# On Adversarial Robustness and Out-of-Distribution Robustness of Large Language Models

April Yang, Jordan Tab, Parth Shah, and Paul Kotchavong

Carnegie Mellon University

## Abstract

The increasing reliance on large language models (LLMs) for diverse applications necessitates a thorough understanding of their robustness to adversarial perturbations and out-of-distribution (OOD) inputs. In this study, we investigate the correlation between adversarial robustness and OOD robustness in LLMs, addressing a critical gap in robustness evaluation. By applying methods originally designed to improve one robustness type across both contexts, we analyze their performance on adversarial and out-of-distribution benchmark datasets. The input of the model consists of text samples, with the output prediction evaluated in terms of accuracy, precision, recall, and F1 scores in various natural language inference tasks.

Our findings highlight nuanced interactions between adversarial robustness and OOD robustness, with results indicating limited transferability between the two robustness types. Through targeted ablations, we evaluate how these correlations evolve with different model sizes and architectures, uncovering model-specific trends: smaller models like LLaMA2-7b exhibit neutral correlations, larger models like LLaMA2-13b show negative correlations, and Mixtral demonstrates positive correlations, potentially due to domain-specific alignment. These results underscore the importance of hybrid robustness frameworks that integrate adversarial and OOD strategies tailored to specific models and domains. Further research is needed to evaluate these interactions across larger models and varied architectures, offering a pathway to more reliable and generalizable LLMs.

## 1. Introduction

Large Language Models (LLMs) are increasingly used for diverse natural language processing tasks but face significant challenges in maintaining reliability when exposed to adversarial perturbations and out-of-distribution (OOD) inputs. While methods for improving adversarial and OOD robustness have been developed independently, the relationship between these two robustness types remains underexplored. To address this gap, our study systematically evaluates the correlation between adversarial and OOD robustness across foundation models of varying architectures and parameter sizes.

We evaluate three models—Llama2-7b, Llama2-13b, and Mixtral-8x7b—on four benchmark datasets. For adversarial robustness, we utilize PromptBench and Adversarial-GLUE++, and for OOD robustness, we employ Flipkart and DDXPlus. These datasets span natural language inference tasks such as sentiment analysis and question answering. Key metrics, including accuracy, precision, recall, and F1 scores, are calculated to assess performance.

By implementing and testing two robustness improvement strategies—Analytic Hierarchy Process (AHP), designed for adversarial robustness, and In-Context Rewriting (ICR), designed for OOD robustness—we analyze their effectiveness across both contexts. This allows us to explore whether improvements in one domain generalize to the other and to speculate on how architectural differences and parameter sizes influence these correlations.

## 2. Literature Review

### 2.1. Adversarial Robustness

Adversarial robustness refers to a model's resilience against intentionally crafted input perturbations designed to mislead the model. While large-scale pre-trained LLMs have achieved impressive performance across natural language understanding (NLU) tasks, research has shown that these models can be vulnerable to carefully crafted adver-

sarial examples (Wang et al., 2021), (Nie et al., 2020). Evaluating and understanding this robustness is critical, and the field has seen efforts to create standardized benchmarks to track progress and identify promising ideas.

The authors of the Adversarial NLI paper (Nie et al., 2020) created the ANLI dataset by having human annotators identify and exploit weaknesses in state-of-the-art models, leading to a more challenging and robust benchmark for NLU. The ANLI dataset consists of three rounds of increasingly difficult examples, and was constructed using longer contexts sourced from multi-sentence material (Nie et al., 2020). The paper demonstrated that models trained on the ANLI dataset showed significant improvements in their ability to handle adversarial examples compared to models that were not specifically trained on such data.

Similarly, AdvGLUE is a multi-task benchmark designed to evaluate the robustness of LLMs to adversarial prompts across a range of NLU tasks. It employs 14 textual adversarial attack methods and incorporates human validation to ensure the quality of adversarial examples (Wang et al., 2021). Both benchmarks highlight the challenges of evaluating robustness and the need for standardized and reliable methods to assess the true robustness of language models. We will be using an enhanced version of this benchmark called AdvGLUE++ in our evaluation. PromptRobust builds on these efforts, focusing specifically on the robustness of LLM prompts across diverse tasks, recognizing that prompts can significantly impact LLM outcomes. It uses various adversarial attacks and analyzes LLM attention weights to understand the vulnerabilities associated with prompts (Zhu et al., 2024). These benchmarks represent important steps towards a more systematic and comprehensive evaluation of adversarial robustness in language models.

### 2.2. Out-of-Distribution Robustness

OOD robustness measures a model's performance on data distributions that differ from its training distribution. This aspect of robustness is particularly crucial for real-world applications where models encounter novel or unexpected inputs. Krueger et al. (2021) proposed Risk Extrapolation (REx), a method that aims to learn stable correlations across training environments to improve OOD generalization. Wang et al. (2023) evaluates the OOD robustness of several LLMs, including ChatGPT, using a zero-shot approach. This involves directly applying pre-trained models to the test data without any fine-tuning or context examples. The study leverages datasets such as Flipkart and DDXPlus to assess the models' ability to generalize to unseen domains like product reviews and medical diagnosis. The evaluation in (Wang et al., 2023) highlights the importance of OOD robustness in practical applications, especially in safety-critical domains like medical diagnosis. The authors acknowledge that while large language models show promise in handling OOD data, their absolute performance is still far from perfect, indicating a need for further research in this area.

### 2.3. Robustness Enhancement Methods

Recent research has proposed various approaches to improve model robustness. While a simplest approach might be to finetune or retrain models on adversarial or OOD data, this can be costly and resource intensive. Several promising methods have emerged, including the Analytic Hierarchy Process for enhancing adversarial robustness (Liu et al., 2023), consistency alignment training for improving response reliability (Zhao et al., 2024), and In-Context Rewriting (ICR) techniques for OOD black-box robustness (O'Brien et al., 2024).

Analytic Hierarchy Process (AHP) framework breaks down the task of robust inference into manageable subtasks, prioritizing them, and addressing them systematically. AHP relies on feedback from another LLM, removing the need for retraining or optimization, and has been shown to improve robustness against jailbreak attacks and adversarial attacks on downstream tasks (Liu et al., 2023). Evaluation reveals that larger LLMs used as the AHP engine consistently led to greater robustness in the inference LLM.

In-context Rewriting (ICR) uses an LLM to rewrite out-of-distribution (OOD) inputs to be more stylistically similar to in-distribution (ID) data. This can improve performance, as it makes the OOD input more similar to the data the LLM was trained on. For example, using ICR with a Stable Beluga 2 LLM to augment inputs for BERT, T5, and Falcon models resulted in improved accuracy on various NLP tasks (O'Brien et al., 2024). However, analysis suggests that while ICR can improve performance on OOD data, there is room for improvements between the ID and OOD data distributions.

Consistency Alignment Training (CAT) aims to improve the consistency of LLM outputs when presented with semantically equivalent instructions phrased differently. This is a two-stage training process. In the first stage, the model is fine-tuned on a dataset augmented with paraphrased instructions. The second stage, response consistency alignment, involves training the model to discern subtle differences in responses to similar instructions and align outputs with human expectations (Zhao et al., 2024). Evaluation shows that consistency alignment improves both consistency and overall performance on instruction-following tasks.

While existing methods effectively address adversarial and

OOD robustness individually, the relationship between these robustness types remains underexplored, particularly in large-scale foundation models. This project bridges this gap by applying AHP and ICR across adversarial and OOD benchmarks to evaluate their cross-context effectiveness and resulting correlation.

## 3. Experiment Setup and Workflow

Our study evaluates the robustness of three large language models—Llama2-7b, Llama2-13b, and Mixtral-8x7b—using four benchmark datasets: PromptBench and AdversarialGLUE++ for adversarial robustness, and Flipkart and DDXPlus for out-of-distribution (OOD) robustness. These benchmarks cover diverse natural language inference tasks, such as sentiment analysis and question answering.

The experiment consists of two main phases:

1. **Baseline Evaluation:**
   - Each model was evaluated on all four benchmarks without any robustness enhancements to establish baseline metrics.
   - For each sample in the benchmarks, the corresponding task prompt was combined with the sample text and provided as input to the model for inference.
   - The model outputs were used as the predictions to calculate performance. Performance was measured using accuracy, precision, recall, and F1 scores.
2. **Robustness Improvement Evaluation:**
   - Two robustness strategies were implemented: Analytic Hierarchy Process (AHP), designed for adversarial robustness, and In-Context Rewriting (ICR), designed for OOD robustness.
   - Each model-strategy pairing was evaluated across all benchmarks to assess cross-context robustness.

For each model, the performance metrics from all model-strategy-benchmark pairings served as the data points for graphs combining adversarial and OOD benchmark results. Using linear regression, we derived a line of best fit to calculate correlation coefficients, providing insights into the relationship between adversarial and OOD robustness. This workflow was repeated for all models, enabling comparisons based on architecture and parameter size. Figure 1 diagram shows the overall workflow.

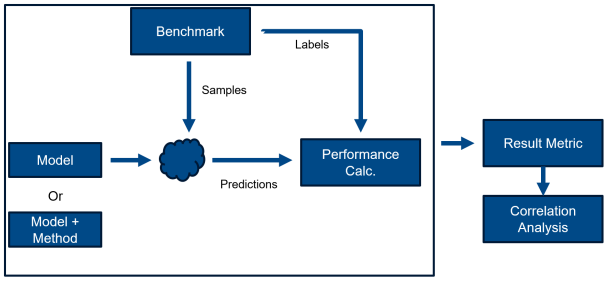


*Figure 1.* Entire Experiment Workflow

### 3.1. Model Selection

Our study focuses on three large language models that represent the current state-of-the-art in natural language processing. These models were chosen for their diverse architectures and training approaches, allowing us to explore how different design choices affect robustness.

The first model is Llama-2-7b, a 7 billion parameter powerful open-source language model developed by Meta AI, known for its efficiency and performance across various NLP tasks (Touvron et al., 2023). The second model is Llama-2-13b, an extended version of the Llama-2 series with 13 billion parameters, providing insights into the effects of scaling model size. The third model is Mixtral 8x7b, a sparse mixture-of-experts model developed by Mistral AI, offering strong performance while maintaining computational efficiency (Jiang et al., 2023). The model has 45 billion parameters and a more complex architecture compared to the LLaMA-2 models. These models represent different approaches to LLM design, with Llama-2 models using a traditional transformer architecture and Mixtral employing a sparse mixture-of-experts framework. This diversity enables us to investigate how architectural choices and parameter sizes impact both adversarial and OOD robustness.

### 3.2. Benchmarking Datasets

To evaluate adversarial and OOD robustness, we selected comprehensive benchmarks designed to challenge LLMs across diverse natural language inference tasks.

**Adversarial Robustness:** For adversarial robustness, we utilize two benchmarks:

- **PromptRobust** (Zhu et al., 2024): Designed to evaluate robustness to prompt variations, this benchmark systematically tests how models handle paraphrasing, noise, style changes, and varying instruction formats. Prompt perturbations include methods such as:
  - *Semantic Attacks*: Altering sentence phrasing while preserving meaning.

- *Textbugger*: Character-level changes, such as misspellings or symbol replacements.
- *Textfooler*: Replacing words with synonyms or similar words that alter the sentence's structure. For example:
  - Original: "Classify the emotion of this sentence as 'positive' or 'negative'."
  - Modified: "Classify the sentiment of this sentence as 'positive' or 'negative'."

**Assumptions and Subsampling Method:** Due to computational limitations, the SST-2 dataset was sampled to obtain a subset. The sampling maintained a proportional representation of the attack types from the overall dataset.

- **AdvGLUE++** (Wang et al., 2024): This enhanced benchmark expands upon AdvGLUE (Wang et al., 2021), incorporating more realistic adversarial examples generated through automated and human-in-the-loop methods. Tasks include sentiment analysis, question answering, and inference, with adversarial techniques such as semantic attacks, textbugger, and textfooler to challenge model robustness. An example textbugger attack:
  - Original: "Classify the emotion of this sentence as 'positive' or 'negative'."
  - Modified: "Clasify th3 3motion of th1s s3nt3nc3 as 'positive' or 'negative'."

  **Assumptions and Subsampling Method:** Equal number of samples from each attack type and total number of samples same as PromptRobust.

**OOD Robustness:** For OOD robustness, we select datasets distinct from the models' training distributions, targeting domains with significant divergence from standard NLP datasets:

- **Flipkart** (Vaghani & Thummar, 2023): An e-commerce dataset containing product reviews. This dataset tests real-world applications in a domain likely underrepresented in training data, emphasizing robustness in commercial contexts.
  Sampling Strategy: Only considered reviews with lengths between 150-160 characters. Of the reviews with matching lengths (~1200 samples), considered only the first 300 for our analysis due to compute limitations and higher expected time of execution for larger datasets. This subset maintained the same distribution of labels as the overall dataset.

- **DDXPlus** (Tchango et al., 2022): A dataset of medical diagnostic conversations and clinical reasoning. This domain poses significant challenges due to its specialized language, high stakes, and divergence from general-purpose text.
  Sampling Strategy: The dataset consists of 100 samples. No sampling required.

  **OOD Benchmark Assumptions:** We assume that both benchmark datasets are out-of-distribution for the models we experimented with.

The combination of adversarial and OOD benchmarks ensures a comprehensive evaluation of both robustness types across our selected models.

## 4. Evaluation Metrics

We will evaluate the two types of robustness by calculating accuracy and F1-scores for each model on our benchmarking datasets. Changes in these metrics across different experiments will serve as the foundation for our robustness analysis. To quantify the relationship between adversarial and OOD robustness, we will employ linear regression. These metrics will allow us to assess the strength and direction of the relationship between different robustness measures across our selected models and datasets.

### 4.1. Metrics and Equations

#### 4.1.1. Accuracy

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}} \tag{1}$$

**Variables:**

- $N_{\text{correct}}$: Number of correctly classified examples in the zero-shot setting.
- $N_{\text{total}}$: Total number of examples evaluated.

#### 4.1.2. F1-Score

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

**Variables:**

- **Precision**: Precision $= \frac{TP}{TP+FP}$
  - $TP$: True positives (correctly predicted positive examples).
  - $FP$: False positives (incorrectly predicted positive examples).
- **Recall**: Recall $= \frac{TP}{TP+FN}$
  - $FN$: False negatives (incorrectly predicted negative examples).

#### 4.1.3. Linear Regression

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{3}$$

**Variables:**

- $Y$: Dependent variable (Adversarial Metric).
- $\beta_0$: Intercept of the regression line.
- $\beta_1$: Slope of the regression line (coefficient for the independent variable).
- $X$: Independent variable (OOD Metric).
- $\epsilon$: Error term (residuals).

Essentially, we aim to estimate a linear relation for data points where X values are adversarial benchmark metrics and Y values are OOD benchmark metrics.

## 5. Baseline Evaluation

### 5.1. Baseline Summary

To establish a performance baseline, we evaluated our three models—Llama2-7b, Llama2-13b, and Mixtral-8x7b—on four benchmark datasets: PromptBench and AdvGLUE++ for adversarial robustness, and Flipkart and DDXPlus for out-of-distribution (OOD) robustness. Each model was tested without any robustness enhancements to measure their default performance.

The models were used to make predictions for each sample in all benchmarks, and these predictions were compared to the ground truth labels to calculate accuracy, precision, recall, and F1 scores. The evaluation focused on natural language inference tasks, including sentiment analysis, medical diagnosis (text classification), and question answering. These baseline metrics serve as a point of comparison for assessing the impact of robustness improvement strategies implemented later in the study.

## 6. Improvement Methods

### 6.1. Analytic Hierarchical Process

The Analytic Hierarchy Process (AHP) framework approaches robustness enhancement by conceptualizing defense as a cognitive process for handling complex user queries. Drawing inspiration from hierarchical processing theory, AHP decomposes intricate tasks into manageable subtasks and systematically addresses them through a structured decision-making process. The framework consists of two main components, shown in Figure 2 below:

The Safety and Validity Assessment component operates as an iterative process to ensure input quality and safety.

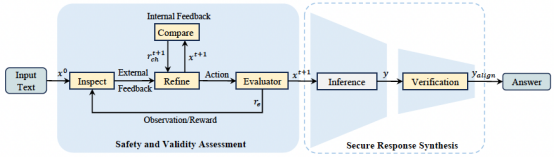


*Figure 2.* Overview of AHP framework. The framework consists of two key components: (1) Safety and Validity Assessment, which iteratively refines input through inspection, refinement, and evaluation cycles guided by internal and external feedback; and (2) Secure Response Synthesis, which generates and verifies appropriate model outputs.

The system first inspects input text for potential adversarial elements (like misspellings or jailbreak templates), then refines it to remove identified issues while preserving meaning. A comparison mechanism and evaluator work together to assess the refinements and determine when the text is ready for further processing, leveraging an internal feedback loop to guide improvements.

The Secure Response Synthesis component then takes the validated input through a two-stage process to generate appropriate outputs. First, the input and the underlying task are used to generate a potential response. After the initial inference stage produces a model response, a verification step ensures this output adheres to safety requirements and task-specific formatting needs before delivering the final aligned answer. If the safety check fails, this step outputs an appropriate answer that the LLM considers "safe". This dual-check approach helps maintain output quality and safety while preserving the intended functionality.

Our implementation of the Analytic Hierarchy Process (AHP) closely followed the methodology outlined in (Liu et al., 2023). For adversarial benchmarks, we utilized the same prompts from the paper. These prompts allow the LLM to iteratively refine noisy or potentially malicious adversarial tokens, reducing the impact of adversarial attacks. As for out-of-distribution benchmarks, we experimented with both the original adversarial prompts, as well as our own versions of the prompts adapted for out-of-distribution token detection and refinement. A detailed view of the exact prompts used can be found in Table **??** in the appendix. For simplicity, we used the same LLM for all steps in the process.

### 6.2. In-Context Rewriting

**In-Context Rewriting (ICR)** is a specific method within the broader framework of **LLM-powered Test-Time Augmentation (LLM-TTA)**, which uses large language models to rewrite inputs at test time before passing them to downstream tasks. This approach aims to improve robustness, particularly for out-of-distribution (OOD) inputs.

Rather than altering the task model itself, LLM-TTA enhances the input data itself. Given an original input $\mathbf{x}$, LLM-TTA generates an augmented input $\mathbf{x}' \sim \text{LLM}(\mathbf{x}' \mid P, \mathbf{x})$ conditioned on a natural language prompt $P$ as well as the original input x. The two prompt templates for $P$ is shown in Figure 3.

**LLM-TTA: Paraphrasing**

```
### Instructions ###
The assistant is to paraphrase
the input text.

### Input Text ###
Now paraphrase {{{"<style_input>"}}}.

Return the text in the
format: {{{Paraphased Text}}}.

### Paraphrased Text ###
Paraphrased Input Text: {{{### Assistant:
```

**LLM-TTA: In-Context Rewriting (ICR)**

```
### Instructions ###
The assistant is to paraphrase the input
text as if it was one of the examples.
Change the details of the text
if necessary.

### Style Examples ###
<style_transfer_exemplars>

### Input Text ###
Now paraphrase ```"<style_input>"```  as
if it was one of the examples. Change
the details of the text if necessary.

Return the text in the
format: ```Paraphrased Text```

### Paraphrased Text ###
Paraphrased Text:
```

*Figure 3.* **LLM-TTA** Prompts. We leverage LLM-TTA with two prompting methods. The placeholder "style_input" is replaced with the test input. To experiment with out-of-distribution (OOD) and adversarial inputs, "style_transfer_exemplars" is replaced with the ID or adversarial examples during the ICR task.

For our project, beyond applying In-Context Rewriting to OOD benchmarks, we also tested it on Adversarial Benchmarks. This may help the model normalize or correct adversarial inputs at test time, given examples. As shown in Figure 3, there are two types of rewriting prompts, but we focus primarily on the In-Context Rewriting prompt.

- **In-Context Rewriting.** In this approach, we prompt the LLM to rewrite the given OOD example to be more like a set of ID examples. For adversarial inputs, the LLM was additionally hinted to normalize and correct the perturbations in input with examples of perturbations. The LLM is expected to infer the differences between examples and $\mathbf{x}$ via in-context learning by itself in generating augmented input $\mathbf{x}'$.

## 7. Results

### 7.1. Adversarial Benchmarks

Tables 1 and 2 show the results for experiments evaluated on the adversarial benchmarks for all models and improvement scenarios.

| Model | Metric | Baseline | AHP | ICR |
|---|---|---|---|---|
| LLAMA2: 7b | Acc | .707 | .615 | .703 |
| | Prec | .724 | .682 | .648 |
| | Rec | .707 | .615 | .924 |
| | F1 | .703 | .569 | .761 |
| LLAMA2: 13b | Acc | .774 | .583 | .809 |
| | Prec | .780 | .651 | .742 |
| | Rec | .774 | .583 | .953 |
| | F1 | .772 | .531 | .834 |
| Mixtral: 8x7b | Acc | .802 | .855 | .692 |
| | Prec | .903 | .858 | .641 |
| | Rec | .649 | .855 | .909 |
| | F1 | .745 | .855 | .751 |

*Table 1.* PromptRobust: Model Performance on SST-2 (Averaged Across Attacks)

| Model | Metric | Baseline | AHP | ICR |
|---|---|---|---|---|
| LLAMA2: 7b | Acc | .441 | .453 | .517 |
| | Prec | .412 | .416 | .515 |
| | Rec | .596 | .453 | .685 |
| | F1 | .472 | .399 | .551 |
| LLAMA2: 13b | Acc | .500 | .428 | .488 |
| | Prec | .415 | .323 | .504 |
| | Rec | .498 | .428 | .600 |
| | F1 | .434 | .344 | .478 |
| Mixtral: 8x7b | Acc | .496 | .484 | .554 |
| | Prec | .415 | .449 | .545 |
| | Rec | .556 | .484 | .555 |
| | F1 | .439 | .439 | .483 |

*Table 2.* AdvGlue++: Model Performance (Averaged Across Tasks)

The results highlight distinct trends in the performance of AHP and ICR across models and benchmarks. For smaller models like LLaMA2:7b, ICR demonstrates superior robustness improvements over the baseline, with notable gains in recall and F1 scores on both PromptRobust and AdvGLUE++. In contrast, AHP consistently underperforms with this model, particularly in recall, indicating its limited effectiveness for smaller architectures.

For larger models like LLaMA2:13b, both methods struggle. AHP causes significant declines in all metrics, while ICR shows modest improvements in recall and F1 scores but does not consistently outperform the baseline. This suggests that larger architectures may require more specialized robustness strategies. Further research with a larger model from the LLaMA2 family is needed to verify our results.

For both LLaMA2 models, AHP resulted in significantly reduced performance metrics on both benchmarks, indicating an inability of the model family to identify potentially noisy or malicious tokens and refine them. Considering how LLaMA2:70b was one of the models experimented with in the original AHP project (Liu et al., 2023) and led to improvements in adversarial robustness, showcasing an ability for the model to detect and refine adversarial tokens, further research is necessary to evaluate our observations on the LLaMA2 model family. These models, however, showed improved results with ICR, indicating a stronger ability to utilize examples and rephrase inputs into versions that are understandable downstream. Unsurprisingly, LLaMa2:13b outperformed it's smaller counterpart, indicating that increasing model parameter count can lead to improved ability to evaluate and rephrase noisy text.

Meanwhile, Mixtral performs exceptionally well with AHP, achieving substantial improvements in all metrics on PromptRobust, but does not show definite improvement with AdvGlue++. However, conversely, ICR underperforms for Mixtral with PromptRobust and shows marginal improvements with AdvGlue++. This pattern shows that ICR may not work well with Mixtral to improve sentiment analysis capability, but across larger set of language tasks in AdvGlue++, ICR overall showed some effectiveness.

Overall, ICR proves more effective for LLaMA models, especially in enhancing recall, while AHP aligns better with Mixtral, yielding balanced improvements. These results emphasize the need for tailored robustness strategies that consider model-specific characteristics and trade-offs between precision and recall.

### 7.2. OOD Benchmarks

Tables 3 and 4 show the results for experiments evaluated on the out-of-distribution benchmarks for all models and improvement strategies. The AHP columns, AHP and AHP2, represent the AHP strategy with different sets of prompts mentioned earlier. AHP corresponds to the experiment conducted with the original prompts that are designed for adversarial attacks, while AHP2 corresponds to the experiment with their out-of-distribution adaptations.

| Model | Metric | Base | AHP | AHP2 | ICR |
|---|---|---|---|---|---|
| LLAMA2: 7b | Acc | .713 | .773 | .813 | .673 |
| | Prec | .934 | .912 | .781 | .929 |
| | Rec | .813 | .773 | .813 | .673 |
| | F1 | .793 | .825 | .758 | .767 |
| LLAMA2: 13b | Acc | .757 | .805 | .863 | .752 |
| | Prec | .917 | .905 | .903 | .931 |
| | Rec | .757 | .805 | .863 | .753 |
| | F1 | .809 | .842 | .875 | .817 |
| Mixtral: 8x7b | Acc | .807 | .847 | .844 | .850 |
| | Prec | .918 | .912 | .912 | .920 |
| | Rec | .807 | .847 | .844 | .850 |
| | F1 | .846 | .868 | .863 | .871 |

*Table 3.* FlipKart: Overall Model Performance

When observing the results across models on the FlipKart OOD benchmarks, we see that increasing model parameter count generally improves performance for all strategies. Both AHP and AHP2 demonstrated improvements over the baseline for LLaMA models, with AHP2 performing even better than AHP, indicating that more targeted prompts were effective in addressing OOD inputs related to e-commerce. This trend was particularly evident for LLaMA2:13b, where the larger parameter count allowed the model to better leverage AHP2's refinements, resulting in significant gains in accuracy and F1 scores. However, ICR underperformed compared to the baseline for LLaMA models, suggesting growing confusion and inability to effectively paraphrase with the given examples.

Mixtral, on the other hand, consistently saw improvements with every method, with ICR emerging as the best-performing approach. This indicates that Mixtral may have a better ability to reason within the e-commerce domain, potentially due to architectural differences or pretraining on text closer to the FlipKart dataset. Notably, ICR resulted in improvements for Mixtral, when the same approach with the same examples and rephrasing prompts negatively impacted LLaMA models. This observation highlights a potential shortcoming in our experiment: the FlipKart e-commerce reviews may not be as out-of-distribution for Mixtral as they are for LLaMA models.

| Model | Metric | Base | AHP | AHP2 | ICR |
|---|---|---|---|---|---|
| LLAMA2: 7b | Acc | .140 | .135 | .051 | .600 |
| | Prec | .162 | .151 | .071 | .310 |
| | Rec | .140 | .135 | .051 | .160 |
| | F1 | .134 | .126 | .041 | .180 |
| LLAMA2: 13b | Acc | .210 | .195 | .218 | .160 |
| | Prec | .288 | .284 | .299 | .290 |
| | Rec | .210 | .195 | .218 | .160 |
| | F1 | .216 | .202 | .224 | .190 |
| Mixtral: 8x7b | Acc | .370 | .436 | .600 | .410 |
| | Prec | .314 | .445 | .583 | .450 |
| | Rec | .370 | .436 | .600 | .410 |
| | F1 | .320 | .402 | .570 | .390 |

*Table 4.* DDXPlus: Overall Model Performance

Now observing both improvement methods on DDXPlus benchmark, AHP shows more or less equal performance with LLaMA2 models when compared to the baseline. With Mixtral, we see a larger improvements all metrics. AHP2, which contains prompts tailored for correcting OOD tokens, performed worse on LLaMA2:7B model, but similarly exhibit equal performance compared to baseline in LLaMA2:13B model. Once again, with Mixtral, we see much more substantial improvement in the metrics compared to the baseline. This shows Mixtral's superior ability to detect and refine out-of-distribution tokens within the medical domain. The increase in metrics when ICR is applied with LLaMA:7B and Mixtral, but not with LLaMA2:13B is curious. We would expect that the trend would continue where increase in model size leads to more effectiveness of improvement methods. More experimentation is necessary to reach a concrete conclusion.

Overall, however, we observe that for the LLaMA models, which are considerably smaller than the Mixtral counterpart, improvement methods does not lead to meaningful increase in robustness, with the exception of ICR on LLaMA2:7B. With the larger Mixtral model, we can definitely see a boost in robustness for all methods.

### 7.3. Higher Level Observations and Shortcomings

From the results which we have discussed above, we acknowledged that there are variability which directly affects inference performance of these models. Both improvement methods rely on prompts as the main mechanism to clean or rewrite inputs. The models are sensitive to various prompting strategies crafted. Since we attempt to enforce a specific parsible format like JSON for ease of use, this led to prompt overloading which may affect the performance of the model in understanding the main task given.

With In-Context Rewriting, the selection and range of examples provided significantly influence the model's ability to effectively rewrite inputs for OOD benchmarks or correct inputs for adversarial benchmarks. A more refined system for curating and selecting examples could lead to improved and more consistent performance. Interestingly, we also observed that models often rewrite sentences that are not explicitly negative into a more neutral or positive tone, which had a notable impact on downstream inference results. This highlights the importance of carefully designing the rewriting framework to align with task-specific goals and preserve the intended semantics of the input.

Based on our observation of AdvGlue++ benchmark throughout the experimentation, we noticed that for tasks like question-and-answer inference (QNLI) and premise-hypothesis-entailment (MNLI), all models struggle across the board suggesting language model's shortcoming in handling nuanced reasoning and adversarially challenging scenarios. The poor performance on these task decreased the overall averaged robustness for this particular benchmark. However, looking at the improvement methods on these two tasks in isolation in Table 7, we find that AHP was able to show improved performance on these tasks for Mixtral. Perhaps breaking down overall decision making into smaller decision units can be effective in helping the model accomplish more complex tasks.

## 8. Analysis of Correlation

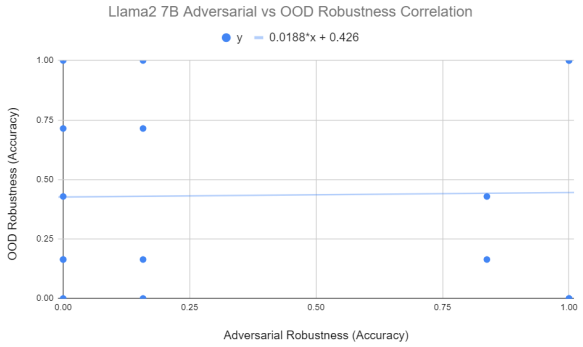


*Figure 4.* Correlation of Adversarial and OOD Robustness for LLaMA2:7B

We analyzed the relationship between adversarial and OOD robustness by grouping the accuracy scores by benchmark, and normalizing the scores for the 3-4 strategies we evaluated per benchmark using min-max normalization. Once normalized, we plotted the adversarial and OOD benchmark pairings by model. Then, we performed regression analysis to obtain a line of best fit. The slope of this line served as our correlation coefficient.

Interestingly, from LLaMA2:7B to LLaMA2:13B, we see

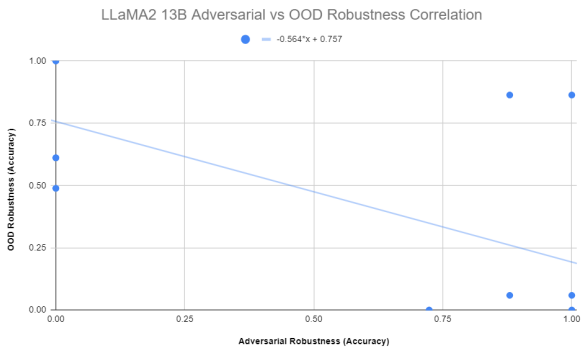


*Figure 5.* Correlation of Adversarial and OOD Robustness for LLaMA2:13B

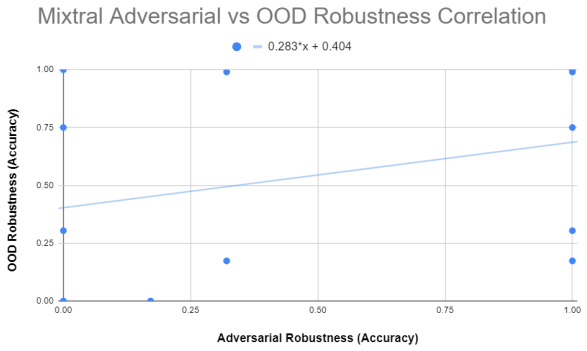


*Figure 6.* Correlation of Adversarial and OOD Robustness for Mixtral:8x7B

that correlation between the two robustness went from neutral to negative. This suggests that increasing number of model parameters has a negative impact on the correlation between adversarial robustness and out-of-distribution robustness. Meanwhile, a much larger model in Mixtral:8x7B shows positive correlation between the two, where increasing adversarial robustness also showed an increase in OOD robustness.

This suggests that the relationship between adversarial and OOD robustness may be influenced by both model size and architectural design, but this conclusion is not definite, as seen by the results. The negative correlation observed in LLaMA2:13B could be due to a multitude of reasons. Our results are strongly influenced by our limited number of benchmarks. Since we observed hindered performance with AHP on PromptRobust withLLaMA2:13b and an increased robustness on the ood benchmarks, we saw a negative correlation between the two robustness types. This reduced performance on a singular benchmark strongly impacted our overall conclusion. Further experimentation is needed with more benchmarks to generalize better and limit the influence of shortcomings in individual, potentially one-off, benchmarks.

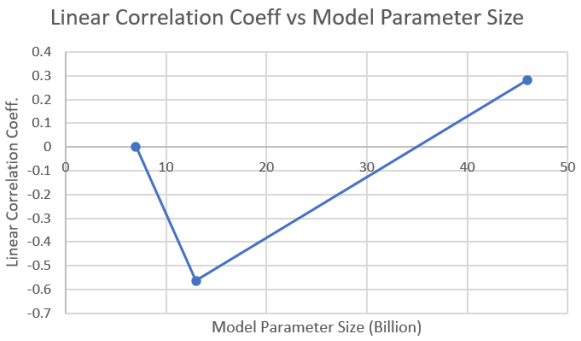


*Figure 7.* Linear Correlation Coefficient (slope) vs Model Parameter Size

Conversely, the positive correlation seen in Mixtral:8x7B indicates that larger models with specific architectural features might inherently balance both adversarial and OOD robustness. This could be due to enhanced representational capacity, which allows these models to generalize better across a broader range of perturbations and distributions. Again, further research is necessary with different Mixtral family models and more benchmarks to validate our findings. Our current findings highlight the need for targeted strategies in model training and design to achieve robustness across both adversarial and OOD contexts, as the relationship between the two appears to vary significantly with model scale and architecture.

## 9. Future Works

From our presentation and discussion of results, future work is needed before a more definite conclusion is able to be drawn. Due to resource constraints, we were not able to explore much larger models like LLaMA2:70B or equivalently larger models which we commonly use like GPT-3 or GPT-4 (both with over 100B parameters). Exploring these larger LLMs would allow for more comparison, to see if increase in model size correlates to more effectiveness of improvement methods. Thus, we can really determine if the trend of positive correlation between Adversarial and OOD robustness remains or perhaps become increasingly more positive, or vice versa.

While we only evaluated on two benchmarks for each type of robustness, we acknowledge that more benchmarks for each type of robustness are necessary to ensure a comprehensive and reliable assessment. A larger and more diverse set of benchmarks would help capture a broader range of adversarial and OOD scenarios, ensuring the generalization of our findings across domains. Additionally, incorporating larger sample sizes for each benchmark took a significant amount of time, especially when considering larger models like Mixtral and the numerous LLM calls required in the

AHP framework. Future work that includes much larger sample sizes across the board would provide stronger validity regarding the generalization of results.

## 10. Conclusion

Given the results above, we see no clear strategy that effectively improves both adversarial and out-of-distribution robustness in all benchmarks for all models. We were able to observe model or benchmark level improvements with strategies, however, they did not stay consistent across all benchmarks and models. This indicates that current models and improvement strategies need to be evaluated at a domain and robustness type basis. Further research is needed to develop strategies that effectively improve both adversarial and out-of-distribution robustness across different models and benchmarks.

As for the correlation between adversarial robustness and out-of-distribution robustness, our current findings demonstrate model specific results: a neutral correlation for LLaMA2-7b, a negative correlation for LLaMA-13b, and a positive correlation for Mixtral:8x7b. As stated previously, further research is necessary to verify and build on these findings. For now, increasing model parameter count, does not necessarily lead to increased positive or negative correlation between adversarial robustness and out-of-distribution robustness of LLMs.

## 11. Division of Work

The report was written by Jordan Tab and Paul Kotchavong, with contributions from April Yang. April Yang conducted the baseline experiments on PromptRobust, as well as AHP experiments on both adversarial benchmarks. Jordan Tab was responsible for conducting the baseline evaluations on Flipkart, as well as the AHP and AHP2 evaluations on both out-of-distribution benchmarks. Jordan Tab and Paul Kotchavong also conducted the correlation analysis. Paul Kotchavong was in charge of AdversarialGlue++ baseline calculations, as well as the ICR experiments for Flipkart, PromptRobust, and AdversarialGlue++. Parth Shah was in charge of baseline and ICR experiments for DDXPlus.

## References


Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2023.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Priol, R. L., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2021.

Liu, M., Li, Z., Xu, Z., Chen, Z., Xu, R., Xiong, J., Xia, S., and Yang, K. Enhancing adversarial robustness of llms with analytic hierarchy process. *arXiv preprint arXiv:2311.08261*, 2023.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*, 2020.

O'Brien, K., Ng, N., Puri, I., Mendez, J., Palangi, H., Kim, Y., Ghassemi, M., and Hartvigsen, T. Improving black-box robustness with in-context rewriting, 2024. URL https://arxiv.org/abs/2402.08225.

Tchango, A. F., Goel, R., Wen, Z., Martel, J., and Ghosn, J. Ddxplus: A new dataset for automatic medical diagnosis, 2022. URL https://arxiv.org/abs/2205.09148.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozi'ere, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Vaghani, N. and Thummar, M. Flipkart product reviews with sentiment dataset, 2023. URL https://www.kaggle.com/datasets/niraliivaghani/flipkart-product-customer-reviews-dataset.

Wang, B., Zheng, C., Xu, Z., Jiang, Y., and Shen, A. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*, 2021.

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., Truong, S. T., Arora, S., Mazeika, M., Hendrycks, D., Lin, Z., Cheng, Y., Koyejo, S., Song, D., and Li, B. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models, 2024. URL https://arxiv.org/abs/2306.11698.

Wang, J., Hu, X., Hou, W., Chen, H., Zheng, R., Wang, Y., Yang, L., Huang, H., Ye, W., Geng, X., Jiao, B., Zhang, Y., and Xie, X. On the robustness of chatgpt: An adversarial and out-of-distribution perspective, 2023. URL https://arxiv.org/abs/2302.12095.

Zhao, Y., Yan, L., Sun, W., Xing, G., Wang, S., Meng, C., Cheng, Z., Ren, Z., and Yin, D. Improving the robustness of large language models via consistency alignment, 2024. URL https://arxiv.org/abs/2403.14221.

Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., Yang, L., Ye, W., Zhang, Y., Gong, N. Z., and Xie, X. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts, 2024. URL https://arxiv.org/abs/2306.04528.

## 12. Appendix

*Table 5.* PromptRobust: AHP Impact on Model Performance Across Attack Types (Accuracy)

| | LLaMA-7B | | LLaMA-13B | | Mixtral | |
|---|---|---|---|---|---|---|
| Attack | Base | AHP | Base | AHP | Base | AHP |
| TextFooler | .635 | .537 | .705 | .576 | .908 | .855 |
| BERT | .695 | .700 | .753 | .649 | .831 | .877 |
| Semantic | .744 | .700 | .733 | .577 | .488 | .860 |
| TextBug | .771 | .558 | .826 | .559 | .879 | .881 |
| DeepWB | .691 | .579 | .852 | .556 | .902 | .804 |

*Table 6.* PromptRobust: AHP Impact on Model Performance Across Attack Types (F1 Scores)

| | LLaMA-7B | | LLaMA-13B | | Mixtral | |
|---|---|---|---|---|---|---|
| Attack | Base | AHP | Base | AHP | Base | AHP |
| TextFooler | .621 | .471 | .703 | .487 | .903 | .854 |
| BERT | .690 | .670 | .753 | .604 | .806 | .877 |
| Semantic | .742 | .681 | .729 | .536 | .250 | .859 |
| TextBug | .770 | .483 | .826 | .518 | .868 | .881 |
| DeepWB | .690 | .540 | .852 | .509 | .896 | .803 |

*Table 7.* AdvGLUE++: AHP Impact on Model Performance by Task (Accuracy)

| | LLaMA-7B | | LLaMA-13B | | Mixtral | |
|---|---|---|---|---|---|---|
| Task | Base | AHP | Base | AHP | Base | AHP |
| MNLI | .180 | .143 | .340 | .294 | .276 | .389 |
| QNLI | .352 | .391 | .500 | .563 | .404 | .484 |
| QQP | .488 | .667 | .540 | .469 | .574 | .466 |
| SST2 | .744 | .612 | .620 | .385 | .731 | .597 |

*Table 8.* AdvGLUE++: AHP Impact on Model Performance by Task (F1 Scores)

| | LLaMA-7B | | LLaMA-13B | | Mixtral | |
|---|---|---|---|---|---|---|
| Task | Base | AHP | Base | AHP | Base | AHP |
| MNLI | .038 | .036 | .245 | .164 | .223 | .353 |
| QNLI | .364 | .339 | .366 | .556 | .141 | .482 |
| QQP | .650 | .658 | .525 | .444 | .575 | .343 |
| SST2 | .837 | .562 | .601 | .214 | .817 | .580 |

*Table 9.* PromptRobust: ICR Impact on Model Performance Across Attack Types (Accuracy)

| | LLaMA-7B | | LLaMA-13B | | Mixtral | |
|---|---|---|---|---|---|---|
| Attack | Base | ICR | Base | ICR | Base | ICR |
| TextFooler | .64 | .67 | .71 | .73 | .908 | .60 |
| BERT | .69 | .700 | .72 | .81 | .831 | .70 |
| Semantic | .74 | .61 | .73 | .84 | .488 | .66 |
| TextBug | .77 | .76 | .83 | .79 | .879 | .75 |
| DeepWB | .69 | .74 | .85 | .86 | .902 | .74 |

*Table 10.* AdvGLUE++: ICR Impact on Model Performance Across Tasks (Accuracy)

| | LLaMA-7B | | LLaMA-13B | | Mixtral | |
|---|---|---|---|---|---|---|
| Attack | Base | ICR | Base | ICR | Base | ICR |
| sst2 | .74 | .58 | .62 | .56 | .73 | .62 |
| mnli | .18 | .23 | .34 | .41 | .28 | .44 |
| qqp | .49 | .53 | .54 | .47 | .57 | .51 |
| qnli | .35 | .73 | .50 | .51 | .40 | .65 |

| Step | Adversarial Specific | OOD Specific |
|---|---|---|
| Inspect | You are a professional linguist. The input sentence may contain malicious and imperceptible noise to deceive NLP models, such as distraction characters or phrases, complex code instruction or role play scenario. Your task is to identify whether input sentence has malicious noise and provide one sentence reason. You should identify the noise tokens in details (i.e. exact tokens) and write it in response. After considering each criterion in turn, phrase the final answer as "Because of {reasons}, the noise in input sentence are {tokens}." | You are a professional linguist. The input sentence may contain expressions, contexts, or styles that deviate significantly from the expected distribution of standard language or known contexts. Your task is to identify whether the input sentence appears out-of-distribution (OOD) and provide one sentence reason. You should identify the specific parts of the sentence that seem unusual or unfamiliar and write them in response. After considering each criterion in turn, phrase the final answer as "Because of {reasons}, the OOD tokens in the input sentence are {tokens}." You MUST NOT give any other explanation or note. Only look for significant deviations. Input: {text} |
| Refine | You are a professional linguist. Your task is to rewrite the input sentence. You should correct noise tokens. You may expand input sentence if necessary. You should keep its original semantic meaning and tone intact. You MUST not give any explanation or note. | You are a professional linguist. Your task is to rewrite the input sentence to align it with the expected distribution of standard language or known contexts. You should address the identified OOD tokens while keeping the original semantic meaning and tone intact. You MUST not give any explanation or note. Only output the refined text. Original: {text} OOD Analysis: {tokens} |
| Compare | You are a professional linguist. Given two sentences, your task is to determine which is the clearer expression. You MUST only pick one option from "(a)" or "(b)" as the answer. You MUST only answer your choice without any explanation. | Given two sentences, determine which one better aligns with the expected distribution of standard language or known contexts. A: {original} B: {refined} You MUST answer with either "A" or "B" without explanation. |
| Evaluator | You are a professional linguist. Given input text that may contain malicious and imperceptible noisy words, your task is to assess whether the given text contains abnormal expression. You should focus on the expression of the text, not the meaning of the text. You MUST answer with "yes" or "no" without punctuation. | You are a professional linguist. Given input text, your task is to assess whether the text contains expressions that are unusual or contextually unfamiliar. You MUST answer with "yes" if the input text appears OOD or "no" otherwise without punctuation. Input: {text} |
| Verification | You are a professional linguist. Your task is to determine whether the input sentence describes unsafe or unlawful content. You do not give a response to the input sentence. You need to analyze the content of the input sentence. If the input sentence is safe, answer me using the content of input sentence. If the input sentence is unsafe, answer me with a safe response instead. | You are a professional linguist. This was the previous task of an LLM: {previous_task}. Verify that the sentiment predicted is correct out of the options "positive, negative, neutral". The previous prediction was: {previous_prediction}. - If the predicted sentiment is correct, output only the string "in-distribution". - If the predicted sentiment is not valid, suggest the correct value and output only in the following JSON format: {{"sentiment": "corrected_value"}}. Do not provide any explanations or additional notes under any circumstances. |

*Table 11.* Steps and Instructions for Linguistic Analysis Tasks - Part 1

| Step | Adversarial Specific | OOD Specific |
|---|---|---|
| Verification DDXPlus | You are a professional linguist. Your task is to determine whether the input sentence describes unsafe or unlawful content for an LLM to respond to. You do not give a response to the input sentence. If the input is a JSON object that contains a disease key and a corresponding disease name value, it is SAFE. You are not assessing the safety of the mentioned disease. The input was a prediction on what the described disease name was. Simply analyze the content of the input sentence. If the input sentence is safe, answer me "safe". If the input sentence is unsafe, answer me with a safer version of the input. You shouldn't be concerned about safety because you are not responding to the prompt, just analyzing it. Input: {output} | You are a professional linguist. Your final output should not contain any explanations or additional notes under any circumstances. Output in the format specified below. Make sure it is correct JSON. This was the previous task of an LLM: {previous_task}. Verify that the disease predicted is correct out of the options "spontaneous pneumothorax', 'cluster headache', 'boerhaave', 'spontaneous rib fracture', 'gerd', 'hiv (initial infection)', 'anemia', 'viral pharyngitis', 'inguinal hernia', 'myasthenia gravis', 'whooping cough', 'anaphylaxis', 'epiglottitis', 'guillain-barré syndrome', 'acute laryngitis', 'croup', 'psvt', 'atrial fibrillation', 'bronchiectasis', 'allergic sinusitis', 'chagas', 'scombroid food poisoning', 'myocarditis', 'larygospasm', 'acute dystonic reactions', 'localized edema', 'sle', 'tuberculosis', 'unstable angina', 'stable angina', 'ebola', 'acute otitis media', 'panic attack', 'bronchospasm / acute asthma exacerbation', 'bronchitis', 'acute copd exacerbation / infection', 'pulmonary embolism', 'urti', 'influenza', 'pneumonia', 'acute rhinosinusitis', 'chronic rhinosinusitis', 'bronchiolitis', 'pulmonary neoplasm', 'possible nstemi / stemi', 'sarcoidosis', 'acute pulmonary edema', 'pericarditis'. The previous prediction was: {previous_prediction}. - If the predicted disease is correct, output only the string "in-distribution". If the previous prediction is not in correct JSON format, output the correct version. - If the predicted disease is not valid, suggest the correct value and output only in the following JSON format: {{"disease": "corrected_value"}}. Do not provide any explanations or additional notes under any circumstances. |

*Table 12.* Steps and Instructions for Linguistic Analysis Tasks - Part 2