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

April Yang Jordan Tab Paul Kotchavong Parth Shah

1. Introduction

Robustness, particularly the ability to generalize to unseen distributions, is a critical concern in the development of large language models in artificial intelligence. While numerous studies have explored adversarial robustness and out-of-distribution (OOD) robustness separately, the relationship between these two forms of robustness remains unclear. This project aims to investigate whether adversarial robustness improvements translate to improved OOD robustness in large language models.

We will focus on two state-of-the-art large language models: Meta-LLaMA/Llama-2-7b and Mixtral 8x7b. By examining these models, we seek to uncover insights into how pre-training data and model architecture influence different aspects of robustness.

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 language models (LLMs) have achieved impressive performance across natural language understanding (NLU) tasks, research has shown that these models can be vulnerable to carefully crafted adversarial 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. 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 large language models, 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 techniques for Adversarial/OOD black-box robustness (O’Brien et al., 2024).

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.

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.

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 the above approaches in this section show promise individually in evaluating Adversarial or OOD robustness, there remains a need for comprehensive studies that examine the relationship between adversarial and OOD robustness, particularly in the context of large-scale foundation models. This project aims to bridge this gap by systematically evaluating and comparing both types of robustness

across multiple foundation models.

3. Evaluation Metrics

We will evaluate two types of robustness by calculating common metrics, such as zero-shot accuracy, F1-score, and others, for each model on a range of 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 both Pearson’s correlation coefficient, Spearman’s rank correlation coefficient (Benesty et al., 2009), and 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.

3.1. Metrics and Equations

3.1.1. ZERO-SHOT ACCURACY

$$\text{Zero-Shot Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}} \quad (1)$$

Variables:

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

3.1.2. F1-SCORE

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

Variables:

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

3.1.3. PEARSON’S CORRELATION COEFFICIENT

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \quad (3)$$

Variables:

- X_i : Value of the i -th observation for Adversarial Metric X .
- Y_i : Value of the i -th observation for OOD Metric Y .
- $\bar{X}$: Mean of values in variable X .
- $\bar{Y}$: Mean of values in variable Y .

3.1.4. SPEARMAN’S RANK CORRELATION COEFFICIENT

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (4)$$

Variables:

- d_i : Difference between the ranks of each observation in the two variables.
- n : Total number of observations.

3.1.5. LINEAR REGRESSION

$$Y = \beta_0 + \beta_1 X + \epsilon \quad (5)$$

Variables:

- Y : Dependent variable (Adversarial Metric).
- β_0 : Intercept of the regression line.
- β_1 : Slope of the regression line (coefficient for the independent variable).
- X : Independent variable (OOD Metric).
- ϵ : Error term (residuals).

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

4. Baseline Evaluation

4.1. Model Selection

Our study will focus on two 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 Meta-LLaMA/Llama-2-7b, a 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 Mixtral 8x7b, a sparse mixture of experts model developed by Mistral AI,

offering strong performance while maintaining computational efficiency (Jiang et al., 2023). These models represent different approaches to large language model design, with Llama-2-7b using a traditional transformer architecture and Mixtral 7x8B employing a sparse mixture of experts. This diversity will allow us to investigate how different architectural choices impact both adversarial and OOD robustness.

4.2. Benchmarking Datasets

To evaluate adversarial robustness and OOD robustness in large language models, we will utilize several comprehensive benchmarks. For adversarial robustness, we will use PromptRobust (Zhu et al., 2024), a comprehensive benchmark designed to evaluate the robustness of LLMs to prompt variations - a critical real-world concern since users rarely phrase their queries in exactly the same way. The benchmark systematically evaluates how models handle different types of prompt perturbations, including paraphrasing, adding noise, changing writing styles, and varying instruction formats. We will also employ AdvGLUE++ (Wang et al., 2024). This is an enhanced version that builds upon the foundation laid by earlier adversarial benchmark, AdvGLUE (Wang et al., 2021). It introduces improvement techniques used in the generation of adversarial examples. The adversarial examples are created through both automated and human-in-the-loop processes, making them more realistic and challenging, and includes diverse tasks such as inference, sentiment, QA, etc.

For OOD robustness, we will select datasets and tasks distinct from Llama-2 and Mixtral’s training data, potentially focusing on specialized domain texts, non-English languages, datasets with temporal shifts, or historical text etc. Flipkart and DDXPlus were chosen for this task, as they were used in OOD benchmarks by Wang et al. (2023) for GPT robustness evaluation. Flipkart (Vaghani & Thummar, 2023) contains e-commerce data and product reviews, which likely differ significantly from the academic/web text commonly used in LLM training. It tests real-world commercial applications, making it practically relevant. DDXPlus (Tchango et al., 2022) contains medical diagnostic conversations and clinical reasoning, which differs substantially from general language use. Errors in this domain are particularly consequential, making it important for robustness testing. This combination of adversarial and carefully selected OOD benchmarks will allow us to comprehensively evaluate both types of robustness across our selected large language models.

5. Initial Results and Conclusion

5.1. Baseline Benchmarking Results

5.1.1. PROMPTROBUST¹

In our experimental evaluation, we conducted extensive testing using a base set of 100 samples for each attack type, with varying numbers of prompts per attack to ensure comprehensive coverage. A detailed statistics of the number of samples tested for each model and attack type can be found in the table below. This diverse range of prompt quantities across different attack types allows for both focused testing of specific attack vectors and broad coverage of potential vulnerabilities, with a total of over 30,000 individual evaluations conducted for each model.

Table 1. Number of Evaluations by Model & Attack. Mixtral shows a slightly higher total with similar distribution across attacks.

Attack	LLaMA-2	Mixtral
TextFooler	500	500
BERTAttack	10,200	11,300
CheckList	5,100	5,100
StressTest	400	400
Semantic	5,900	5,900
TextBugger	4,500	4,900
DeepWordBug	2,100	2,100
Total	28,700	30,200

Table 2. PromptRobust Results by Model. Mixtral outperforms LLaMA-2 on most metrics. PDR stands for Performance Drop Rate and refers to how much a model’s performance drop from evaluating original sample to perturbed sample

Metric	LLaMA-2	Mixtral
Accuracy	0.772	0.845
Precision	0.794	0.959
Recall	0.701	0.732
F1 Score	0.744	0.841
PDR	0.297	0.293
Original Accuracy	0.804	0.863
Perturbed Accuracy	0.769	0.843

¹ Author April Yang performed this benchmark of both models

Table 3. Perturbed F1 Scores by Model and Attack. Mixtral generally achieved higher scores.

Attack	LLaMA-2	Mixtral
TextFooler	0.851	0.903
BERTAttack	0.793	0.806
CheckList	0.874	0.928
StressTest	0.885	0.939
Semantic	0.250	0.250
TextBugger	0.872	0.868
DeepWordBug	0.827	0.896

Our experimental results reveal significant differences in robustness between Mixtral and LLaMA-2 across various adversarial attacks. Mixtral demonstrates superior overall performance, achieving higher accuracy (84.5% vs 77.2%) and notably better precision (95.9% vs 79.4%) compared to LLaMA-2. Both models show strong resilience against StressTest and CheckList attacks, with Mixtral achieving F1 scores of 0.939 and 0.928 respectively, while LLaMA-2 maintains respectable F1 scores of 0.885 and 0.874. However, both models exhibit substantial vulnerability to semantic attacks, with F1 scores dropping dramatically to 0.250. For traditional adversarial attacks like TextFooler and BERTAttack, Mixtral maintains better resilience (F1 scores of 0.903 and 0.806) compared to LLaMA-2 (0.851 and 0.793). The robustness analysis reveals that Mixtral better preserves its performance under attack, showing only a 2% drop in accuracy (86.3% to 84.3%) compared to LLaMA-2’s 3.5% drop (80.4% to 76.9%). These findings suggest that while larger models like Mixtral demonstrate improved robustness against adversarial attacks, significant vulnerabilities persist, particularly against semantic-level perturbations, highlighting the need for continued research in this direction.

5.1.2. ADVGLUE++²

The AdvGLUE++ dataset builds upon the original AdvGLUE dataset, enhancing it with adversarial texts generated using three open-source autoregressive models: Alpaca-7B, Vicuna-13B, and StableVicuna-13B (Wang et al., 2024). Similar to PromptRobust, adversarial samples are crafted using various attack methods, such as TextFooler, which utilizes embedding-based similarity, and TextBugger, which introduces typos. Following the structure of AdvGLUE, AdvGLUE++ is organized by task (Wang et al., 2021), including:

- **SST-2** for sentiment analysis,
- **QQP** for paraphrase similarity (determining whether

² Author Paul Kotchavong performed this benchmark of both models

two questions are semantically similar),

- **MNLI** for Natural Language Inference (identifying relationships such as entailment, contradiction, or neutrality between two sentences),
- **QNLI** for question-answering (verifying whether a given context sentence answers a question).

For our experiments, the dataset was made readily accessible, allowing us to evaluate models across multiple metrics, including zero-shot accuracy, F1-score, and Attack Success Rate. **Zero-shot accuracy** represents the model’s performance when prompted to complete each task without prior examples and how much the labels from adversarial labels match expected label. **Attack Success Rate** measures the effectiveness of adversarial attacks in disrupting the model’s predictions, providing insight into its resilience under challenging conditions. The Attack Success Rate (ASR) is calculated as:

$$\text{ASR} = \frac{\sum_{(x,y) \in D} \mathbb{1}[f(A(x)) \neq y]}{\sum_{(x,y) \in D} \mathbb{1}[f(x) = y]}$$

Given a benign dataset $D = \{(x^{(i)}, y^{(i)})\}_{i=1}^N$ consisting of N pairs of sample $x^{(i)}$ and ground truth $y^{(i)}$, for an adversarial attack method A that generates an adversarial example $A(x)$ given an input x to attack a surrogate model f where $\mathbb{1}$ is the indicator function. It is essentially the ratio of number of times the attack was successful (the model generates label opposite of the expected label) to number of times the model generates correct label.

Table 4. Zero-Shot Accuracy by Model and Type of Task

Model	SST-2	QQP	MNLI	QNLI	Avg. Acc
LLaMA-2	0.744	0.488	0.180	0.352	0.441
Mixtral	0.731	0.574	0.276	0.404	0.496

Table 5. Zero-Shot F1-Score by Model and Type of Task

Model	SST-2	QQP	MNLI	QNLI	Avg. F1
LLaMA-2	0.837	0.650	0.116	0.364	0.492
Mixtral	0.817	0.575	0.325	0.141	0.465

Table 4 and Table 5 highlights results obtained from the benchmark. The resulting metrics don’t match exactly to the results obtained by AdvGLUE++ paper, but since we are using different language models for our baseline evaluation, discrepancy in results are expected. Overall, Mixtral was able to achieve average zero-shot accuracy

higher than LLaMA-2 across all tasks (49.6% vs 44.1%). This suggests that Mixtral model may be more robust than LLaMa-2 in handling adversarial prompts, but not by a large margin. Both models performed well in handling adversarial prompts for sentiment analysis task with LLaMA’s 74.4% accuracy and 83.7% F1-Score compared to Mixtral’s 73.1% accuracy and 81.7% F1-Score. However, for Natural Language Inference tasks, such as MNLI and QNLI, both models are particularly vulnerable to adversarial attacks as shown by their low accuracy and F1-Score. These tasks will be of interest when we examine methods to improve a model’s robustness.

Additionally, the model sometimes struggles to respond reliably to prompts designed to elicit straightforward, singular answers like “yes”, “no”, “positive”, “negative”, or “neutral.” This inconsistency suggests potential success of the adversarial attacks in prompting the model to misinterpret or overcomplicate simple instructions. However, it’s also important to recognize that some of these response variations may stem from the model’s inherent architecture and response-generation process, which we cannot fully control or modify.

Table 6. Attack Success Rate (%) by Model and Type of Task

Model	SST-2	QQP	MNLI	QNLI	Avg. ASR
LLaMA-2	20.6	6.36	4.33	48.1	19.8
Mixtral	24.3	33.4	44.4	39.6	35.4

Our evaluation found that the attack success rate for Mixtral aligns closely with the rates reported in the AdvGLUE++ benchmark paper, demonstrating approximately 35-40% success. This consistency suggests that the adversarial attacks meaningfully disrupt the model’s responses in a significant portion of cases. Interestingly, although we observed that the Mixtral model performed a little better than LLaMA in terms of accuracy, the accuracy is relatively low and does suggest that substantial attacks were successful. LLaMA-2 exhibit low ASR, implying the model should have better accuracy which initial results doesn’t reflect. Further investigation is needed for the next step. Overall, both models demonstrate vulnerability to adversarial attacks, particularly in the context of Natural Language Inference (NLI) tasks. There is significant potential for the improvement methods to prove their effectiveness, and we will explore this further in the next phase of our investigation

5.1.3. FLIPKART³

The Flipkart dataset was collected by scraping product listings from the Flipkart e-commerce website, capturing detailed information on product attributes and customer feedback. It includes structured data such as product titles, categories, and more. For our analysis, we focused specifically on the product reviews and their corresponding sentiment labels. The dataset contains over 205,000 product reviews, from which we sampled only those with lengths between 150 and 160 characters, inclusive, resulting in 1,237 reviews. Of the reviews in our subset, roughly 75% were labeled 'positive', 22% were labeled 'negative', and 3% were labeled 'neutral'. We used both models to predict the sentiment of each review, categorizing it as 'positive,' 'neutral,' or 'negative.' In cases where the model output did not match the expected format, the prediction was counted as an 'error,' under the assumption that the model failed to follow directions as intended and should therefore be penalized.

Table 7. Model Performance on Flipkart Dataset

Metric	LLaMA-2	Mixtral
Accuracy	0.710	0.637
Precision	0.927	0.894
Recall	0.710	0.637
F1 Score	0.791	0.729

The results presented in Table 7 highlight the comparative performance of the LLaMA-2 and Mixtral models on the Flipkart dataset across four evaluation metrics: Accuracy, Precision, Recall, and F1 Score. When calculating recall, precision, and F1 scores, we used a weighted average due to the imbalanced nature of our dataset labels. The data indicates that LLaMA-2 consistently outperforms Mixtral 8x7B in all evaluated metrics, showing a particularly notable advantage in both Precision (0.927 vs. 0.894) and F1 Score (0.791 vs. 0.729). While both models achieved decent levels of performance, LLaMA-2's higher scores suggest it has a better ability to make accurate and reliable predictions on this dataset, despite its smaller size and simpler architecture, suggesting that more complex models are not necessarily needed to achieve better performance. Given that the Accuracy and F1 Scores for both models are relatively modest, further work is needed to enhance their performance on this benchmark. This improvement could provide a stronger foundation for applying these models to similar tasks.

³ Author Jordan Tab performed this benchmark of both models5.1.4. DDXPLUS⁴

The DDXPlus dataset was curated to support differential diagnosis tasks, containing detailed information on patient symptoms, diagnoses, and other clinical features. It includes a column titled Evidences, which we utilize to create prompts simulating dialogues between a doctor and a patient. In these dialogues, the doctor interrogates the patient with questions related to diagnosis, simulating a real-world clinical setting where symptoms are explored interactively. The DDXPlus dataset contains 1.3 million data-points, but for our use case we have utilized only 100 data-points to create prompts. This dataset enables us to model and analyze conversational approaches in diagnosis, aiding in the development of systems that can predict diagnoses based on structured symptom inputs. Our analysis focused on generating meaningful dialogue-based prompts, which were subsequently used to evaluate large models on recognizing medical conditions based on patient input. The model is asked to return a diagnosis from a set of 49 medical conditions. When the model's output did not align with the expected format, it was penalized, based on the assumption that it had not adhered to the intended instructions and thus warranted a penalty.

Table 8. Model Performance on DDXPlus Dataset

Metric	LLaMA-2	Mixtral	gpt-3.5-turbo
Accuracy	0.08	0	0.43
Precision	0.1016666667	0	0.491
Recall	0.08	0	0.43
F1 Score	0.08406926407	0	0.4330779221

The results in Table 8 present the comparative performance of LLaMA-2, Mixtral, and GPT-3.5-turbo on the DDXPlus dataset across four key evaluation metrics: Accuracy, Precision, Recall, and F1 Score. The DDXPlus dataset, with its specialized medical language and terminology, poses a unique challenge for general-purpose language models, as medical datasets of this nature are not widely available or commonly used in model training. The LLaMA-2 and Mixtral models evaluated here are both 7B parameter versions, chosen due to hardware limitations. However, to assess whether a larger model could yield better results, we also evaluated GPT-3.5-turbo, which has 135B parameters, using OpenAI's API. Among the models, GPT-3.5-turbo achieved the highest scores across all metrics, with an Accuracy of 0.43 and F1 Score of 0.433, suggesting a significantly stronger performance on this medical domain dataset than the 7B parameter models. LLaMA-2 showed limited performance, while Mixtral achieved 0 accuracy. These findings indicate that larger models with higher pa-

⁴ Author Parth Shah performed this benchmark of both models

parameter counts, such as GPT-3.5-turbo, may have a greater capacity for handling domain-specific language and complex medical terminology, underscoring the potential benefits of scaling up model size when working with specialized datasets.

5.2. Summary of Initial Results

Based on our comprehensive evaluation across multiple benchmarks, several key patterns emerge:

- **Mixed Performance Patterns:** While Mixtral shows superior performance on PromptRobust benchmarks (84.5% vs 77.2% accuracy), the advantage doesn't consistently translate across all evaluation scenarios. On AdvGLUE++, both models show task-dependent performance, with Mixtral having better average accuracy (49.6% vs 44.1%) but showing mixed results on individual tasks. This suggests that robustness advantages may be highly context-dependent.
- **Common Vulnerabilities:** Both models demonstrate similar weaknesses, particularly against semantic attacks (both showing F1 scores of 0.250), indicating that current architectural differences may not address fundamental vulnerabilities to semantic-level perturbations.
- **OOD Performance Reversal:** Interestingly, on the Flipkart OOD dataset, LLaMA-2 outperforms Mixtral on all metrics, suggesting that superior adversarial robustness doesn't necessarily translate to better OOD performance. This highlights the potential independence of these two types of robustness.
- **Attack Type Sensitivity:** Both models show varying levels of resilience to different attack types, with strongest performance against StressTest and CheckList attacks (F1 scores > 0.87) and weakest against semantic attacks. This pattern suggests that current model architectures may be better equipped to handle surface-level perturbations compared to deeper semantic modifications.

6. Next Steps

6.1. Correlation Analysis

- **Calculation of Pearson and Spearman correlation coefficients:** We will analyze correlations between adversarial and OOD performance metrics, examine performance patterns across different attack types, and investigate relationships between model architecture characteristics and robustness measures.
- **Analysis of potential factors affecting robustness relationships:** We will investigate how model architecture

(dense vs. sparse), pre-training data distribution, and model scale influence the relationship between adversarial and OOD robustness. This analysis will help identify key architectural decisions that contribute to improved robustness. For this section, we are planning on utilizing the Mistral-7b model as well, to allow us to compare different architectures, training data distributions, and model scale. We will compare the performance of Mixtral-8x7b and Mistral-7b to keep training data distribution consistent and draw conclusions on model architecture and scale. We will compare Mistral-7b to LLaMA-2-7b to analyze the impact of different training data distributions.

6.2. Implementation of Improvement Methods

- **Analytic Hierarchy Process (AHP):** We will implement the hierarchical decision-making framework and evaluate its impact on both adversarial and OOD performance. The effectiveness of AHP will be compared across different attack and task types to understand its generalizability and limitations.
- **In-Context Rewriting:** We will implement LLM-TTA, which uses Large Language Models for Test-Time Augmentation (LLM-TTA) (O'Brien et al., 2024) of inputs and evaluate its effectiveness on both adversarial and OOD samples. Special attention will be paid to analyzing the impact on different types of distribution shifts and understanding how the rewriting process affects model robustness.

6.3. Comprehensive Evaluation

- **Comparative Analysis:** We will assess the effectiveness of each improvement method across different attack types, various OOD scenarios, and model architectures. This will provide insights into which methods work best in specific contexts.
- **Trade-off Analysis:** We will examine the balance between adversarial and OOD robustness, evaluate performance versus computational cost considerations, and study the relationship between generalization and specific task performance.
- **Recommendations Development:** Based on our findings, we will develop concrete recommendations for optimal robustness enhancement strategies, guide model selection for different use cases, and identify promising directions for future research.

References

- Benesty, J., Chen, J., Huang, Y., and Cohen, I. Pearson correlation coefficient. *Noise reduction in speech processing*, pp. 1–4, 2009.

- 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., Binias, 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., Rozière, 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>.