [1.6, 12.5]	4.5 [1.6, 12.5]	4.5 [1.6, 12.5]	5.7 [2.3, 14.0]
	Conjunction	3.0 [1.0, 8.5]	6.0 [2.8, 12.5]	1.0 [0.2, 5.4]	1.0 [0.2, 5.4]	3.0 [1.0, 8.5]	3.0 [1.0, 8.5]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	2.1 [0.8, 6.1]
Semantic	**Concept**	6.0 [2.8, 12.5]	6.0 [2.8, 12.5]	5.0 [2.2, 11.2]	4.0 [1.6, 9.8]	4.0 [1.6, 9.8]	2.0 [0.6, 7.0]	1.0 [0.2, 5.4]	5.0 [2.2, 11.2]	4.1 [1.7, 9.9]
	Negation	22.9 [15.6, 32.3]	20.8 [13.9, 30.0]	25.0 [17.4, 34.5]	16.7 [10.5, 25.4]	17.7 [11.4, 26.5]	15.6 [9.7, 24.2]	16.7 [10.5, 25.4]	17.7 [11.4, 26.5]	19.1 [12.6, 28.1]
Pragmatic	**Discourse**	3.0 [1.0, 8.5]	6.1 [2.8, 12.6]	4.0 [1.6, 9.9]	1.0 [0.2, 5.5]	12.1 [7.1, 20.0]	3.0 [1.0, 8.5]	2.0 [0.6, 7.1]	3.0 [1.0, 8.5]	4.3 [1.9, 10.1]
	Sentiment	19.0 [12.5, 27.8]	16.0 [10.1, 24.4]	12.0 [7.0, 19.8]	14.0 [8.5, 22.1]	15.0 [9.3, 23.3]	16.0 [10.1, 24.4]	11.0 [6.3, 18.6]	10.0 [5.5, 17.4]	14.1 [8.7, 22.2]
Genre	**Casual**	9.0 [4.8, 16.2]	9.0 [4.8, 16.2]	5.0 [2.2, 11.2]	3.0 [1.0, 8.5]	7.0 [3.4, 13.7]	6.0 [2.8, 12.5]	3.0 [1.0, 8.5]	2.0 [0.6, 7.0]	5.5 [2.6, 11.7]
	Dialectal	9.8 [5.4, 17.1]	9.8 [5.4, 17.1]	7.8 [4.0, 14.7]	4.9 [2.1, 11.0]	6.9 [3.4, 13.5]	5.9 [2.7, 12.2]	2.9 [1.0, 8.3]	3.9 [1.5, 9.7]	6.5 [3.2, 12.9]
Average		6.5 [3.5, 12.7]	7.6 [4.0, 14.5]	5.3 [2.7, 10.8]	4.2 [2.0, 9.5]	8.0 [4.2, 15.1]	4.8 [2.3, 10.8]	3.6 [1.7, 8.5]	4.4 [2.1, 9.8]	5.6 [2.8, 11.5]

Table 16: Sentiment Analysis: Unrobustness (U, %) by model and modification

Category	Modification	GPT-4o	Claude-3.5	Llama 3.1	GPT-5	DS R1	Avg
Bias	**Temporal**	1.0 [0.2, 5.4]	2.0 [0.6, 7.0]	4.0 [1.6, 9.8]	0.0 [0.0, 0.0]	1.0 [0.2, 5.4]	1.6 [0.5, 5.5]
	Geographical	5.0 [2.2, 11.2]	5.0 [2.2, 11.2]	7.0 [3.4, 13.7]	1.0 [0.2, 5.4]	2.0 [0.6, 7.0]	4.0 [1.7, 9.7]
	Length	4.0 [1.6, 9.8]	2.0 [0.6, 7.0]	2.0 [0.6, 7.0]	0.0 [0.0, 0.0]	2.0 [0.6, 7.0]	2.0 [0.6, 6.2]
Orthographic	**Spelling**	1.0 [0.2, 5.4]	0.0 [0.0, 0.0]	2.0 [0.6, 7.0]	1.0 [0.2, 5.4]	0.0 [0.0, 0.0]	0.8 [0.2, 3.6]
	Capitalization	3.0 [1.0, 8.5]	1.0 [0.2, 5.4]	5.0 [2.2, 11.2]	1.0 [0.2, 5.4]	1.0 [0.2, 5.4]	2.2 [0.7, 7.2]
	Punctuation	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	5.0 [2.2, 11.2]	1.0 [0.2, 5.4]	1.0 [0.2, 5.4]	1.4 [0.5, 4.4]
Semantic	**Concept**	1.0 [0.2, 5.4]	0.0 [0.0, 0.0]	5.0 [2.2, 11.2]	2.0 [0.6, 7.0]	1.0 [0.2, 5.4]	1.8 [0.6, 5.8]
	Negation	15.0 [9.3, 23.3]	15.0 [9.3, 23.3]	18.0 [11.7, 26.7]	7.0 [3.4, 13.7]	9.0 [4.8, 16.2]	12.8 [7.7, 20.6]
Discourse	**Appraisal**	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	1.0 [0.2, 5.4]	1.0 [0.2, 5.4]	1.0 [0.2, 5.4]	0.6 [0.1, 3.3]
Varieties	**Style**	4.0 [1.6, 9.8]	3.0 [1.0, 8.5]	6.0 [2.8, 12.5]	4.0 [1.6, 9.8]	3.0 [1.0, 8.5]	4.0 [1.6, 9.8]
	Dialect	2.0 [0.6, 7.0]	4.0 [1.6, 9.8]	6.0 [2.8, 12.5]	1.0 [0.2, 5.4]	2.0 [0.6, 7.0]	3.0 [1.1, 8.4]
Syntactic	**Conjunction**	0.0 [0.0, 0.0]	1.0 [0.2, 5.4]	4.0 [1.6, 9.8]	1.0 [0.2, 5.4]	1.0 [0.2, 5.4]	1.4 [0.4, 5.2]
	Voice	2.0 [0.6, 7.0]	2.0 [0.6, 7.0]	6.0 [2.8, 12.5]	1.0 [0.2, 5.4]	2.0 [0.6, 7.0]	2.6 [0.9, 7.8]
Average		2.9 [1.3, 7.1]	2.7 [1.2, 6.5]	5.5 [2.6, 11.6]	1.6 [0.5, 5.7]	2.0 [0.7, 6.6]	2.9 [1.3, 7.5]

Table 17: GSM: Unrobustness (U, %) by model and modification

Category	Modification	GPT-4o	Claude-3.5	Llama 3.1	GPT-5	DS R1	Avg
Bias	**Temporal**	11.1 [6.3, 18.8]	3.0 [1.0, 8.5]	11.1 [6.3, 18.8]	5.1 [2.2, 11.3]	12.1 [7.1, 20.0]	8.5 [4.6, 15.5]
	Geographical	9.0 [4.8, 16.2]	12.0 [7.0, 19.8]	14.0 [8.5, 22.1]	9.0 [4.8, 16.2]	10.0 [5.5, 17.4]	10.8 [6.1, 18.4]
	Length	11.0 [6.3, 18.6]	10.0 [5.5, 17.4]	17.0 [10.9, 25.5]	6.0 [2.8, 12.5]	16.0 [10.1, 24.4]	12.0 [7.1, 19.7]
Orthography	**Capitalization**	8.1 [4.2, 15.1]	5.1 [2.2, 11.3]	10.1 [5.6, 17.6]	9.1 [4.9, 16.4]	14.1 [8.6, 22.3]	9.3 [5.1, 16.6]
	Punctuation	7.1 [3.5, 13.9]	3.0 [1.0, 8.5]	13.1 [7.8, 21.2]	8.1 [4.2, 15.1]	12.1 [7.1, 20.0]	8.7 [4.7, 15.7]
	Spelling	4.1 [1.6, 10.1]	4.1 [1.6, 10.1]	7.2 [3.5, 14.2]	6.2 [2.9, 12.8]	8.2 [4.2, 15.4]	6.0 [2.8, 12.5]
Syntax	**Conjunction**	11.0 [6.3, 18.6]	8.0 [4.1, 15.0]	6.0 [2.8, 12.5]	8.0 [4.1, 15.0]	11.0 [6.3, 18.6]	8.8 [4.7, 15.9]
	Voice	9.0 [4.8, 16.2]	9.0 [4.8, 16.2]	11.0 [6.3, 18.6]	6.0 [2.8, 12.5]	15.0 [9.3, 23.3]	10.0 [5.6, 17.4]
Semantics	**Concept**	9.1 [4.9, 16.4]	7.1 [3.5, 13.9]	11.1 [6.3, 18.8]	7.1 [3.5, 13.9]	8.1 [4.2, 15.1]	8.5 [4.5, 15.6]
	Negation	25.0 [17.5, 34.3]	22.0 [15.0, 31.1]	24.0 [16.7, 33.2]	23.0 [15.8, 32.2]	24.0 [16.7, 33.2]	23.6 [16.4, 32.8]
Discourse	**Appraisal**	7.1 [3.5, 14.0]	4.1 [1.6, 10.0]	12.2 [7.1, 20.2]	7.1 [3.5, 14.0]	6.1 [2.8, 12.7]	7.3 [3.7, 14.2]
Varieties	**Style**	18.0 [11.7, 26.7]	7.0 [3.4, 13.7]	12.0 [7.0, 19.8]	7.0 [3.4, 13.7]	11.0 [6.3, 18.6]	11.0 [6.4, 18.5]
	Dialect	12.0 [7.0, 19.8]	13.0 [7.8, 21.0]	12.0 [7.0, 19.8]	11.0 [6.3, 18.6]	15.0 [9.3, 23.3]	12.6 [7.5, 20.5]
Average		10.9 [6.3, 18.4]	8.3 [4.5, 15.1]	12.4 [7.4, 20.2]	8.7 [4.7, 15.7]	12.5 [7.5, 20.4]	10.5 [6.1, 17.9]

Table 18: IFEval: Unrobustness (U, %) by model and modification

Category	Modification	Llama-8b	Llama-70b	Llama-405b	Avg
Bias	**Temporal**	8.0 [4.1, 15.0]	5.0 [2.2, 11.2]	4.0 [1.6, 9.8]	5.7 [2.6, 12.0]
	Geographical	8.0 [4.1, 15.0]	3.0 [1.0, 8.5]	7.0 [3.4, 13.7]	6.0 [2.9, 12.4]
	Length	8.0 [4.1, 15.0]	2.0 [0.6, 7.0]	2.0 [0.6, 7.0]	4.0 [1.7, 9.7]
Orthographic	**Spelling**	10.0 [5.5, 17.4]	2.0 [0.6, 7.0]	2.0 [0.6, 7.0]	4.7 [2.2, 10.5]
	Capitalization	11.0 [6.3, 18.6]	2.0 [0.6, 7.0]	5.0 [2.2, 11.2]	6.0 [3.0, 12.3]
	Punctuation	5.0 [2.2, 11.2]	1.0 [0.2, 5.4]	5.0 [2.2, 11.2]	3.7 [1.5, 9.3]
Semantic	**Concept**	12.0 [7.0, 19.8]	1.0 [0.2, 5.4]	5.0 [2.2, 11.2]	6.0 [3.1, 12.1]
	Negation	22.0 [15.0, 31.1]	17.0 [10.9, 25.5]	18.0 [11.7, 26.7]	19.0 [12.5, 27.8]
Discourse	**Appraisal**	7.0 [3.4, 13.7]	5.0 [2.2, 11.2]	1.0 [0.2, 5.4]	4.3 [1.9, 10.1]
Varieties	**Style**	16.0 [10.1, 24.4]	6.0 [2.8, 12.5]	6.0 [2.8, 12.5]	9.3 [5.2, 16.5]
	Dialect	10.0 [5.5, 17.4]	6.0 [2.8, 12.5]	6.0 [2.8, 12.5]	7.3 [3.7, 14.1]
Syntactic	**Conjunction**	12.0 [7.0, 19.8]	2.0 [0.6, 7.0]	4.0 [1.6, 9.8]	6.0 [3.0, 12.2]
	Voice	11.0 [6.3, 18.6]	2.0 [0.6, 7.0]	6.0 [2.8, 12.5]	6.3 [3.2, 12.7]
Average		10.8 [6.2, 18.2]	4.2 [1.9, 9.8]	5.5 [2.6, 11.6]	6.8 [3.6, 13.2]

Table 19: GSM Scaling: Unrobustness (U, %) by Llama model size and modification

Category	Modification	BERT	GPT-2	T5	GPT-4o	Claude-3.5	Llama 3.1	GPT-5	DS R1	Avg
Bias	**Temporal**	6.4 [3.1, 12.3]	4.5 [1.8, 10.0]	4.3 [1.9, 9.3]	4.7 [2.3, 9.5]	3.9 [1.5, 9.2]	4.3 [2.0, 9.5]	3.6 [1.6, 8.0]	6.1 [3.2, 11.9]	4.7 [2.2, 10.0]
	Geographical	11.2 [7.2, 17.6]	13.6 [9.1, 20.5]	13.2 [9.1, 19.9]	11.4 [7.1, 18.3]	11.7 [7.3, 18.8]	12.0 [7.6, 18.9]	10.1 [6.3, 16.5]	11.8 [7.7, 18.5]	11.9 [7.7, 18.6]
	Length	11.0 [6.4, 17.9]	8.8 [5.0, 15.2]	8.8 [5.1, 15.1]	8.3 [4.6, 14.8]	10.5 [6.1, 17.6]	7.6 [4.4, 13.5]	5.1 [2.5, 10.1]	8.6 [5.0, 15.0]	8.6 [4.9, 14.9]
Orthographic	**Spelling**	5.0 [2.2, 11.2]	5.0 [2.2, 11.2]	2.0 [0.8, 4.9]	3.0 [1.4, 6.8]	3.0 [1.3, 7.0]	2.3 [0.7, 7.5]	0.7 [0.1, 3.6]	2.0 [0.9, 4.2]	2.9 [1.2, 7.0]
	Capitalization	0.0 [0.0, 0.0]	4.6 [1.9, 10.8]	4.6 [1.9, 10.8]	2.4 [0.8, 7.6]	3.4 [1.4, 8.9]	3.4 [1.3, 9.0]	1.7 [0.6, 5.2]	1.7 [0.5, 6.5]	2.7 [1.1, 7.3]
	Punctuation	2.5 [0.9, 7.6]	2.0 [0.6, 7.0]	4.0 [2.0, 7.5]	1.3 [0.4, 4.6]	2.0 [0.7, 5.7]	3.0 [1.1, 8.4]	2.3 [0.9, 7.3]	1.3 [0.3, 5.9]	2.3 [0.9, 6.8]
Semantic	**Concept**	6.5 [3.1, 13.1]	5.5 [2.5, 11.8]	4.5 [1.9, 10.5]	4.3 [2.0, 10.1]	3.7 [1.7, 7.8]	3.0 [1.1, 8.4]	4.3 [2.1, 9.9]	4.0 [1.7, 9.7]	4.5 [2.0, 10.2]
	Negation	25.9 [18.3, 35.4]	29.9 [21.9, 39.4]	28.0 [20.1, 37.5]	20.9 [14.2, 29.8]	20.9 [14.2, 29.8]	21.2 [14.4, 30.2]	16.2 [10.5, 24.5]	17.2 [11.2, 25.7]	22.5 [15.6, 31.5]
Discourse	**Appraisal**	5.2 [2.5, 9.9]	5.3 [2.4, 10.0]	4.4 [2.0, 8.8]	4.6 [2.2, 8.6]	3.5 [1.5, 7.4]	6.3 [3.2, 11.8]	5.0 [2.5, 10.2]	7.0 [3.9, 12.8]	5.2 [2.6, 9.9]
Varieties	**Style**	13.9 [8.3, 21.0]	15.6 [9.9, 22.8]	15.6 [10.1, 22.5]	11.2 [6.9, 17.5]	7.2 [3.8, 13.1]	9.6 [5.3, 16.2]	8.1 [4.4, 14.2]	9.2 [5.3, 15.6]	11.3 [6.8, 17.9]
	Dialect	8.3 [4.4, 15.1]	26.9 [19.4, 35.9]	22.2 [15.4, 30.9]	12.8 [8.3, 20.2]	9.6 [5.4, 16.7]	11.6 [6.8, 19.1]	5.0 [2.5, 10.8]	9.9 [5.8, 17.0]	13.3 [8.5, 20.7]
Syntactic	**Conjunction**	4.0 [1.6, 9.8]	5.0 [2.2, 11.2]	4.0 [1.8, 9.6]	2.3 [1.0, 6.0]	1.7 [0.5, 6.4]	3.0 [1.1, 8.4]	1.7 [0.6, 5.1]	1.7 [0.6, 5.1]	2.9 [1.2, 7.7]
	Voice	9.0 [4.8, 16.2]	10.5 [5.9, 18.0]	8.0 [4.1, 14.9]	4.3 [1.9, 10.2]	4.7 [2.2, 10.5]	4.0 [1.6, 9.8]	5.0 [2.4, 10.9]	3.7 [1.5, 9.3]	6.1 [3.0, 12.5]
Average	8.4 [4.8, 14.4]	10.6 [6.5, 17.2]	9.5 [5.9, 15.5]	7.0 [4.1, 12.6]	6.6 [3.7, 12.2]	7.0 [3.9, 13.1]	5.3 [2.9, 10.5]	6.5 [3.7, 12.1]	7.6 [4.4, 13.5]	

Table 20: All Tasks: Unrobustness (U, %) by model and modification (averaged across tasks)