FACTEVAL: Evaluating the Robustness of Fact Verification Systems in the Era of Large Language Models

Mamta and **Oana Cocarascu**
King's College London
{mamta.name, oana.cocarascu}@kcl.ac.uk

Abstract

Whilst large language models (LLMs) have made significant advances in every natural language processing task, studies have shown that these models are vulnerable to small perturbations in the inputs, raising concerns about their robustness in the real-world. Given the rise of misinformation online and its significant impact on society, fact verification is one area in which assessing the robustness of models developed for this task is crucial. However, the robustness of LLMs in fact verification remains largely unexplored. In this paper, we introduce FACTEVAL, a novel large-scale benchmark for extensive evaluation of LLMs in the fact verification domain covering 16 realistic word-level and character-level perturbations and 4 types of subpopulations. We investigate the robustness of several LLMs in zero-shot, few-shot, and chain-of-thought prompting. Our analysis using FEVER, one of the largest and most widely-used datasets for fact verification, reveals that LLMs are brittle to small input changes and also exhibit performance variations across different subpopulations.

1 Introduction

Large language models (LLMs) have marked a significant milestone in the natural language processing (NLP) field. However, despite their impressive capabilities across a variety of tasks (Cao et al., 2023; Zhang et al., 2024; Li et al., 2024a), it has been shown that these models exhibit vulnerabilities to adversarial samples – subtle alterations to the input that can be easily understood by humans (Wang et al., 2023). This has raised concerns about the robustness of these models when deployed in the real world.

In the era where vast amounts of information are generated and shared rapidly, fact verification has become an increasingly important area of research in NLP as it is one of the key methods for detecting misleading and false claims Guo et al. (2022); Akhtar et al. (2023). Given the widespread influence of misinformation, which can have significant political, societal, and economic impact, assessing the robustness of fact verification systems and having a comprehensive understanding of these models' failures becomes crucial.

Numerous efforts have been made to evaluate the robustness of NLP models in a variety of tasks, including sentiment analysis and natural language inference (Wang et al., 2022; Mamta and Ekbal, 2022; Goyal et al., 2023). While there have been some studies examining adversarial attacks on fact verification models by manipulating either the claims (Thorne et al., 2019; Hidey et al., 2020; Atanasova et al., 2020) or the evidence (Abdelnabi and Fritz, 2023), their scope remains limited to traditional task-specific fine-tuned models.

In this paper, we propose FACTEVAL[1], a novel benchmark for evaluating the robustness of LLMs for fact verification. To the best of our knowledge, we are the first to study LLM-based fact verification models; prior studies have been limited to traditional BERT style models.

FACTEVAL evaluates the robustness of fact verification models through *perturbations* and *subpopulations* applied to the claims, leaving the evidence unchanged. This allows us to focus on evaluating the model's ability to handle variations in the claims without adding noise to the evidence.

We introduce 16 word-level and character-level perturbations used to create adversarial datasets. We incorporate visual perturbations (Boucher et al., 2022; Eger et al., 2019), expanding the scope of attacks on fact verification models. Recognizing that regional accents and the influence of native languages impact the writing style of non-native English speakers in informal conversations (Le et al., 2022), we integrate phonetic perturbations for a comprehensive evaluation.

[1]Code and data available at https://github.com/TRAI-group/FactEval

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 10647–10660

Furthermore, we designed subpopulations based on demographic factors, temporal information, sentiment, and named entities. These subpopulations consist of distinct subgroups of data that may differ in specific properties. Subpopulation analysis helps ensure that the model is not only accurate on the overall dataset but also performs consistently across all subgroups. This is not only important for fairness but also crucial for enhancing the overall reliability and effectiveness of the systems.

With FACTEVAL, we examine the robustness of LLM models for fact verification under zero-shot, few-shot, and chain-of-thought prompting. We evaluate three LLMs: Llama3 8B (Dubey et al., 2024), Mistral 7B (Jiang et al., 2023), and Gemma 7B (Mesnard et al., 2024). For comparison, we also investigate the robustness of transformer-based pre-trained language models (PLMs) such as BERT (Devlin et al., 2019).

Our experiments on the FEVER dataset, one of the largest and most widely-used benchmarks for fact verification (Thorne et al., 2018a), reveal that PLMs and LLMs based fact verification systems severely lack robustness against even simple perturbations, raising concerns about the trustworthiness of these systems and emphasizing the need for more robust methodologies. Addressing these challenges is essential for ensuring the reliability and effectiveness of fact verification models in real-world applications.

To summarize, our **contributions** are:

1. We propose FACTEVAL, a novel large-scale benchmark for comprehensive evaluation of the robustness of fact verification systems, covering 16 realistic word-level and character-level perturbations and 4 types of subpopulations.

2. We evaluate the performance of PLMs and LLMs, specifically Llama, Gemma, and Mistral, in zero-shot, few-shot, and chain-of-thought prompting.

3. Our comprehensive evaluation shows that PLMs and LLMs based fact verification systems are not resilient to minor perturbations.

Many perturbations in FACTEVAL can be applied to tasks other than fact verification, making it easy to incorporate in evaluations beyond just reporting standard metrics on held-out data.

2 Related Work

2.1 Robustness of NLP models

Despite having achieved great progress on standard benchmarks, NLP models still struggle in the presence of small changes to the input. Several works have shown that high-performing transformer-based models are brittle to adversarial attacks and small perturbations to the inputs (Alzantot et al., 2018; Lin et al., 2021; Neerudu et al., 2023). Recent studies on robustness of LLMs investigated out-of-distribution datasets (Yuan et al., 2023; Gupta et al., 2024) as well as challenge test sets, behavioral testing, contrast sets, and adversarial inputs (Gupta et al., 2024).

Various tasks have been explored in studies on NLP robustness, including sentiment analysis (Jin et al., 2020; Kiela et al., 2021; Yuan et al., 2023; Mamta et al., 2023), toxic content detection (Li et al., 2019; Yuan et al., 2023), argument mining (Sofi et al., 2022; Mayer et al., 2020), machine translation (Sai et al., 2021a; Wang et al., 2021; Morris et al., 2020), question answering (Goel et al., 2021; Moradi and Samwald, 2021; Kiela et al., 2021; Yuan et al., 2023; Gupta et al., 2024), natural language inference (Wu et al., 2021; Morris et al., 2020; Li et al., 2021; Yuan et al., 2023), and dialogue generation (Sai et al., 2020; Li et al., 2023). Perturbations are applied at character-level, word-level, or sentence-level (Wang et al., 2022).

Few works created test sets by constructing subpopulations based on various input characteristics (Mille et al., 2021; Goel et al., 2021). These subpopulations were defined by features such as topic, entity names, input length, presence of pronouns for a particular gender. allowing for a more granular analysis of model performance across different subsets of data.

2.2 Robustness of Fact Verification Systems

There have been few attempts to assess the robustness of fact verification systems against adversarial attacks. The majority of these works manipulate the claims to create adversarial examples. Niewinski et al. (2019) introduced a generative enhanced model (GEM), a modified and fine-tuned GPT-2 language model used to generate adversarial claims. Other works (Thorne et al., 2019; Hidey et al., 2020) evaluated the robustness of the BERT model (Devlin et al., 2019). Thorne et al. (2019) evaluated three adversarial attacks by altering the claim and without modifying the evidence. The attacks in-

clude rules-based meaning-preserving transformations to induce classification errors, hand-crafted rules exploiting common patterns and constructions in the FEVER claims and a paraphrase model to generate new instances. Hidey et al. (2020) created an adversarial dataset of 1000 examples that targeted models' weaknesses in (i) multi-hop reasoning by augmenting existing claims with conjunctions or relative clauses, (ii) temporal reasoning by manipulating claims using date-alteration rules, and (iii) named entity ambiguity and lexical variation using the using the lexical substitution method of Alzantot et al. (2018).

Atanasova et al. (2020) evaluated the robustness of the RoBERTa model (Liu et al., 2019) using gold evidence from the dataset. They constructed universal adversarial triggers which are n-grams that, when appended to the actual claim, can change the model's prediction from the source to a target class. These trigger tokens were generated using the HotFlip algorithm (Ebrahimi et al., 2018) which updates the embeddings of the trigger tokens to minimize the loss for the target class.

Some works (Du et al., 2022; Abdelnabi and Fritz, 2023) modified the evidence repository to mislead the retrieval system by adding, removing, or altering evidence, and evaluated their attacks on BERT and RoBERTa-based models.

Prior studies on robustness for fact verification have focused on traditional task-specific fine-tuned models (e.g. BERT style models). There is a noticeable gap in the literature regarding the investigation of the robustness of LLMs within the context of fact verification, which we address in this paper.

2.3 Fact Verification

Recent efforts have leveraged LLMs to solve the fact verification task (Vykopal et al., 2024). Lee et al. (2021) used GPT-2 (Radford et al., 2019) to assess the factuality of the claim based on the perplexity of evidence-conditioned claim generation. Tang et al. (2024) used GPT models to create synthetic training data to enhance the performance of LLMs in fact verification tasks, whereas Li et al. (2024b) proposed a self-sufficient approach to claim verification using prompting instructions across multiple language models. Zeng and Gao (2023a) leveraged a consistency mechanism to improve the performance of fact verification models by constructing 3 variants of the original prompt based on three logical relations (confirmation, negation, uncertainty) and fine-tuned the model on these variants. Yao et al. (2023); Ni et al. (2024) leveraged chain-of-thought prompting to effectively verifying complex claims using reasoning steps.

3 FACTEVAL

We present FACTEVAL, a framework for evaluating the robustness of fact verification models. Our framework incorporates word-level and character-level perturbations, as well as visual and phonetic perturbations. Furthermore, we design subpopulations of the data based on demographic factors, temporal information, sentiment, and named entities. FACTEVAL contains 16 realistic perturbations and 4 types of subpopulations.

3.1 Task Definition

Given a claim and some evidence, the aim of a fact verification model is to determine whether the evidence supports, refutes, or does not provide enough information to reach a verdict.

To verify the robustness of a fact verification model, we assume no black-box or white-box access to the model. Hence, we modify the input by applying perturbations before passing it to the model. In particular, for a claim C consisting of n tokens $c_1, c_2, c_3, ..., c_n$, m pieces of evidence $E = E_1, E_2, ..E_m$ with label y, the objective is to apply *label-preserving* perturbations to C and test whether these mislead the target model FV, i.e., $FV(C', E) \neq y$. Perturbations are applied to the claims, while the evidence remains unchanged in order to evaluate how well the model can handle variations in the claims without introducing noise into the evidence.

3.2 Perturbations

FACTEVAL contains several adversarial datasets obtained using word-level and character-level perturbations. Examples are presented in Table 1.

3.2.1 Word-Level Perturbations

Contractions/Expansions Contractions are words obtained by shortening and combining two words, often using an apostrophe (e.g. *do not* → *don't*). Expansions are the opposite, where a contraction is written out in its full form (e.g. *don't* → *do not*). These perturbations help evaluate how well models handle linguistic variations and maintain meaning across different writing styles, such as formal and informal language. For this pertur-

Perturbation	Original Claim	Perturbed Claim
Contractions	Oscar Isaac **did not** act in X-Men	Oscar Isaac **didn't act** in X-Men
Expansions	Henry Cavill **didn't** play Superman.	Henry Cavill **did not** play Superman.
Jumbling	Oscar Isaac **did not act** in X-Men	Oscar Isaac **not did act** in X-Men
Num to Words	Southern Hospitality fell to number **23** on the Top 40.	Southern Hospitality fell to number **twenty-three** on the Top 40.
Repeat Phrases	South Island is sometimes referred to as the ""mainland"" of New Zealand.	South Island is sometimes referred to as the " mainland " of New Zealand . **South Island is sometimes**
Subject Verb Disagreement	One Dance **was** by Drake.	One Dance **were** by Drake.
Typos	Richard Kuklinski is **a innocent** man.	Richard Kuklinski is **ainnocent** man.
Word Repetitiom	Trouble with the Curve is a cat	Trouble with the Curve **Curve** is a cat
Synonyms	A **good** Day to Die Hard stars Bruce Willis as John McClane	A **nice** Day to Die Hard stars Bruce Willis as John McClane
Tautology	Benjamin Franklin was a person.	Benjamin Franklin was a person. **and true is true. and true is true. and true is true.**
Character Swap	LinkedIn **is based in Russia.**	Liknedin **is based in russia.**
Character Repetition	The Burj Khalifa contains zero escalators	The Burj Khalifa **contaains** zero escalators
Character Insertion	Trouble with the Curve is a cat.	Trouble **wtith** the Curve is a cat .
Character Deletion	The Catcher in the Rye is a young adult **novel**.	The Catcher in the Rye is a young adult **novl**.
Phonetic	**Spider-Man** 2 **was** written by Donald Trump.	**Spiderman** 2 **wasss** written by Donald Trump.
Homoglyph	**Annie** is the title of a work.	**Aññie is the title of a work.**
LEET	**Scandal** is a **Mexican** band.	**5candal is a M3xican band.**

Table 1: Perturbation examples.

bation, we use a dictionary containing all possible expansions and contractions (Sai et al., 2021b).

Jumbling Word Order We perturb the claim by randomly changing the order of its words. This perturbation tests how well models handle variations in syntax and word arrangement.

Numbers to Words In real-world data, numbers can appear as numerals or words. To evaluate the versatility of a model, we convert all numbers (e.g. "2") to their word equivalents (e.g. "two").

Repeat Phrases To test how robust a model is to repetitive patterns, we perturb a sentence by adding the first quarter of the claim to the end of the original claim.

Subject Verb Disagreement As grammar errors occur frequently in real-world data, we evaluate the robustness of models in understanding grammatical structure. We follow Sai et al. (2021b) and create syntactically incorrect sentences based on subject-verb disagreement (i.e. singular vs plural).

Typos We introduce a typographical error (typo) into a claim by swapping adjacent characters. This perturbation simulates realistic typing errors, which is essential for creating more resilient applications capable of handling noisy or imperfect data.

Word Repetition We randomly select a word from a sentence and insert it immediately after the selected word.

Synonym Adjectives We select the adjectives in a claim and replace them with their synonyms from WordNet (Fellbaum, 1998). The goal is to create semantically equivalent but lexically varied versions of the claims.

Addition of Tautology To create this adversarial dataset, we append *and true is true* three times at the end of the claim.

Phonetic Perturbations Non-native English speakers may pronounce words differently as they apply the speech rules of their first language. This may also affect their writing style. Using a dictionary of human-written phonetic perturbations (Le et al., 2022), we introduce phonetic perturbations using 25% and 50% budget values ($x\%$ budget value means perturbing $x\%$ words of the claim).

3.2.2 Character-Level Perturbations

Character Swapping For this perturbation, we randomly swap adjacent characters within a word.

Character Repetition We randomly select a character (except the first/last characters) from a random word in the claim and insert it directly after the selected character. This mimics common typographical errors where a key is accidentally pressed twice in quick succession. This perturbation introduces subtle noise into the text while preserving the readability of the word.

Character Insertion We select a random character (except the first/last characters) from a randomly chosen word of at least three characters in the claim. We then insert the caracter into a randomly chosen position within the word. The insertion occurs at any position except the first and last, mimicking typographical errors such as unintentional key presses during typing.

Character Deletion We randomly select a character (except the first/last characters) from a random word in the claim and delete it. This simulates common typing errors where characters are accidentally omitted. This perturbation preserves the

general structure of the word and sentence, introducing noise in a controlled way.

Homoglyph Perturbations Homoglyphs are characters that look similar or identical to other characters (e.g. n and $\tilde{n}$). We perturb characters based on the dictionary from Unicode Security.[2] We experiment with various perturbation budgets (i.e. k = 25% and 50%) to evaluate the robustness of models against these changes.

LEET Perturbations LEET perturbations amount to replacing letters with their visually similar counterparts (i.e. numbers, special characters, or other symbols) and are often used as a distinct writing style online. For example, 'A' can be replaced with '4', 'E' with '3', and 'I' with '1'. We use a pre-defined dictionary (Eger et al., 2019; Eger and Benz, 2020) to perturb claims and experiment with character perturbation ratios of 25% and 50%.

3.3 Subpopulations

Subpopulations represent distinct subgroups of data. We design four types of subpopulations to ensure that models perform robustly and fairly across diverse groups of data.

Demographic We create subpopulations based on *nationality*, *ethnicity*, and *gender* to analyze a model's behavior across different demographic groups. We identify person entities and retrieve their ethnicity, gender, and nationality from their corresponding Wikipedia pages, if available. If a claim contains multiple nationalities, we include it in all relevant subpopulations.

Temporal Information We design temporal subpopulations based on the presence or absence of date entities (i.e. calendar dates, years, months, days, and general time expressions such as "the 90s"). This is useful to assess how well the model handles claims that rely on accurate temporal understanding.

Sentiment Similarly to Mamta et al. (2020) where sentences are annotated based on the event described, we create subpopulations related to the claim's sentiment (i.e. *positive*, *negative*, or *neutral*) using VADER (Hutto and Gilbert, 2014). This allows us to evaluate whether the model correctly classifies claims with positive sentiments more effectively than those with negative or neutral sentiments.

Named Entities We design subpopulations related to *person*, *location*, and *organization* entities to evaluate model performance on claims mentioning persons compared to those mentioning organizations or locations.

[2] www.unicode.org/Public/security/latest/confusables.txt

	Supported	Refuted	NEI
Train	80,035	29,775	35,639
Test	3,333	3,333	3,333

Table 2: Data distribution in FEVER.

4 Experiments

4.1 Dataset

For our experiments, we use FEVER, one of the largest and most widely-used benchmarks for fact verification (Thorne et al., 2018a,b). As one of the first large-scale datasets for fact verification (Guo et al., 2022), FEVER contains 185,000 annotated claims, each accompanied by evidence from Wikipedia. The claims are classified as *supported*, *refuted* or *not enough information (NEI)*. FEVER provides gold evidence for *supported* and *refuted* classes only. We follow Zeng and Gao (2023b) to provide evidence for samples in the *NEI* class.

Table 2 shows the data distribution in FEVER. We further split the training data into train (80%) and development (20%) sets to fine-tune pre-trained language models.

4.2 Models

We experiment with pre-trained language models which can be directly fine-tuned, as well as large language models.

Pre-trained Language Models. We perform task-specific fine-tuning of BERT (Devlin et al., 2019) by adding an output layer on top of these models. The input sequence is constructed by separating the claim C and the evidences E using a separator token. This input sequence is passed to the model and the final sentence representation is fed into an output layer for the classification task.

Large Language Models. We evaluate Llama3 8B (base model) (Dubey et al., 2024), Mistral 7B Instruct (Jiang et al., 2023), and Gemma 7B Instruct (Mesnard et al., 2024) models using zero-shot, few-shot, and chain-of-thought prompting.

Zero-Shot Prompting Zero-shot prompting relies solely on the pre-trained knowledge and generalization abilities of LLMs. We provide task instructions and class definitions to ensure that LLMs

understand the task at hand. The instructions clarify the expected input (the claim and supporting evidence) and the required answer (classification into one of the predefined classes).

Few-Shot Prompting In the few-shot setting, the model is provided with a small set of labeled examples to guide its response, leveraging in-context learning (Xun et al., 2017). The model observes a few labeled instances from the dataset and uses these examples to infer the class at inference time. We provide task instructions and randomly select two examples from each class to include in the prompt.[3]

Chain-of-Thought Prompting We use Chain-of-Thought (CoT) along with the task definition in zero-shot setting. We follow Wei et al. (2022) for CoT and add *Think step by step* at the end of each zero-shot prompt.

5 Results and Discussion

To evaluate the robustness of LLMs and PLMs in different attack settings, we compute accuracy, F_1, and Attack Success Rate (ASR). Accuracy and F_1 are calculated on the FEVER test set as well as the adversarial test sets. ASR is calculated as the percentage of adversarial examples that can successfully attack the target model.

5.1 Adversarial Robustness

Which model is more/least robust among LLMs? Table 3 shows the results for zero-shot and few-shot learning across Llama, Mistral, and Gemma models. In the zero-shot setting, Gemma outperforms both Llama and Mistral on the FEVER test set, indicating that Gemma has a stronger ability to generalize in the absence of labeled examples. However, all models, despite their good performance on the FEVER test set, are vulnerable to small perturbations in the input data, as indicated by their ASR.

Among the models, Llama is the most vulnerable (i.e. having a higher ASR when exposed to adversarial perturbations). It is interesting to note that even character swapping in the input text significantly affects the performance of all models. While few-shot models perform notably better than zero-shot models in terms of accuracy, they still remain vulnerable to adversarial examples.

In the few-shot setting, Llama has a higher ASR than both Mistral and Gemma in most perturbations. Gemma, on the other hand, is the most robust model across the majority of perturbations. For example, in homoglyph and LEET perturbations, Gemma demonstrates a lower ASR than both Llama and Mistral, showing its resilience against these types of adversarial attacks.

We also investigate whether CoT makes the best two models, Mistral and Gemma, more robust to perturbations. The results in Table 4 indicate that both models remain vulnerable to perturbation in the CoT setting. Gemma has a higher ASR than Mistral for the majority of perturbations, suggesting that Mistral is more resilient to perturbations compared to Gemma in this setting. However, Gemma has lower ASR for LEET perturbations compared to Mistral.

How do LLMs compare to PLMs in terms of robustness? Table 5 shows that BERT performs better compared to LLMs, highlighting the importance of task-specific fine-tuning. However, despite its good performance, BERT remains vulnerable to all types of adversarial perturbations. While BERT shows greater robustness than the Llama zero-shot and few-shot (base model) variants, the instruction-tuned models, Mistral and Gemma, have even higher resilience compared to BERT.

Which perturbations are more challenging? The LLama zero-shot model is highly vulnerable to almost all types of perturbations, with ASR ranging from 13% to 37%. Homoglyph and LEET perturbations have the highest ASR across all zero-shot and few-shot models, indicating that these models are particularly susceptible to such attacks. Additionally, all models show increased vulnerability to perturbations involving typos, tautology, phonetic variations, and character swapping. Zero-shot models, in particular, are more affected by phonetic perturbations than their few-shot models. Interestingly, the few-shot LLama model has a higher ASR in the presenece of tautology perturbations compared to its zero-shot variant. This also indicates that the presence of tautology in in-context learning (ICL) can confuse the Llama model about the patterns it has learned from the few-shot samples, leading to a higher ASR.

5.2 Subpopulations

Table 6 shows the results for different subpopulations in zero-shot, few-shot, and CoT setting for Llama, Mistral, and Gemma. We present the data distributions for subpopulations in Section C.

How do models behave across demographic sub-

[3]The prompts are provided in the Appendix.

	Perturbation	Zero-shot									Few-shot								
		Llama			Mistral			Gemma			Llama			Mistral			Gemma		
		Acc	F1	ASR	Acc	F1	ASR	Acc	F1	ASR	Acc	F1	ASR	Acc	F1	ASR	Acc	F1	ASR
	None	57.43	54.88	–	72.73	69.8	—	77.44	74.82	–	66.63	60.69	–	83.82	83.18	–	61.25	59.03	–
Word	Contractions	47.55	42.7	17.21	70.32	67.18	2.94	75.74	72.61	2.19	63.25	55.91	4.58	81.87	81.14	2.13	49.54	46.52	7.11
Word	Expansions	47.3	42.29	17.64	70.38	67.29	2.85	75.89	72.85	2	63.44	56.28	4.28	81.92	81.2	2.07	49.41	46.39	7.36
Word	Jumbling	35.71	26.69	**37.82**	60.33	53.47	**17.05**	65.04	57.37	**16**	48.69	38.64	**26.81**	59.85	56.1	**28.59**	45.02	40.55	**15.58**
Word	Number to Words	47.21	42.27	17.8	70.22	67.08	2.07	74.7	71.59	3.25	63.18	55.91	4.68	81.8	81.07	2.21	49.2	46.14	7.5
Word	Repeat Phrases	42.12	36.52	26.66	68.48	65.32	5.47	74.88	71.52	3.31	58.55	50.17	11.77	80	79.26	4.46	48.86	45.41	8.73
Word	SVD	44.38	38.97	22.72	68.16	65.47	4.91	74.01	70.76	4.14	61.33	53.54	7.52	80.3	79.49	4.01	47.78	44.46	10.42
Word	Typos	44.22	38.66	22.99	66.98	63.17	7.53	72.1	68.15	6.72	59.03	51.89	10.96	78.18	77.2	6.53	49	45.91	**8.12**
Word	Word Repetition	43.21	37.69	24.76	68.1	64.53	6	72.66	69.27	5.59	60.38	53.47	9.39	79.28	78.63	5.03	47.97	44.66	10.06
Word	Synonyms	46.82	41.78	18.48	69.72	66.5	3.76	74.26	71.16	3.68	63.19	56.03	4.66	81.22	80.46	2.91	49	45.91	8.12
Word	Tautology	44.73	40.04	22.1	66.51	61.9	7.81	74.14	70.44	4.3	41.18	32.45	**38.17**	77.51	76.29	7.33	56.66	53.98	7.48
Character	Phonetic (0.25)	42.61	36.79	25.81	66.46	62.33	7.87	71.35	69.59	7.28	61.58	54.45	6.58	79.01	78.3	4.96	57.69	55.2	5.16
Character	Phonetic (0.50)	39.64	32.45	**30.96**	64.96	60.45	10.3	70.4	65.49	**9.09**	60.85	53.87	7.17	78.04	77.33	5.93	55.86	53.41	6.89
Character	Char Swap	39.05	31.57	**31.99**	65.68	61.05	9.69	70.42	65.67	9.08	47.59	38.71	**28.58**	74.63	73.27	**10.77**	56.37	53.01	7.96
Character	Char Repetition	42.84	37.18	25.4	68.26	64.55	6.15	73.12	69.74	5.29	58.91	52.84	11.09	79.62	78.92	4.82	57.98	55.14	5.33
Character	Char Insertion	42.27	36.44	26.39	66.64	62.74	8.01	73.32	69.62	5.31	58.71	51.53	11.45	78.47	77.67	6.18	57.72	54.87	5.75
Character	Char Deletion	42.67	36.91	25.7	66.37	61.33	7.37	72.88	69.21	5.88	57.73	50.5	12.35	78.21	77.44	6.49	57.18	54.37	6.63
Character	Homoglyph (0.25)	49.75	45.76	13.37	65	60.11	10.25	71.12	67.06	7.01	55.56	46.9	16.62	76.22	75.08	9.06	56.35	53.77	8
Character	Homoglyph (0.50)	47.53	42.73	17.24	62.58	57.27	**13.96**	66.66	60.68	**14.01**	51.18	41.79	23.18	69.65	67.68	**16.9**	54.24	51.69	**11.44**
Character	LEET (0.25)	41.85	36.07	27.13	53.49	48.37	**26.46**	58.45	51.39	**24.51**	39.8	28.82	**40.26**	51.72	47.87	**38.29**	49.66	47.44	**18.91**
Character	LEET (0.50)	37.45	31.5	**35.01**	46.58	41.58	**35.94**	50.89	43.52	**34.28**	38.2	26.57	**42.66**	42.98	36.23	**48.72**	44.3	42.71	**27.66**

Table 3: Results on adversarial perturbations in zero-shot and few-shot setting. Here, SVD: Subject Verb Disagreement, ASR: Attack Success Rate, Acc: Accuracy. *None* denotes accuracy on the original FEVER test set. The top 5 perturbations in terms of ASR are highlighted for each setting.

	Perturbation	Mistral			Gemma		
		Acc	F1	ASR	Acc	F1	ASR
	None	81.11	80.85	–	72.07	71.51	–
Word	Contractions	75.99	75.54	6.3	66.39	65.77	7.24
Word	Expansions	76.33	75.89	5.88	66.47	65.86	7.13
Word	Jumbling	58.08	56.18	**28.38**	52.22	49.88	**27.05**
Word	Number to Words	75.9	75.46	6.42	66.27	65.65	7.41
Word	Repeat Phrases	75.79	75.43	6.55	64.14	63.55	10.39
Word	SVD	75.5	75.09	6.91	64.4	63.85	10.02
Word	Typos	73.5	73.11	9.38	62.87	62.12	12.16
Word	Word Repetition	75.08	74.72	7.43	63.78	63.25	11.47
Word	Synonyms	75.12	74.68	7.39	65.42	64.79	8.59
Word	Tautology	76.05	76.61	6.03	65.76	65.06	8.75
Character	Phonetic (0.25)	76.05	75.61	6.12	66.54	65.92	7.66
Character	Phonetic (0.50)	74.23	73.41	8.34	65.41	64.75	8.85
Character	Char Swapping	71.89	71.42	11.35	62.87	62.1	12.76
Character	Char Repetition	74.37	73.96	8.31	62.33	61.86	13.04
Character	Char Insertion	73.68	73.22	9.16	62.07	61.43	13.23
Character	Char Deletion	73.17	72.69	9.77	61.42	60.85	14.14
Character	Homoglyph (0.25)	70.49	69.95	**13.08**	61.94	61.16	14.06
Character	Homoglyph (0.50)	65.25	64.06	**19.54**	57.05	55.65	**20.85**
Character	LEET (0.25)	52.22	46.71	**39.35**	49.11	47.07	**31.86**
Character	LEET (0.50)	42.7	37.78	**47.35**	44.26	41.73	**38.58**

Table 4: Results on adversarial perturbations in CoT zero-shot setting. Here, SVD: Subject Verb Disagreement, ASR: Attack Success Rate, Acc: Accuracy. *None* denotes accuracy on the original FEVER test set.

	Perturbation	BERT		
		Acc	F1	ASR
	None	94.01	94	–
Word	Contractions	81.36	81.22	13.45
Word	Expansions	81.82	81.69	12.96
Word	Jumbling	73.15	72.1	22.18
Word	Number to Words	81.58	81.44	13.22
Word	Repeat Phrases	78.81	78.62	16.16
Word	SVD	81.57	81.42	13.23
Word	Typos	75.03	74.5	20.18
Word	Word Repetition	81.16	81.03	13.65
Word	Synonyms	81.35	81.21	13.46
Word	Tautology	80.26	80.17	14.61
Character	Phonetic (0.25)	69.06	68.14	23.11
Character	Phonetic (0.50)	64	63.56	**28.48**
Character	Char Swapping	71.94	70.74	**23.47**
Character	Char Repetition	75.55	75.1	19.63
Character	Char Insertion	75.42	74.9	19.77
Character	Char Deletion	75.46	74.98	19.73
Character	Homoglyph (0.25)	75.58	74.96	19.6
Character	Homoglyph (0.50)	70.87	69.97	**24.61**
Character	LEET (0.25)	51.67	46.66	**45.03**
Character	LEET (0.50)	48.65	42.71	**48.24**

Table 5: Results on adversarial perturbations for BERT model. Here, SVD: Subject Verb Disagreement, ASR: Attack Success Rate, Acc: Accuracy. *None* denotes accuracy on the original FEVER test set.

populations? The models exhibit varying performance across different demographic groups, suggesting that models do not treat each group equally, as certain subgroups achieve higher accuracy compared to others. For instance, the zero-shot Llama model has the lowest classification performance on the *Mexican* subgroup, whereas the few-shot Llama model underperforms on the *Spanish* subgroup.

How does temporal information affect the performance of the models? All models exhibit superior performance when claims do not contain any mention of temporal information. This indicates that the models struggle to interpret and reason about temporal information.

How does sentiment affect the performance of the models? The models' performance is significantly influenced by the sentiment of the claims. Claims with positive sentiments achieve higher accuracy compared to those with negative or neutral sentiments across all the models.

How do models perform on named entities subpopulations? The results indicate that the models can effectively handle various types of named entities mentioned in claims. Both the zero-shot and few-shot Llama models perform significantly better on claims that include location information compared to those mentioning person and organization entities.

5.3 Qualitative Analysis

Figure 1 shows a few examples for the zero-shot setting where models fail to correctly classify adversarial samples. When the input text is perturbed,

		Llama						Mistral						Gemma					
		Zero-shot		CoT		Few-shot		Zero-shot		CoT		Few-shot		Zero-shot		CoT		Few-shot	
Type	Subpopulation	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Gen	Male	58.07	54.18	47.49	42.54	67.78	61.13	73.2	70.33	81.67	81.33	85.84	84.93	63.83	65.09	72.71	71.89	79.95	76.18
	Female	57.01	53.36	46.46	42.04	67.8	60.07	72.4	68.05	81.95	81.44	88.08	87.29	65.09	61.42	76.65	76.27	78.11	75.62
Nationality	Chinese	45.71	47.43	37.14	32.72	45.71	45.13	65.71	66.62	68.57	70.71	80	80.13	45.71	47.51	82.85	82.53	71.42	72.07
	British	60.86	56.43	48.06	41.67	69.08	62.28	73.18	70.25	83.57	83.18	80.43	80.43	65.94	3.04	74.15	73.24	79.22	74.91
	Spanish	46.87	39.58	53.12	45.79	59.37	45.64	68.75	65.05	78.12	72.54	75	69.51	53.12	48.92	68.75	67.79	65.62	59.87
	Indian	61.9	52.28	49.2	44.48	70.63	64.34	78.57	75.33	80.95	78.9	89.68	88.42	63.49	60.25	73.01	71.67	82.53	79.16
	American	58.18	53.53	46.92	41.87	68.14	60.46	73.41	69.6	82.43	82.05	87.31	86.26	65.36	61.88	73.75	72.84	79.85	75.78
	Australian	50.01	43.8	56.66	56.65	71.66	66.93	75.01	75.25	83.33	82.87	88.33	87.99	61.66	59.9	86.66	86.46	83.33	80.94
	Mexican	36.58	23.67	51.21	47.1	58.52	54.85	65.85	67.57	80.48	80.41	85.36	85.06	65.85	65.91	75.6	73.72	80.42	76.32
	Canadian	60.37	59.51	37.73	36.05	58.49	51.99	83.01	82.65	88.67	88.55	88.67	88.61	67.92	66.74	83.01	83.05	90.56	90.41
	French	58.52	53.46	54.87	49.8	70.73	58.8	78.04	69.42	73.17	68.92	81.7	77.99	67.07	60.21	80.48	78.89	76.82	67.49
Ethnicity	English	65.78	61.74	39.47	34.83	71.05	59.89	76.31	71.76	78.94	72.45	92.1	92.2	65.78	62.62	81.57	81.29	81.57	81.29
	British American	55.17	42.73	37.93	26.15	62.06	47.17	72.22	71.6	75.86	72.25	79.31	75.19	55.17	53.25	68.96	65.41	82.75	77.05
	African American	61.11	49.56	52.77	45.43	80.55	75.94	80.55	81.81	86.11	86.14	88.88	88.15	66.66	65.47	72.22	70.71	88.87	88.45
	Ashkenazi Jews	53.7	47.28	42.59	41.47	52	38.88	75.92	73.2	79.62	79.68	90.74	90.42	55.55	54.42	77.77	77.68	81.48	79.67
	Chinese	47.61	49.89	33.33	28.87	42.85	41.81	66.66	67.59	80.95	82.14	90.47	90.7	42.85	42.35	76.19	75.79	66.66	67.97
	Mexican Americans	57.14	36.98	47.61	27.35	80.95	55.1	66.66	50.28	71.42	65.07	90.47	86.34	71.42	51.55	85.71	86.9	90.47	89.28
Temp	Yes	54.65	53.53	45.31	40.97	57.54	53.22	72.61	68.72	80.41	79.99	77.87	76.23	61.79	58.21	69.15	69.06	78.91	78.48
	No	58.06	55.11	47.62	45.04	68.68	62.4	73.29	72.92	84.19	84.37	83.59	83.73	59.44	57.56	72.73	72.03	77.11	73.71
Sent	Negative	54.16	53.69	40.83	38.39	59.93	57.3	66.17	66.21	80.58	79.09	82.89	83.28	54.77	55.66	71.72	71.71	72.49	72.1
	Neutral	56.7	53.86	45.74	41.8	66.4	60.18	73.22	70.18	80.13	79.85	83.68	82.94	61.63	59.16	72.75	72.19	78.1	75.45
	Positive	62.43	58.71	49.16	43.93	72.2	64.44	75.55	70.61	81.38	80.03	84.99	83.34	64.38	60.29	69.83	68.09	78.46	73.5
NER	Person	57.69	54.3	46.53	42.06	66.79	60.29	72.08	68.9	81.44	81.11	85.32	84.53	63.35	60.81	73.25	72.66	78.34	75.12
	Location	59.32	56.42	45.99	41.19	69.62	63.4	73.13	69.79	81.35	80.98	83.88	82.94	60.91	58.44	71.21	70.47	77.3	74.28
	Organization	58.79	56.79	45.55	41.46	67.32	61.77	72.59	69.54	81.22	80.97	83.75	83.35	59.67	57.32	71.42	71.07	77.68	75.4

Table 6: Subpopulation Performance across Zero-Shot, Few-Shot, and Chain-of-Thought (CoT) Setting for Llama, Mistral and Gemma models. Here, Gen: Gender, Temp: Temporal, Sent: Sentiment, Acc: Accuracy.

Orig Claim	Adv Claim	Gold	Llama Orig / Adv	Mistral Orig / Adv	Gemma Orig / Adv
Matches were contested at SummerSlam.	Matc̣hes were contested at SummerSlam.	S	S / NEI	S / NEI	S / NEI
Annie is the title of a work.	Annіe is the title of a work.	S	S / R	S / NEI	S / NEI
The Wonder Years was only a book.	Thẹ Wender Years was only a book.	R	R / NEI	R / NEI	R / S
Benjamin Franklin was born in 1790.	Benjɑmiɳ Franklin was born in 1790.	R	R / NEI	R / NEI	R / NEI
Spider-Man 2 was destroyed in 2004.	Spidėr-Ṃan 2 was destroyed in 2004.	NEI	NEI / R	R / NEI	NEI / NEI

Figure 1: Predictions for *homoglyph* in the zero-shot setting on the orig(inal) and adv(ersarial) claims.

Orig Claim	Adv Claim	Gold	Llama Orig / Adv	Mistral Orig / Adv	Gemma Orig / Adv
John S. McCain Jr. went to school.	jhon .s mccain jr. went to school.	S	S / NEI	S / NEI	S / S
1997 was the year No Way Out was released.	1979 aws the year no way out aws released.	S	S / R	S / NEI	S / R
Ares is a Senator.	aers si a senator.	NEI	NEI / R	NEI / NEI	NEI / R
LinkedIn is based in Russia.	liknedin si based in russia.	R	R / S	R / NEI	R / NEI
2 Hearts is a song by Nirvana.	2 herats si a song by nirvana.	R	R / S	R / S	R / R

Figure 2: Predictions for *character swap* in the few-shot setting on the orig(inal) and adv(ersarial) claims.

the models often misclassify supported/refuted classes as NEI (first 3 rows), suggesting that the models struggle to maintain their understanding of the original input when we apply subtle changes. Gemma and Mistral are more robust compared to Llama – example 5 illustrates a case where Gemma and Mistral models show resilience to homoglyph perturbations as opposed to Llama.

Figure 2 shows a few examples of character swapping perturbations for few-shot setting. We can see that (i) Llama and Mistral are fooled by character swaps, while Gemma is robust (examples 1 and 5); (ii) all models are misled by character-swap perturbations (examples 2 and 4); (iii) only Mistral is robust (example 3).

5.4 Adversarial Training

To investigate the impact of adversarial training on PLMs and LLMs, we conduct several experiments on BERT and Mistral using the perturbations with a high success rate. For BERT model, we generate adversarial data by applying perturbations to the training data. The adversarial data is then combined with the original training data, and the model is re-trained on the augmented dataset. Table 7 shows the results for BERT. We observe that adversarial training improves the robustness of the BERT model. For example, in the case of LEET, phonetic, and homoglyph perturbations, there is a significant improvement compared to other perturbations.

Similarly, for the Mistral model, we incorporate adversarial examples for each selected perturbation in the prompts, along with their corresponding examples. A total of 12 adversarial examples were added to the few-shot prompt, consisting of 6 original examples and 6 adversarial examples. Table 8 shows the results for Mistral. It can be seen that the addition of adversarial prompts increases the robustness of the Mistral model against only a few of the perturbations (highlighted). There is an increase in ASR for the remaining perturbations, which can be attributed to the added noise in the prompts.

6 Conclusion

We introduced FACTEVAL, a novel large-scale benchmark designed to evaluate the robustness of large language models in the fact verification do-

		Acc	F1	ASR	w/o Adv.	Acc. TS
Word	Phonetic (0.25)	78.99	78.7	**15.89**	23.11	93.93
	Phonetic (0.50)	78.56	78.23	**16.35**	28.48	93.12
Character	Char Swapping	71.85	70.64	23.56	23.47	93.41
	Char Repetition	76.32	75.93	**18.74**	19.63	93.56
	Char Insertion	76.09	75.71	**18.99**	19.77	93.48
	Char Deletion	76.07	75.22	**19.01**	19.73	93.89
	Homoglyph (0.25)	78.65	78.33	**16.04**	19.6	93.69
	Homoglyph (0.50)	75.93	75.29	**18.95**	24.61	93.7
	LEET (0.25)	78.96	78.68	**15.1**	45.03	93.27
	LEET (0.50)	77.12	76.76	**17.08**	48.24	93.01

Table 7: Results on adversarial perturbations using BERT after adversarial training on augmented data. Here w/o Adv is ASR without adversarial training, TS: accuracy on actual test set.

	Perturbation	Mistral			w/o Adv.	Acc. TS
	Acc	F1	ASR	ASR		
Word	Contractions	80.71	80.11	3.71	**2.13**	87.4
	Expansions	80.67	80	3.76	**2.07**	87.46
	Jumbling	69.01	67.34	**17.67**	28.59	85.21
	Number to Words	80.45	79.84	4.02	**2.21**	87.35
	Repeat Phrases	79.5	79.01	5.14	**4.46**	87.44
	SVD	80.22	79.63	4.3	**4.01**	87.73
	Typos-1	77.98	77.43	6.96	**6.53**	87.03
	Word Repetition	78.18	77.82	6.72	**5.01**	86.92
	Synonyms	79.89	79.28	4.68	**2.91**	87.37
	Tautology	74.55	73.24	11.06	**7.33**	77.41
	Phonetic (0.25)	77	76.5	8.14	**4.96**	82.55
	Phonetic (0.50)	76.95	75.76	8.19	**5.93**	82.67
Character	Char Swapping	76.34	75.56	8.92	**10.77**	82.88
	Char Repetition	76.46	76.17	8.78	**4.82**	85.1
	Char Insertion	74.25	74.11	11.42	**6.18**	83.18
	Char Deletion	74.64	74.42	10.95	**6.49**	82.95
	Homoglyph (0.25)	76.82	75.79	**8.35**	9.06	86.64
	Homoglyph (0.50)	70.88	69.2	**15.43**	16.9	85.55
	LEET (0.25)	64.18	63.44	**23.42**	38.29	86.6
	LEET (0.50)	51.08	51.08	**39.05**	48.72	85.74

Table 8: Results on adversarial perturbations for Mistral after adding adversarial prompts.

main. We developed 16 word-level and character-level perturbations, along with 4 types of subpopulations. Our comprehensive evaluation covered zero-shot, few-shot, and chain of thought prompting methods across Llama, Mistral, and Gemma models. Our analysis revealed that LLMs are not robust to these perturbations, with the Llama model being the least robust compared to Mistral and Gemma. Furthermore, our experiments show that these models show varying performance on subpopulations such as demographic groups, sentiment, entity types, and temporal information. This indicates that more work needs to be done to improve the robustness of the models.

In future work, we plan to extend our framework to other monolingual and code-mixed languages, allowing for a more comprehensive evaluation of robustness across diverse linguistic contexts. We also aim to explore providing explanations for model failures through model explainability techniques, which can offer valuable insights into the model's behavior and help identify areas for improvement.

Limitations

Like many studies, this research has certain limitations that future investigations could address. Specifically, our current work is focused exclusively on the English language and textual inputs. Multi-lingual and multi-modal data ought to be also investigated. In addition, we have assumed no white-box or black-box access to the model, and thus, adversarial samples were generated by randomly selecting words in the claim. This approach may have perturbed words that are not critical to the classification task. In future work, this limitation can be addressed by assuming black-box access and applying perturbations only to words which are crucial for the model's decision-making. Moreover, all subpopulation subsets are derived from a single, existing dataset, limiting the range of subpopulations we could analyze. Some subpopulations (demographic) with a small number of samples were excluded from this analysis.

Ethics Statement

We utilize publicly available datasets for our experiments. These datasets are used solely for academic purposes and in full compliance with their licensing agreements.

Acknowledgement

This research was supported by EPSRC (grant number EP/X04162X/1).

References

Sahar Abdelnabi and Mario Fritz. 2023. Fact-saboteurs: A taxonomy of evidence manipulation attacks against fact-verification systems. In *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, pages 6719–6736. USENIX Association.

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal automated fact-checking: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5430–5448, Singapore. Association for Computational Linguistics.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. 2020. Generating label cohesive and well-formed adversarial claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3168–3177, Online. Association for Computational Linguistics.

Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. 2022. Bad characters: Imperceptible nlp attacks. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1987–2004. IEEE.

Han Cao, Lingwei Wei, Mengyang Chen, Wei Zhou, and Songlin Hu. 2023. Are large language models good fact checkers: A preliminary study. *arXiv preprint arXiv:2311.17355*.

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

Yibing Du, Antoine Bosselut, and Christopher D. Manning. 2022. Synthetic disinformation attacks on automated fact verification systems. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10581–10589. AAAI Press.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

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. Association for Computational Linguistics.

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.

Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. Text processing like humans do: Visually attacking and shielding NLP systems. 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 1634–1647, Minneapolis, Minnesota. Association for Computational Linguistics.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database. *MIT Press google schola*, 2:678–686.

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

Shreya Goyal, Sumanth Doddapaneni, Mitesh M. Khapra, and Balaraman Ravindran. 2023. A survey of adversarial defenses and robustness in NLP. *ACM Comput. Surv.*, 55(14s):332:1–332:39.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Ashim Gupta, Rishanth Rajendhran, Nathan Stringham, Vivek Srikumar, and Ana Marasovic. 2024. Whispers of doubt amidst echoes of triumph in NLP robustness. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5533–5590, Mexico City, Mexico. Association for Computational Linguistics.

Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, and Smaranda Muresan. 2020. DeSePtion: Dual sequence prediction and adversarial examples for improved fact-checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8593–8606, Online. Association for Computational Linguistics.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

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

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Thai Le, Jooyoung Lee, Kevin Yen, Yifan Hu, and Dongwon Lee. 2022. Perturbations in the wild: Leveraging human-written text perturbations for realistic adversarial attack and defense. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2953–2965, Dublin, Ireland. Association for Computational Linguistics.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021. Towards few-shot fact-checking via perplexity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1971–1981, Online. Association for Computational Linguistics.

Chen Li, Meishan Zhang, Xuebo Liu, Zhaocong Li, Derek Wong, and Min Zhang. 2024a. Towards demonstration-aware large language models for machine translation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13868–13881.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2021. Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5053–5069, Online. Association for Computational Linguistics.

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

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2024b. Self-checker: Plug-and-play modules for fact-checking with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 163–181, Mexico City, Mexico. Association for Computational Linguistics.

Yufei Li, Zexin Li, Yingfan Gao, and Cong Liu. 2023. White-box multi-objective adversarial attack on dialogue generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1792, Toronto, Canada. Association for Computational Linguistics.

Bill Yuchen Lin, Wenyang Gao, Jun Yan, Ryan Moreno, and Xiang Ren. 2021. RockNER: A simple method to create adversarial examples for evaluating the robustness of named entity recognition models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3728–3737, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Mamta and Asif Ekbal. 2022. Adversarial sample generation for aspect based sentiment classification. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 478–492, Online only. Association for Computational Linguistics.

Mamta, Asif Ekbal, Pushpak Bhattacharyya, Shikha Srivastava, Alka Kumar, and Tista Saha. 2020. Multi-domain tweet corpora for sentiment analysis: Resource creation and evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5046–5054, Marseille, France. European Language Resources Association.

Mamta Mamta, Zishan Ahmad, and Asif Ekbal. 2023. Elevating code-mixed text handling through auditory information of words. In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing, pages 15918–15932, Singapore. Association for Computational Linguistics.

Tobias Mayer, Santiago Marro, Elena Cabrio, and Serena Villata. 2020. Generating adversarial examples for topic-dependent argument classification 1. In *Computational Models of Argument*, pages 33–44. IOS Press.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Cristian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

Simon Mille, Kaustubh D. Dhole, Saad Mahamood, Laura Perez-Beltrachini, Varun Gangal, Mihir Sanjay Kale, Emiel van Miltenburg, and Sebastian Gehrmann. 2021. Automatic construction of evaluation suites for natural language generation datasets. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks*.

Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1570, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Pavan Kalyan Reddy Neerudu, Subba Oota, Mounika Marreddy, Venkateswara Kagita, and Manish Gupta. 2023. On robustness of finetuned transformer-based NLP models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7180–7195, Singapore. Association for Computational Linguistics.

Jingwei Ni, Minjing Shi, Dominik Stammbach, Mrinmaya Sachan, Elliott Ash, and Markus Leippold. 2024. AFaCTA: Assisting the annotation of factual claim detection with reliable LLM annotators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1890–1912, Bangkok, Thailand. Association for Computational Linguistics.

Piotr Niewinski, Maria Pszona, and Maria Janicka. 2019. GEM: Generative enhanced model for adversarial attacks. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 20–26, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Ananya B Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M Khapra. 2021a. Perturbation checklists for evaluating nlg evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234.

Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. 2021b. Perturbation CheckLists for evaluating NLG evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ananya B Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

Mehmet Sofi, Matteo Fortier, and Oana Cocarascu. 2022. A robustness evaluation framework for argument mining. In *Proceedings of the 9th Workshop on Argument Mining*, pages 171–180.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. *ArXiv*, abs/2404.10774.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Evaluating adversarial attacks against multiple fact verification systems. 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 2944–2953, Hong Kong, China. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Ivan Vykopal, Matúš Pikuliak, Simon Ostermann, and Marián Šimko. 2024. Generative large language models in automated fact-checking: A survey. *arXiv preprint arXiv:2407.02351*.

Jiongxiao Wang, Zichen Liu, Keun Hee Park, Muhao Chen, and Chaowei Xiao. 2023. Adversarial demonstration attacks on large language models. *CoRR*, abs/2305.14950.

Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, et al. 2021. Textflint: Unified multilingual robustness evaluation toolkit for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 347–355.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. 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 6707–6723.

Guangxu Xun, Xiaowei Jia, Vishrawas Gopalakrishnan, and Aidong Zhang. 2017. A survey on context learning. *IEEE Trans. Knowl. Data Eng.*, 29(1):38–56.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Revisiting out-of-distribution robustness in NLP: benchmarks, analysis, and llms evaluations. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Fengzhu Zeng and Wei Gao. 2023a. Prompt to be consistent is better than self-consistent? few-shot and zero-shot fact verification with pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4555–4569, Toronto, Canada. Association for Computational Linguistics.

Fengzhu Zeng and Wei Gao. 2023b. Prompt to be consistent is better than self-consistent? few-shot and zero-shot fact verification with pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4555–4569.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

A Implementation Details

All models were implemented using PyTorch and HuggingFace's (Wolf, 2019) for Llama, Mistral, Gemma, and BERT models. The BERT-base model has 12 transformer layers, a hidden size of 768, and 12 self-attention heads, with a total of 110 million trainable parameters. We optimized the BERT model using the Adam optimizer, with weight updates computed based on categorical cross-entropy loss. All computations were performed on an NVIDIA A100-SXM4 GPU with 40 GB of memory. Named entities are extracted using spaCy Named Entity Recognition.[4]

B Prompts

The zero-shot and few-shot prompts are presented in Figure 3.

C Subpopulations

We include the dataset distribution across each subpopulation in Table 9. Our findings from Section

[4] https://spacy.io/api/entityrecognizer

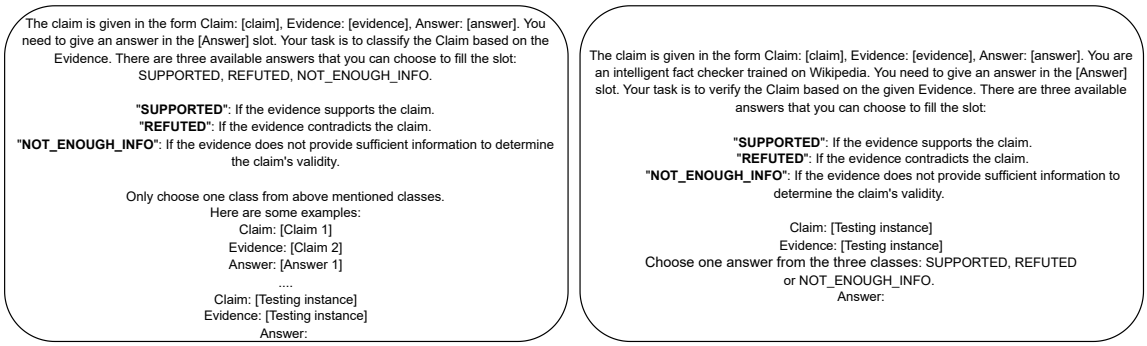

Figure 3: Prompt in the few-shot (left) and the zero-shot setting (right).

5.2 indicate that the performance of LLMs models on various subpopulations is influenced by the pre-training dataset used for each model. Since we are not fine-tuning these models on specific fact-checking datasets such as FEVER, their performance is impacted by the biases and composition of their pre-trained corpora, which can vary across models. For example, the sensitivity to sentiment or demographic factors in these models may be more pronounced due to the representation (or lack thereof) of certain sentiments or demographic groups in their pre-training data.

Subpopulation	Samples
Male	2657
Female	848
LOC	1824
PER	4614
ORG	2733
Positive	1824
Negative	4614
Neutral	2733
Temporal (Yes)	1835
Temporal (No)	8150
Chinese	35
British	414
Spanish	32
Indian	126
American	2050
Australia	60
Mexico	41
Canada	53
France	82
English people	38
British America	29
African American	36
Ashkenazi Jews	54
Chinese People	21
Mexican Americans	21

Table 9: Data distribution across subpopulations.