

Comparing LLM Text Annotation Skills: A Study on Human Rights Violations in Social Media Data

Poli Apollinaire Nemkova,¹ Solomon Ubani,² Mark V. Albert¹

¹College of Computer Science and Engineering, University of North Texas, USA

²Intuit, USA

poli.nemkova@unt.edu, solomon_ubani@intuit.com, mark.albert@unt.edu

Abstract

Large language models (LLMs) have shown promise in tasks requiring nuanced textual understanding. This study evaluates state-of-the-art LLMs—GPT-3.5, GPT-4, LLaMA3, Mistral 7B, and Claude-2—for annotating a multilingual dataset of social media posts in Russian and Ukrainian, focusing on identifying human rights violations. Model performance is compared against a human-annotated gold standard across 1000 samples, with analysis of zero-shot and few-shot prompting in English and Russian.

Key findings show GPT-4.0 excels in precision and recall, making it ideal for high-stakes tasks. Aligning prompt language with dataset language improves performance, particularly for open-source models like LLaMA3 and Mistral-7B, which benefit from few-shot prompting but require additional validation for sensitive applications. Models, like humans, tend to struggle with ambiguous cases, particularly those involving indirect or nuanced references. Closed-source models demonstrate robust zero-shot capabilities with marginal gains from few-shot setups.

This study highlights the importance of linguistic alignment, prompt strategy, and model selection, offering practical insights for deploying LLMs in multilingual, high-impact domains like human rights monitoring.

1 Introduction

The emergence of large language models (LLMs) has revolutionized natural language processing (NLP), pushing the boundaries of what machines can achieve in understanding, generating, and classifying text. Models such as GPT-3.5, GPT-4, Claude-2, and open-source alternatives like LLaMA and Mistral-7B have demonstrated extraordinary capabilities across a wide range of applications, from translation to content summarization. However, the deployment of LLMs for high-stakes, domain-specific tasks, such as identifying references to human rights violations in multilingual social media data, remains a critical yet underexplored area of research.

The domain of human rights monitoring presents unique challenges. It involves the analysis of contextually rich, multilingual, and emotionally charged texts, often requiring nuanced comprehension and interpretation of implicit or in-

direct references. Traditionally, these tasks have been conducted by human annotators, whose expertise ensures high accuracy but incurs significant costs in time and resources. LLMs offer an unprecedented opportunity to automate and scale such tasks, but their ability to handle the inherent subjectivity and complexity of human rights-related data is not well understood. Moreover, questions remain about how factors such as prompt design, linguistic alignment, and model architecture influence performance.

In this study, we systematically evaluate the performance of leading LLMs and open-source alternatives on the binary classification task of detecting references to human rights violations in social media posts written in Russian and Ukrainian. Using a dataset curated from public Telegram channels, we compare models under two prompting configurations—zero-shot and few-shot learning—and investigate the effect of prompt language (English vs. Russian) on model performance. Additionally, we examine differences between closed-source models, such as GPT-4 and Claude-2, and open-source models, such as LLaMA and Mistral-7B, to understand the trade-offs between cost, accessibility, and performance.

Key contributions of this study:

- **Comprehensive Benchmarking Across Models and Prompts:** We provide a comparative analysis of closed- and open-source LLMs under different prompting styles (zero-shot vs. few-shot) and languages (English vs. Russian), highlighting performance gaps and strengths.
- **Insights Into Prompting Strategies:** The study demonstrates that open-source models benefit significantly from few-shot prompting with examples, whereas closed-source models exhibit robust performance across both zero-shot and few-shot settings.
- **Impact of Linguistic Alignment:** We show that aligning the prompt language with the data language (e.g., Russian prompts for Russian texts) consistently improves model performance across architectures, emphasizing the importance of multilingual capability in real-world tasks.
- **Error Analysis and Practical Guidance:** Through a detailed analysis of error types and model disagreements with human annotations, we provide actionable insights for selecting and optimizing LLMs in high-context, sensitive domains.

- **Cost-Effectiveness of Open-Source Models:** Despite their lower baseline performance, open-source models demonstrate strong potential when optimized with appropriate prompting strategies, making them viable alternatives for resource-constrained scenarios.

This work advances the understanding of LLM capabilities in multilingual, high-stakes applications and offers practical guidelines for leveraging these models effectively in real-world human rights monitoring tasks. By addressing critical gaps in model evaluation and exploring the interplay of architecture, prompt design, and language alignment, this study lays the groundwork for future research and development in this domain.

2 Literature Review

The emergence of Large Language Models (LLMs) has transformed Natural Language Processing (NLP), particularly in text annotation, enabling automation at a scale and speed unattainable through traditional methods. This review explores the comparative effectiveness of LLMs and human annotators in text annotation tasks (Aldeen et al. 2023; Nasution and Onan 2024), highlighting their strengths, limitations, and the pressing need for further research into their labeling performance, especially for complex and high-context tasks.

Efficiency and Cost-Effectiveness. LLMs have demonstrated substantial potential to reduce the time and cost associated with text annotation. For instance, integrating LLMs with human expertise in medical information extraction has been shown to significantly reduce manual effort while maintaining high accuracy, facilitating the rapid creation of labeled datasets (Goel et al. 2023). Similarly, the CoAnnotating framework leverages LLMs to complement human annotators by allocating annotation tasks based on uncertainty, achieving a 21% performance improvement over random baselines (Li et al. 2023).

However, these gains are task-dependent. For more nuanced annotations, such as legal or ethical contexts, the costs of fine-tuning or prompt engineering can be substantial, offsetting the initial efficiency benefits (Brown and Others 2020; Bommasani et al. 2021). This highlights a critical need to understand where LLMs can outperform or meaningfully complement traditional annotation workflows.

Accuracy and Reliability. While LLMs excel in generalization, particularly in zero-shot and few-shot learning, their performance varies significantly across tasks and datasets. For example, ChatGPT-4 has outperformed both expert classifiers and crowd workers in annotating political Twitter messages, demonstrating higher accuracy and consistency (Törnberg 2023). Similarly, GPT-3 and GPT-4 have shown strong performance in summarizing legal documents and medical records, often rivaling domain experts (Šavelka 2023; Shaib et al. 2023; Takagi et al. 2023).

However, studies have also highlighted limitations. For instance, LLMs frequently underperform on tasks requiring deep contextual understanding or domain-specific expertise, often lagging behind fine-tuned smaller models trained on expert-annotated data (Plaza-del Arco, Nozza, and Hovy

2023; Pangakis, Wolken, and Fasching 2023). Furthermore, while LLMs achieve high accuracy in standard tasks, their susceptibility to generating false positives or negatives in edge cases or nuanced scenarios remains a critical limitation.

Domain-Specific Knowledge. Human annotators possess domain-specific knowledge that is often critical for accurate text annotation. In tasks requiring subtle contextual judgments, such as medical diagnosis or human rights monitoring, small models trained on curated datasets have been shown to outperform general-purpose LLMs (Lu et al. 2023). These findings underscore the limitations of pre-trained LLMs when applied to tasks requiring specialized knowledge and suggest that domain adaptation remains a key challenge.

Research also indicates that while in-context learning (e.g., few-shot prompting) improves LLM performance in specialized domains, it is not always sufficient to match the nuanced understanding of human annotators (Shin et al. 2020). This highlights a need for further exploration into how prompting strategies can bridge this gap, particularly in complex labeling tasks.

Bias and Ethical Considerations The use of LLMs in text annotation raises significant concerns about bias and ethical implications. While models like ChatGPT-4 have shown reduced bias compared to human annotators in some tasks (Törnberg 2023), they are far from neutral. Biases inherent in training data often manifest in annotations, particularly in sensitive or controversial domains such as politics, law, or human rights (Fisher et al. 2024; Bender et al. 2021; Zimmer 2010).

Auditing frameworks such as ALLURE aim to address these challenges by incorporating failure cases into the evaluator through in-context learning, highlighting the importance of iterative improvement (Hasanbeig et al. 2023). However, such frameworks remain experimental, and the broader challenges of transparency and accountability in LLM-based annotation persist (Bommasani et al. 2021).

Open-Source vs. Proprietary Models. Open-source models such as LLaMA, Mistral, or FLAN-T5 have emerged as viable alternatives to proprietary systems like GPT-4, offering cost-effectiveness, transparency, and superior data protection (Touvron et al. 2023; Raffel et al. 2020; Jiang et al. 2023). Studies have shown that open-source LLMs can outperform crowd-sourced services like Amazon MTurk in specific tasks and even achieve competitive performance against proprietary models in domains like sentiment analysis and legal classification (Wong 2024).

However, open-source models are often more sensitive to prompting strategies and lack the robust pretraining data that benefits proprietary systems, particularly in multilingual and context-sensitive tasks. This creates a trade-off between cost and performance that must be better understood through systematic benchmarking (Bommasani et al. 2021).

Need for Research on LLM Labeling. Despite their promise, the use of LLMs in labeling tasks remains underexplored, particularly for complex scenarios requiring nuanced understanding. Most existing studies focus on standard tasks (e.g., sentiment analysis, named entity recogni-

tion), where LLMs often achieve near-human performance. However, tasks involving ambiguity, indirect references, or ethical considerations, such as detecting human rights violations in multilingual social media posts, are far more challenging and less studied (Artstein and Poesio 2008; Törnberg 2023).

Understanding how LLMs navigate these challenges—particularly in zero-shot and few-shot settings—is critical for assessing their suitability for real-world applications. This study addresses this gap by evaluating LLM performance in labeling complex, multilingual datasets, exploring the interplay of prompt design, linguistic alignment, and model architecture. By providing insights into model strengths, limitations, and error patterns, this work contributes to the growing body of research on optimizing LLMs for domain-specific annotation tasks.

3 Experiment Design

General Framework

This study assesses the capability of large language models (LLMs) to annotate a dataset of Russian and Ukrainian social media posts for references to human rights violations, comparing their performance against a gold-standard dataset created by human annotators. The dataset underwent a rigorous double-annotation process, with disagreements resolved by a senior adjudicator, resulting in a finalized benchmark label set.

The initial Cohen’s Kappa score (before adjudication) for annotator agreement was 0.63, indicating substantial agreement. Out of 1000 samples, human annotators disagreed on 184 instances, which were subsequently resolved by the adjudicator. Despite the substantial agreement, the annotation task remains challenging for humans.

By comparing the outputs of LLMs with the human-labeled data, this study explores the extent to which these models can replicate human annotations in a multilingual, context-sensitive domain. Furthermore, the study investigates the performance of LLMs on the 816 fully agreed-upon samples and the 184 samples with initial disagreements separately, providing a nuanced evaluation of their capabilities.

3.1 Dataset

To evaluate model annotation performance, we used a sample from a larger Human Rights Violations (HRV) dataset (Nemkova et al. 2023). The dataset comprises social media posts collected from public Telegram¹ news channels, primarily focused on the Russia-Ukraine conflict. These posts are written predominantly in Russian, with a smaller proportion in Ukrainian (966 post in Russian, 34 in Ukrainian). Telegram has been increasingly used for public discourse and news dissemination, particularly in the context of conflict reporting².

Each post was annotated for the presence or absence of references to situations falling under the category of hu-

man rights violations. The annotation process involved two trained volunteers who are native speakers of either Russian or Ukrainian, with proficiency in understanding both languages. The label is binary: the positive class indicates the presence of a reference to a human rights violation (HRV), while the negative class covers posts where HRV references are absent or unclear. In cases where the two annotators disagreed, a third annotator with domain-specific expertise adjudicated the final label. Such adjudication processes are widely used in NLP to resolve disagreements and ensure high-quality ground truth annotations (Artstein and Poesio 2008).

Inter-annotator agreement was measured using Cohen’s Kappa (Cohen 1960) with value 0.63, which indicated substantial agreement between the two primary annotators. This level of agreement reflects the inherent complexity of the task, as detecting references to human rights violations often involves interpreting nuanced or implicit contextual cues. However, the adjudication process ensured the creation of a high-quality gold-standard dataset suitable for benchmarking.

For this study, a sample of 1000 posts was selected from the dataset. The sample is moderately imbalanced, with 517 of posts labeled as belonging to the positive class (HRV present) and the remaining 483 labeled as negative (no HRV or unclear). Class imbalance is a common challenge in such tasks and can impact model performance (Japkowicz and Stephen 2002). This curated sample served as the benchmark against which the LLMs’ annotation performance was evaluated.

3.2 Method

We evaluated the following LLMs to capture a diverse range of capabilities, including both state-of-the-art proprietary systems and open-source models:

GPT-4.0 and GPT-3.5 Turbo (OpenAI): GPT-4.0 represents the current cutting-edge in language modeling, demonstrating superior performance across a range of multilingual NLP tasks (OpenAI 2023). GPT-3.5 Turbo, while less powerful, is more cost-effective and serves as a strong baseline for comparison in terms of practical deployment.

Claude-2 (Anthropic): Claude-2 is designed with a focus on safety and alignment, particularly relevant in domains involving sensitive or ethical considerations (Anthropic 2023).

LLaMA-3.2-1B (Meta): This smaller-scale open-source model (Touvron et al. 2023) was included to evaluate the feasibility of using lightweight, publicly available models for multilingual annotation tasks.

Mistral-7B: Another open-source model, Mistral-7B (Mistral AI 2023), was selected for its promising performance on various NLP tasks while maintaining a smaller computational footprint.

The selection of models reflects a balance between high-performing proprietary systems and resource-efficient open-source models, providing a comprehensive perspective on LLM performance across different architectures and scales.

Experimental Setup. All experiments were conducted in Python in a Google Colab environment to ensure accessibility and reproducibility. High RAM A100 was utilized. None

¹<https://telegram.org/>

²<https://time.com/6158437/telegram-russia-ukraine-information-war/>

of the models were fine-tuned on the dataset. Instead, we utilized their zero-shot and few-shot capabilities to assess their out-of-the-box performance, a common real-world scenario where task-specific fine-tuning is infeasible due to data limitations or computational constraints.

*Prompting and Evaluation.*³ Two prompting strategies were employed:

- *Zero-shot prompting:* Direct task instructions without providing examples. This evaluates the models’ ability to generalize to the task based on pretraining alone.
- *Few-shot prompting:* Instructions augmented with multiple labeled examples. This approach leverages in-context learning to improve task-specific alignment (Brown and Others 2020).

Prompts were tested in both English and Russian, aligning with the linguistic characteristics of the dataset. Russian prompts generally outperformed English prompts across models, demonstrating the importance of prompt language alignment in multilingual tasks. The best-performing prompts for each model and setting are detailed in the GitHub.

Performance was evaluated using precision, recall, F1 score, and accuracy, with a focus on balancing sensitivity and specificity, particularly given the dataset’s moderate positive class imbalance.

Reproducibility. To ensure replicability, all experiments were conducted with fixed random seeds. The seeds, along with detailed prompt designs and hyperparameter settings, are provided in the project GitHub. This adherence to reproducibility standards aligns with best practices in the NLP community (Pineau et al. 2021).

4 Results

The performance metrics for all models are detailed in Figure 1. Among the evaluated models, **GPT-4.0** demonstrated the highest overall performance, particularly excelling in zero-shot prompting with English prompts, where it achieved the top F1 score of 0.84 and an accuracy of 0.82. This performance highlighted its strong balance between precision (0.78) and recall (0.92). Similarly, **GPT-3.5** performed well, particularly in few-shot prompting with Russian prompts, achieving an F1 score of 0.76 and an accuracy of 0.70, showcasing its capability in scenarios where high recall (0.92) is essential.

In contrast, **LLaMA-3** and **Claude-2** exhibited limitations. LLaMA-3 struggled with consistently low accuracy and F1 scores, particularly in zero-shot Russian prompting (F1 = 0.26, Accuracy = 0.50). Claude-2 showed imbalanced performance, achieving moderate precision but relatively lower recall across all setups. For example, in few-shot prompting with English, it reached an F1 score of 0.58 but had recall as low as 0.45, signaling challenges in achieving generalizable results.

³All prompts used in this experiment are available at the experiment GitHub link: https://GitHub.com/PoliNemkova/LLM_labeling_skills

The results emphasize the importance of aligning prompt language with the dataset. Russian prompts generally outperformed English prompts across models, particularly with GPT-3.5 and GPT-4. Few-shot prompting tended to enhance recall, as seen in GPT-3.5 and Mistral-7B, but sometimes led to trade-offs in precision. Meanwhile, zero-shot prompting with GPT-4 delivered the most balanced and robust outcomes, making it a reliable choice for diverse task requirements.

5 Error Analysis and Discussion

5.1 Model Performance Overview

This study evaluated the performance of multiple language models (LLMs) on the task of detecting human rights violations in a multilingual dataset comprising Russian and Ukrainian social media posts. The dataset’s positive class ratio reflects a moderate imbalance, necessitating careful consideration of both recall (to minimize false negatives) and precision (to limit false positives). The analysis underscores the critical impact of model architecture, prompting strategy, and prompt language alignment on model performance.

Best Performing Models

GPT-4.0 emerged as the strongest performer across all settings, demonstrating consistent superiority in both zero-shot and few-shot scenarios. In particular, zero-shot prompting with English prompts achieved the highest overall F1 score (0.84) and accuracy (0.82), showcasing GPT-4’s exceptional ability to generalize effectively from minimal context. This performance reflects a robust balance between precision (0.78) and recall (0.92), which is critical for tasks involving human rights violations where both false positives and false negatives carry significant consequences.

GPT-3.5 also showed competitive performance, particularly in few-shot prompting with Russian prompts (F1 = 0.76, Recall = 0.92), demonstrating its ability to capture a high proportion of true positives. While GPT-3.5 fell short of GPT-4’s overall performance, its cost-efficiency and recall-centric behavior in certain settings make it a strong candidate for tasks prioritizing coverage.

Both GPT-4 and GPT-3.5 exhibited a pronounced benefit from using Russian prompts, suggesting their enhanced alignment with the linguistic and contextual features of the dataset. These results emphasize the importance of prompt language matching, particularly when working with models trained primarily on multilingual or English-centric corpora.

Underperforming Models

LLaMA-3 and **Claude-2** consistently underperformed across most settings. LLaMA-3 struggled with balancing precision and recall, achieving an F1 score as low as 0.26 and an accuracy of 0.50 in zero-shot Russian prompting. While its recall improved dramatically in few-shot settings with Russian prompts (0.99), the corresponding drop in precision (0.52) undermines its practical applicability. Similarly, Claude-2 exhibited imbalanced behavior, achieving moderate F1 scores in few-shot English prompting (F1 = 0.58) but struggling with recall (0.45), indicating an inability to generalize effectively.

Model and Prompt	P (Precision)	R (Recall)	F_1	Accuracy
<i>Few-Shot Prompting: GPT 3.5</i>				
English Prompt	0.61	0.95	0.74	0.66
Russian Prompt	0.65	0.92	0.76	0.70
<i>Zero-Shot Prompting: GPT 3.5</i>				
English Prompt	0.68	0.89	0.77	0.73
Russian Prompt	0.69	0.86	0.76	0.72
<i>Few-Shot Prompting: GPT 4.0</i>				
English Prompt	0.86	0.79	0.82	0.82
Russian Prompt	0.87	0.78	0.82	0.83
<i>Zero-Shot Prompting: GPT 4.0</i>				
English Prompt	0.78	0.92	0.84	0.82
Russian Prompt	0.77	0.89	0.83	0.81
<i>Few-Shot Prompting: LLaMA-3</i>				
English Prompt	0.52	0.98	0.68	0.53
Russian Prompt	0.52	0.99	0.68	0.52
<i>Zero-Shot Prompting: LLaMA-3</i>				
English Prompt	0.52	0.97	0.68	0.52
Russian Prompt	0.55	0.17	0.26	0.50
<i>Few-Shot Prompting: Mistral-7B</i>				
English Prompt	0.57	0.79	0.66	0.58
Russian Prompt	0.52	0.89	0.66	0.52
<i>Zero-Shot Prompting: Mistral-7B</i>				
English Prompt	0.46	0.08	0.13	0.48
Russian Prompt	0.52	0.83	0.64	0.51
<i>Few-Shot Prompting: Claude-2</i>				
English Prompt	0.82	0.45	0.58	0.60
Russian Prompt	0.61	0.55	0.58	0.50
<i>Zero-Shot Prompting: Claude-2</i>				
English Prompt	0.79	0.50	0.61	0.67
Russian Prompt	0.45	0.44	0.44	0.43

Figure 1: Performance metrics for GPT 3.5, GPT 4.0, LLaMA-3, Mistral-7B, and Claude-2 in Few-Shot and Zero-Shot Prompting scenarios using English and Russian prompts. Bold values indicate the highest performance for each column across all rows.

These results highlight that both models lack the robustness required for complex tasks in multilingual and context-sensitive domains. Notably, Claude-2 and LLaMA-3 were

particularly sensitive to prompt design, and their performance varied significantly based on the prompting language, with both models generally performing worse on Russian

prompts.

Prompting Strategy

The results reveal notable differences in model performance based on prompting strategy. **Few-shot prompting** often achieved higher recall across models, as seen with GPT-3.5 (Russian Few-Shot Recall = 0.92) and LLaMA-3 (Russian Few-Shot Recall = 0.99). This indicates that few-shot prompting can provide models with a stronger inductive bias toward identifying true positives in imbalanced datasets. However, the trade-off was reduced precision in many cases, as seen with LLaMA-3 (Russian Few-Shot Precision = 0.52).

In contrast, **zero-shot prompting** tended to produce more balanced results, yielding higher F1 scores and accuracy. GPT-4, for instance, achieved its best performance in the zero-shot English prompt setting (F1 = 0.84), reflecting its ability to effectively leverage minimal contextual information without overfitting to specific patterns in the prompt.

Implications for Multilingual NLP

The significant performance gap between Russian and English prompts across all models reinforces the *critical role of linguistic alignment in multilingual tasks*. Even models designed with multilingual capabilities, such as GPT-4, displayed notable improvements when prompted in Russian, which aligns directly with the dataset's language. This suggests that for multilingual or non-English tasks, careful prompt design in the target language is essential to achieving optimal results.

Furthermore, the stark underperformance of Claude-2 and LLaMA-3 highlights the limitations of smaller or less advanced models in generalizing effectively for non-English, context-sensitive tasks. This finding raises important questions about the scalability and robustness of smaller LLMs in real-world applications, particularly when deployed in low-resource or non-English domains.

Model Disagreement and Error Analysis This section examines where and why the models diverged in their predictions, focusing on key sources of disagreement and common error patterns.

Error analysis identified key challenges for the models:

- **False Positives:** Frequently occurred in cases involving events such as reports of military activity or protests, where human rights violations were implied but not explicitly mentioned.
- **False Negatives:** Often observed in posts containing indirect references or systemic violations, requiring deeper contextual understanding.
- **Disagreement on Ambiguity:** Posts with the highest disagreement involved ambiguous language or context, where models diverged in their interpretation of human rights implications.

Ablation Study

Two of the best-performing models, GPT-4.0 (proprietary) and LLaMA-3 (open-source), were evaluated on two distinct subsets of the original dataset: one where human annotators were in full agreement on the labels, and another where annotators disagreed, requiring adjudication. The goal of this analysis was to determine whether large

language models (LLMs) face similar challenges to those encountered by humans. Results are shown in Figure 2.

Do Models Face the Same Challenges as Humans?

1- Shared Struggles

- Disagreement cases likely involve ambiguous language, edge cases, or subtle contextual nuances that make labeling more challenging.
- Both LLaMA-3 and GPT-4.0 exhibit significant drops in Precision on these cases, suggesting an increased likelihood of false positives when ambiguity is present.

2 - GPT-4.0 Shows Greater Resilience - GPT-4.0 demonstrates a smaller decrease in F1 Score and Accuracy compared to LLaMA-3, indicating that it handles disagreement cases more effectively. - This suggests that GPT-4.0 may better capture nuanced patterns and context, reducing errors that LLaMA-3 struggles with in these challenging scenarios.

Key Findings

1. **Advanced Models Perform Better:** GPT-4.0 should be prioritized for tasks requiring high precision and recall. Its robust performance across multiple settings demonstrates the value of state-of-the-art LLMs.
2. **Alignment of Prompt Language Improves Performance:** Models consistently performed better with prompts in Russian, reflecting the importance of linguistic alignment when working with multilingual datasets.
3. **Few-Shot Prompting Boosts Open-Source Model Performance:** Open-source models showed substantial improvements with few-shot prompting compared to zero-shot prompting. In contrast, closed-source models like GPT-4.0 and GPT-3.5 displayed strong zero-shot capabilities, with marginal gains in few-shot settings.
4. **Additional Validation for Open-Source Models:** The moderate precision rates of open-source models, such as Mistral-7B Few-Shot (Russian) and LLaMA-3 Few-Shot (Russian), warrant an additional manual review step to mitigate false positives in sensitive applications.

Future Directions Future work should focus on enhancing the multilingual capabilities of smaller LLMs like LLaMA and Mistral through *fine-tuning* strategies or domain-specific pretraining. Additionally, investigating advanced *prompt engineering* techniques and dataset augmentation could help mitigate the performance gaps of underperforming models, especially in context-sensitive and low-resource scenarios. Expanding experiments to include a *broader range of languages* will also provide valuable insights into the adaptability and generalization of these models across diverse linguistic contexts.

6 Limitations

Despite promising findings, this study has several limitations that should be addressed in future work.

Language Coverage: The dataset focuses on Russian and Ukrainian posts, which limits generalizability to other lan-

Set	Model and Prompt	P (Precision)	R (Recall)	F_1	Accuracy
4*Agreement Set	Few-Shot Prompting: LLaMA-3 (English)	0.54	0.99	0.70	0.54
	Few-Shot Prompting: LLaMA-3 (Russian)	0.53	1.00	0.69	0.53
	Zero-Shot Prompting: LLaMA-3 (English)	0.53	0.98	0.69	0.53
	Zero-Shot Prompting: LLaMA-3 (Russian)	0.55	0.17	0.25	0.48
	Few-Shot Prompting: GPT-4.0 (English)	0.90	0.83	0.86	0.86
	Few-Shot Prompting: GPT-4.0 (Russian)	0.92	0.81	0.86	0.86
	Zero-Shot Prompting: GPT-4.0 (English)	0.82	0.92	0.86	0.85
	Zero-Shot Prompting: GPT-4.0 (Russian)	0.82	0.90	0.86	0.84
4*Disagreement Set	Few-Shot Prompting: LLaMA-3 (English)	0.45	0.94	0.60	0.45
	Few-Shot Prompting: LLaMA-3 (Russian)	0.45	1.00	0.62	0.45
	Zero-Shot Prompting: LLaMA-3 (English)	0.44	0.94	0.60	0.45
	Zero-Shot Prompting: LLaMA-3 (Russian)	0.57	0.20	0.29	0.58
	Few-Shot Prompting: GPT-4.0 (English)	0.65	0.60	0.62	0.68
	Few-Shot Prompting: GPT-4.0 (Russian)	0.65	0.59	0.62	0.67
	Zero-Shot Prompting: GPT-4.0 (English)	0.55	0.80	0.65	0.62
	Zero-Shot Prompting: GPT-4.0 (Russian)	0.82	0.90	0.86	0.84

Figure 2: Performance metrics for LLaMA-3 and GPT-4.0 on the agreement and disagreement sets in Few-Shot and Zero-Shot Prompting scenarios using English and Russian prompts. Bold values indicate the highest performance for each column across all rows.

guages or dialects. Although Russian prompts generally outperformed English prompts, additional analysis is needed to understand model performance on datasets with multiple or less-resourced languages.

Prompt Design Limitations: The study used standardized zero-shot and few-shot prompts, but the potential of prompt engineering remains underexplored. Optimizing prompt structure, content, or complexity could further enhance performance, particularly for underperforming models.

Sample Size: The study utilized a random sample of 1000 posts to evaluate model performance. While a larger dataset would enable more robust and generalizable conclusions, the computational cost associated with employing models such as GPTs or Claude makes the use of an expanded dataset significantly more expensive.

Computational Costs: While GPT-4 and GPT-3.5 achieved strong results, their computational demands and associated costs may limit their scalability in real-world applications. Smaller models like LLaMA-3.2-1B, despite their lower performance, offer cost advantages that could be valuable in resource-constrained environments.

7 Conclusion

This study systematically evaluates the performance of both paid and open-source language models for identifying references to human rights violations in a multilingual dataset. By analyzing the effects of prompting style, language alignment, and model architecture, we provide actionable insights for selecting and optimizing models under varying resource constraints.

Our findings reveal that paid models, particularly GPT-4.0, excel in both precision and recall, demonstrating robust generalization across languages and task configura-

tions. Open-source models, such as Llama3 and Mistral7B, perform competitively with carefully designed prompts, especially in few-shot and language-aligned settings. However, their sensitivity to context and higher error rates underscore the need for task-specific fine-tuning.

The disparity in performance between Russian and English prompts highlights the importance of language alignment in multilingual tasks. Furthermore, error analysis shows that ambiguity and indirect references pose significant challenges, leading to model disagreements and underscoring the complexity of the human rights domain.

These insights offer practical guidance for practitioners: leveraging few-shot prompting for open-source models in low-budget settings and prioritizing GPT-4.0 for high-stakes applications where accuracy is paramount. Our study paves the way for future work to refine LLMs for nuanced, high-impact tasks such as monitoring and analyzing human rights violations globally.

References

- Aldeen, M.; Luo, J.; Lian, A.; Zheng, V.; Hong, A.; Yetukuri, P.; and Cheng, L. 2023. ChatGPT vs. Human Annotators: A Comprehensive Analysis of ChatGPT for Text Annotation. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, 602–609.
- Anthropic. 2023. Claude: Safe and Aligned Language Models. Available at <https://www.anthropic.com/claude>.
- Artstein, R.; and Poesio, M. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4): 555–596.
- Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of*

the 2021 ACM conference on fairness, accountability, and transparency, 610–623.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brown, T.; and Others. 2020. Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1): 37–46.

Fisher, J.; Feng, S.; Aron, R.; Richardson, T.; Choi, Y.; Fisher, D. W.; Pan, J.; Tsvetkov, Y.; and Reinecke, K. 2024. Biased ai can influence political decision-making. *arXiv preprint arXiv:2410.06415*.

Goel, A.; Gueta, A.; Gilon, O.; Liu, C.; Erell, S.; Nguyen, L.; Hao, X.; Jaber, B.; Reddy, S.; Kartha, R.; Steiner, J.; Laish, I.; and Feder, A. 2023. LLMs Accelerate Annotation for Medical Information Extraction. *ArXiv*, abs/2312.02296.

Hasanbeig, H.; Sharma, H.; Betthausen, L.; Frujeri, F.; and Momennejad, I. 2023. ALLURE: Auditing and Improving LLM-based Evaluation of Text using Iterative In-Context-Learning. *ArXiv*, abs/2309.13701.

Japkowicz, N.; and Stephen, S. 2002. The Class Imbalance Problem: A Systematic Study. In *Proceedings of the 1st International Conference on Machine Learning (ICML)*, 253–270.

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. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Li, M.; Shi, T.; Ziems, C.; Kan, M.-Y.; Chen, N. F.; Liu, Z.; and Yang, D. 2023. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. *arXiv preprint arXiv:2310.15638*.

Lu, Y.; Yao, B.; Zhang, S.; Wang, Y.; Zhang, P.; Lu, T.; Li, T.; and Wang, D. 2023. Human Still Wins over LLM: An Empirical Study of Active Learning on Domain-Specific Annotation Tasks. *ArXiv*, abs/2311.09825.

Mistral AI. 2023. Mistral 7B: A New Generation of Open-Source Language Models. Available at <https://mistral.ai>.

Nasution, A. H.; and Onan, A. 2024. ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-resource Language NLP Tasks. *IEEE Access*.

Nemkova, P.; Ubani, S.; Polat, S. O.; Kim, N.; and Nielsen, R. D. 2023. Detecting Human Rights Violations on Social Media during Russia-Ukraine War. *ArXiv*, abs/2306.05370.

OpenAI. 2023. GPT-4 Technical Report. Available at <https://openai.com/research/gpt-4>.

Pangakis, N.; Wolken, S.; and Fasching, N. 2023. Automated Annotation with Generative AI Requires Validation. *ArXiv*, abs/2306.00176.

Pineau, J.; LeBlanc, A.; Raffel, C.; Maddison, C.; Sinha, K.; Roelofs, R.; Cheung, V.; Dodge, J.; Kuznetsova, A.; and Program, N. R. 2021. Improving Reproducibility in Machine Learning Research: A Report from the NeurIPS 2019 Reproducibility Program. *Communications of the ACM*, 64(4): 76–84.

Plaza-del Arco, F. M.; Nozza, D.; and Hovy, D. 2023. Leveraging label variation in large language models for zero-shot text classification. *arXiv preprint arXiv:2307.12973*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Shaib, C.; Li, M.; Joseph, S. A.; Marshall, I.; Li, J. J.; and Wallace, B. 2023. Summarizing, Simplifying, and Synthesizing Medical Evidence using GPT-3 (with Varying Success). 1387–1407.

Shin, T.; Razeghi, Y.; Logan IV, R. L.; Wallace, E.; and Singh, S. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4222–4235. Online: Association for Computational Linguistics.

Takagi, S.; Watari, T.; Erabi, A.; and Sakaguchi, K. 2023. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. *JMIR Medical Education*, 9.

Törnberg, P. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Roziere, B.; Wightman, D.; Lachaux, M.-A.; Kolesnikov, A.; Labatut, M.; Mazué, F.; Usunier, N.; Synnaeve, G.; Verbeek, J.; and Jégou, H. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*. Available at <https://arxiv.org/abs/2302.13971>.

Törnberg, P. 2023. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. *ArXiv*, abs/2304.06588.

Wong, E. 2024. Comparative Analysis of Open Source and Proprietary Large Language Models: Performance and Accessibility. *Advances in Computer Sciences*, 7(1): 1–7.

Zimmer, M. 2010. “But the data is already public”: on the ethics of research in Facebook. *Ethics and Information Technology*, 12: 313–325.

Šavelka, J. 2023. Unlocking Practical Applications in Legal Domain: Evaluation of GPT for Zero-Shot Semantic Annotation of Legal Texts. *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*.