Prompt-driven Detection of Offensive Urdu Language using Large Language Models

Iffat Maab[1], Usman Haider[2], Junichi Yamagishi[1]

[1]Yamagishi Lab, National Institute of Informatics (NII), Tokyo

[2]University of Galway, Ireland

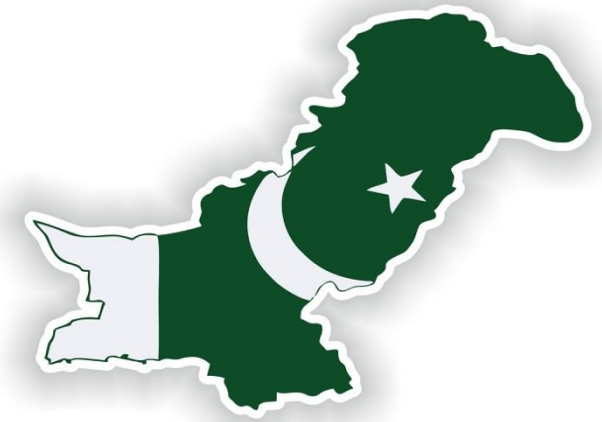
Motivation

- **Offensive language detection** is crucial for safer online spaces (Ullah et al., 2023).
- Social media platforms can **amplify hate and acts of violence.**
- **Traditional NLP approaches** have typically dominated the field of hate speech detection, which requires extensive tuning and careful model design.
- **Resource gaps are severe** for addressing offensive language, particularly when it is transcribed in non-native scripts, such as Roman Urdu and Urdu.
- Prompt-based LLMs may offer a more flexible, **rapidly adaptable** solution for offensive language detection.

Can LLMs offer a scalable, explainable alternative?

Faizad Ullah, Ubaid Azam, Ali Faheem, Faisal Kamiran, and Asim Karim. 2023. Comparing prompt-based and standard fine-tuning for urdu text classification. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 6747–6754

Urdu Writing Systems and Code-Switching



- Urdu, **spoken by more than 70 million people**, has a complex
- morphological structure.
- **Urdu is Pakistan's national language** and is central to cultural and social communication.
- Urdu uses an **abjad script** derived from **Persian**, with historical roots in the **Arabic writing tradition**.
- In daily online communication, people widely use **Roman Urdu (Latin script)** and frequently **code-switch between Urdu and English**, even within the same sentence.

Roman Urdu:	**Tum bilkul jahil ho**, itni simple baat bhi samajh nahi aati.
Urdu:	تم بالکل جاہل ہو، اتنی سادہ بات بھی سمجھ نہیں آتی
English:	*You are completely ignorant; you can't even understand such a simple thing.*

Problems Statement

- **Low-resource setting:** Urdu has limited labeled data and quite a few prompt-based ICL studies compared to high-resource languages.
- **Multi-script reality and code-mixing is common:** Urdu appears in both **Urdu script** and **Roman Urdu**, requiring models to handle both reliably.
- **Urdu Script →** Formal, standardized, used in media and official contexts
- **Roman Urdu →** Informal, non-standardized, common in social media and messaging
- **Roman Urdu is highly noisy:** it lacks standardized spelling, uses inconsistent abbreviations, and employs informal grammar, making modeling difficult.
- **Supervised methods are costly:** Fine-tuning strong classifiers requires large annotated datasets and careful training design.
- **Limited generalization:** Supervised systems can become **dataset-specific** and may reflect annotation or domain biases.

Related Work

- Research on **in-context learning (ICL) and prompt engineering for Urdu text classification** is quite limited.
- **Rizwan et al. (2020):** introduced **RUHSOLD** and use supervised baselines for Roman Urdu hate/offensive detection.
- **Ullah et al. (2023):** studied fine-tuning vs prompt-based fine-tuning using BERT-family models.
- **Arif et al. (2024):** evaluated generalist LLMs (GPT-4, Llama-8B) only on Urdu-script abuse detection.

- Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. 2020. Hate-speech and offensive language detection in roman urdu. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pages 2512–2522.
- Faizad Ullah, Ubaid Azam, Ali Faheem, Faisal Kamiran, and Asim Karim. 2023. Comparing prompt-based and standard fine-tuning for urdu text classification. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 6747–6754.
- Samee Arif, Abdul Hameed Azeemi, Agha Ali Raza,and Awais Athar. 2024. Generalists vs. specialists: Evaluating large language models for urdu. arXiv preprint arXiv:2407.04459.

Our Contribution

- Focuses on **inference-time prompting + ICL** with state-of-the-art LLMs
- Prior studies prompt only in **English**
  - **Systematic Prompt Language Comparison:** We test prompts in **English, Roman Urdu, and Urdu**
  - **Multi-Script Robustness Analysis:** Evaluates **multiple scripts and datasets** (Roman Urdu + Urdu script)
- Direct comparison with strong supervised baselines.
- The terms abusive and "offensive" are resolved and used interchangeably.
- **Need for practical evaluation:** We still lack clear answers about:
  - Which LLMs and fine-tuned approaches work best for Urdu? Does k-shot help? Which script/prompt language works best?

Datasets

We evaluate LLM-based offensive language detection on **three Roman Urdu/Urdu datasets**:

- **RUHSOLD** *(Roman Urdu Tweets; Rizwan et al., 2020)*
  - **Coarse-grained:** Abusive/Offensive vs Normal
  - **Fine-grained:** Abusive, Religious Hate, Sexual Hate, Profane/Untargeted
  - ~10K samples per setting
- **RUTUBE** *(Roman Urdu YouTube Comments; Akhter et al., 2020)*
  - 10K sampled comments (~147K)
  - Binary: Abusive vs Non-abusive
- **USADC** *(Urdu Script multiple sources; Arif et al., 2024)*
  - Urdu script (not Roman Urdu)
  - Binary: Abusive vs Non-abusive

Dataset Statistics and Prompt Variants

PER-CLASS	RUHSOLD		RUTUBE	USADC
Labels	**Coarse**	**Fine**	**Binary**	**Binary**
Normal	5342	4917	5000	2816
Abusive	4650	2208	5000	2854
Sexual Hate	-	770	-	-
Religious Hate	-	719	-	-
Profane	-	582	-	-
Prompts Tested				
English	✓	✓	✓	✓
Roman Urdu	✓	✓	-	-
Urdu	✓	✓	-	-

Table: Summary of the data available for coarse- and fine-grained tasks using RUHSOLD, and binary tasks for RUTUBE and USADC, and prompt-based methodologies across each dataset.

Prompt Formulation (English, Roman Urdu, Urdu)

	Type	Prompt Template	Labels
Coarse-Grained	English	System: <As a responsible AI tool, you are an abusive language detector.> User: <Your job is to classify the following "*sentence*" as either c_a, or c_b. Only output one of these two labels.> Sentence: <sentence>	c_a: Abusive c_b: Not abusive
	Roman Urdu	System: <Ek zimmedar AI tool ke tor par, aap ek gaali detektor hain> User: <Barah-e-karam darj zael jumlay: "*sentence*" ko c_a, ya c_b, ke tor par darja bandi karein. Inn darja bandiyon kay ilawa koi izafi malomat na den.> Jumla: <sentence>	c_a: Khair ikhlaqi c_b: Ikhlaqi
	Urdu	System: ایک ذمہ دار اے آئی ٹول کے طور پر، آپ ایک گالی ڈیٹیکٹر ہیں۔ User: براہ کرم درج ذیل جملے "*sentence*" کو c_a, ya c_b کے طور پر درجہ بندی کریں۔ جملہ: ان درجہ بندیوں کے علاوہ کوئی اضافی معلومات نہ دیں۔ <sentence>	c_a: غیر اخلاقی c_b: اخلاقی
Fine-Grained	English	System: <As a responsible AI tool, you are an abusive language detector.> User: <Classify the following "*sentence*" as c_a, c_b, c_c, c_d, or c_e. Only output one of these five labels.> Sentence: <sentence>	c_a: Abusive c_b: Not abusive c_c: Sexual hate c_d: Religious hate c_e: Profane
	Roman Urdu	System: <Ek zimmedar AI tool ke tor par, aap ek gaali detektor hain> User: <Aap ek gaali detektor hain. Barah-e-karam darj zael jumlay: "*sentence*" ko c_a, c_b, c_c, c_d, ya c_e, ke tor par darja bandi karein. Inn darja bandiyon kay ilawa koi izafi malomat na den.> Jumla: <sentence>	c_a: Khair ikhlaqi c_b: Ikhlaqi c_c: Jinsi istehsal c_d: Mazhabi nafrat c_e: Be-adab
	Urdu	System: ایک ذمہ دار اے آئی ٹول کے طور پر، آپ ایک گالی ڈیٹیکٹر ہیں۔ User: براہِ کرم درج ذیل جملہ کو درجِ ذیل لیبلز میں سے کسی ایک کے طور "*sentence*" ان درجہ بندیوں کے علاوہ کوئی اضافی c_a, c_b, c_c, c_d, ya c_e پر درجہ بندی کریں جملہ: معلومات نہ دیں۔ <sentence>	c_a: خیر اخلاقی c_b: اخلاقی c_c: جنسی استحصال c_d: مذہبی نفرت c_e: بے ادب

Test studies to make sure the category labels are **natural, understandable, and culturally appropriate** in each language.

Few-shot Examples for RUSOLD, RUTUBE, and USADC

TASK	Prompt Template	Labels
COARSE-GRAINED RUHSOLD	Sentence: "lanati slected government murdabad." Label: l_a Sentence: "bc hn sary bas khan per bhokny ky paisy mil rhy sab ko." Label: l_a Sentence: "lagta hsi man chaha job nae de rahy bhai ko." Label: l_b Sentence: "acha ha g bohat acha ha." Label: l_b Sentence: "s", Label:	l_a: "Abusive" l_a: "Abusive" l_b: "Not Abusive" l_b: "Not Abusive"
FINE-GRAINED RUHSOLD	Sentence: "or haan tm pe or uss ghaleez soowr pashteen per bahrpoooor lanat." Label: l_a Sentence: "kasy btaaa du itna mushkil hai aesy uska direct nam ly kar btana." Label: l_b Sentence: "rt : tum jese khusray maa bhen ki baatein hi krty hain oqat dekh r baatein dekho gandi naali k keeray ki." Label: l_c Sentence: "is molvi ne yahoodi dress coat pehna hua hay is k iman ka b mujhy shak hay." Label: l_d Sentence: "kyaaaa musibbaattt ha yr sara mood khraab kr diya ha bc." Label: l_e Sentence: "s", Label:	l_a:"Abusive" l_b: "Not Abusive" l_c: "Sexual Hate" l_d: "Religious Hate" l_e: "Profane"
RUTUBE	Sentence: "Inhi jahilon ki Waja say Nawaz shreef select hota hai" Label: l_a Sentence: "bharwy phely to sham ki level pe aa phr akar bt kariyo bc randi ki nasal" Label: l_a Sentence: "kia bat hay larke you are our jani." Label: l_b Sentence: "Karachi Me Aashura K Mauqay Per Mukammal Aman Raha." Label: l_b Sentence: "s", Label:	l_a: "Abusive" l_a: "Abusive" l_b: "Not Abusive" l_b: "Not Abusive"
USADC	Sentence: "بکواس بند کر حرامی کتے" Label: l_a Sentence: "لعنت اس بیغیرت کے منہ پے، کنجروں کے پاس اختیارات ہونے کے باوجود یہ حال ہے، اس بیغیرت کو عوام کو حوصلہ دینا چاہیے" Label: l_a Sentence: "میرا اطلاعات کا وزیر کدہر ہے" Label: l_b Sentence: "پیارے بچو آج کے دن برصغیر کا مسلمان آزاد نہیں صرف تقسیم ہؤا تھا" Label: l_b Sentence: "s", Label:	l_a: "Abusive" l_b: "Abusive" l_a: "Not Abusive" l_a: "Not Abusive"

Results using RUHSOLD (Coarse-Grained Task)

Model	COARSE-GRAINED English Acc.	English F1	Roman Ur. Acc.	Roman Ur. F1	Urdu Acc.	Urdu F1
Llama-3.2-1B	57.76	54.47	51.87	57.50	53.43	45.43
+ K-SHOTS	**60.20**	**60.20**	52.90	58.80	52.19	48.63
Llama-3.2-3B	63.60	64.08	55.93	50.87	54.67	48.60
+ K-SHOTS	**66.47**	**67.12**	56.25	53.53	56.77	49.55
Llama-3-8B	65.22	65.30	63.88	63.88	55.79	44.69
+ K-SHOTS	**70.88**	**69.82**	66.66	65.02	59.12	49.79
Llama-3-70B	74.36	74.36	70.26	70.76	70.83	68.88
+ K-SHOTS	**78.41**	**78.41**	73.77	73.11	73.68	71.07
Qwen-2-7B	73.21	73.21	69.17	67.40	-	-
+ K-SHOTS	**76.90**	**75.77**	70.03	69.80	-	-
Qwen-2-72B	82.07	81.11	72.32	75.91	-	-
+ K-SHOTS	**84.53**	**85.75**	74.55	76.40	-	-
Lughaat-1-8B	57.76	49.96	52.35	50.85	49.84	51.18
+ K-SHOTS	**61.58**	**58.75**	54.09	45.49	49.28	51.40
GPT-4	91.38	92.17	84.46	84.46	88.78	87.78
+ K-SHOTS	**92.91**	**92.58**	86.05	85.37	90.14	89.05
Supervised						
BERT-M Ullah et al. (2023)	54.20	-	-	-	-	-
DISTIL-BERT Ullah et al. (2023)	52.80	-	-	-	-	-
XLM-R Ullah et al. (2023)	57.20	-	-	-	-	-
XLM-Roberta+CNN Rizwan et al. (2020)	88.00	88.00	87.21*	87.14*	-	-
RomUrEm+CNN Rizwan et al. (2020)	89.00	89.00	61.90*	60.92*	-	-
BERT+CNN Rizwan et al. (2020)	90.00	90.00	**89.21***	**89.17***	-	-

English prompts are strongest overall, but Roman Urdu prompts are surprisingly competitive, and Urdu-script prompts tend to be weakest.

Table: Summary of RUHSOLD results for coarse-grained tasks under zero- and few-shot settings across English, Roman Urdu, and Urdu prompts. Results marked with (*) indicate fine-tuned models that we reimplemented for consistent evaluation.

Results using RUHSOLD (Fine-Grained Task)

Model	FINE-GRAINED English Acc.	English F1	Roman Ur. Acc.	Roman Ur. F1	Urdu Acc.	Urdu F1
Llama-3-8B	59.11	53.33	21.27	20.95	15.11	15.38
+ K-SHOTS	**58.15**	**53.15**	48.55	47.08	31.77	31.70
Llama-3-70B	67.34	62.04	62.79	60.48	57.59	55.36
+ K-SHOTS	**72.67**	**69.25**	69.21	68.86	68.34	68.33
Qwen-2-7B	52.16	49.90	50.34	51.60	-	-
+ K-SHOTS	**53.11**	**49.63**	58.90	59.15	-	-
Qwen-2-72B	73.96	68.31	59.71	59.39	-	-
+ K-SHOTS	**75.77**	**70.30**	62.20	60.89	-	-
Lughaat-1-8B	44.45	44.50	29.77	29.80	**49.59**	**49.81**
+ K-SHOTS	29.59	32.81	22.23	22.81	42.84	43.47
GPT-4	75.50	75.02	68.32	68.28	63.27	62.98
+ K-SHOTS	**78.84**	**77.56**	70.52	69.90	68.24	65.09
Supervised						
XLM-RoBERTa+CNN Rizwan et al. (2020)	81.00	72.00	**79.57***	69.39*	-	-
RomUrEm+CNN Rizwan et al. (2020)	75.00	64.00	55.92*	51.27*	-	-
BERT+CNN Rizwan et al. (2020)	82.00	75.00	78.47*	**75.83***	-	-

Table: Summary of RUHSOLD results for fine-grained tasks under zero- and few-shot settings across English, Roman Urdu, and Urdu prompts. Results marked with (*) indicate fine-tuned models that we reimplemented for consistent evaluation.

RUTUBE versus USADC (Binary Classification)

Model	RUTUBE Acc.	RUTUBE F1	USADC Acc.	USADC F1
Llama-3-8B	66.99	65.95	50.83	47.80
+ K-SHOTS	66.48	68.12	69.08	68.70
Llama-3-70B	72.70	72.07	73.83	72.83
+ K-SHOTS	76.00	74.55	79.90	78.51
Qwen-2-7B	69.01	66.20	67.76	66.97
+ K-SHOTS	69.35	68.60	72.88	72.45
Qwen-2-72B	79.99	79.20	79.34	78.89
+ K-SHOTS	83.89	82.46	81.24	81.24
Lughaat-1-8B	52.59	11.56	52.34	47.82
+ K-SHOTS	78.34	75.36	56.79	55.20
GPT-4	86.70	85.95	86.41	86.34
+ K-SHOTS	**87.33**	**86.15**	**88.85**	**88.78**
Supervised				
SimpleLogistic Reg. Akhter et al. (2020)	85.50*	85.61*	74.56*	77.86*
LogitBoost Reg. Akhter et al. (2020)	85.40*	85.35*	75.43*	77.77*
XLM-Roberta+CNN Rizwan et al. (2020)	90.15*	90.11*	**84.21***	**84.18***
RomUrEm+CNN Rizwan et al. (2020)	68.60*	68.46*	70.17*	69.16*
BERT+CNN Rizwan et al. (2020)	**92.55***	**92.53***	78.94*	77.97*
No Fine-tuning				
Llama-3-8b Arif et al. (2024) (0-shot)	-	-	-	44.73
Llama-3-8b Arif et al. (2024) (6-shot)	-	-	-	71.64
GPT-4 Arif et al. (2024) (0-shot)	-	-	-	86.27
GPT-4 Arif et al. (2024) (6-shot)	-	-	-	88.71

Table: Comparison of RUTUBE and USADC. Results marked with (*) indicate fine-tuned models that we reimplemented for consistent evaluation.

Cross-Lingual Prompt Responsiveness of LLMs

	Responsiveness to prompts provided in		Given English prompt can process	
Model	**Roman Urdu**	**Urdu**	**Roman Urdu**	**Urdu**
Llama-3	✓	✓	✓	✓
Qwen-2	✓	X	✓	✓
FLAN-T5	Partial	X	✓	Partial
Falcon	Partial	X	✓	✓
GPT-4	✓	✓	✓	✓

Table: Comparative analysis of various LLMs responsiveness to prompts provided in Roman Urdu and Urdu language, while also detailing each model's ability to interpret and process prompts provided in English across these languages

Influence of `k'-shots

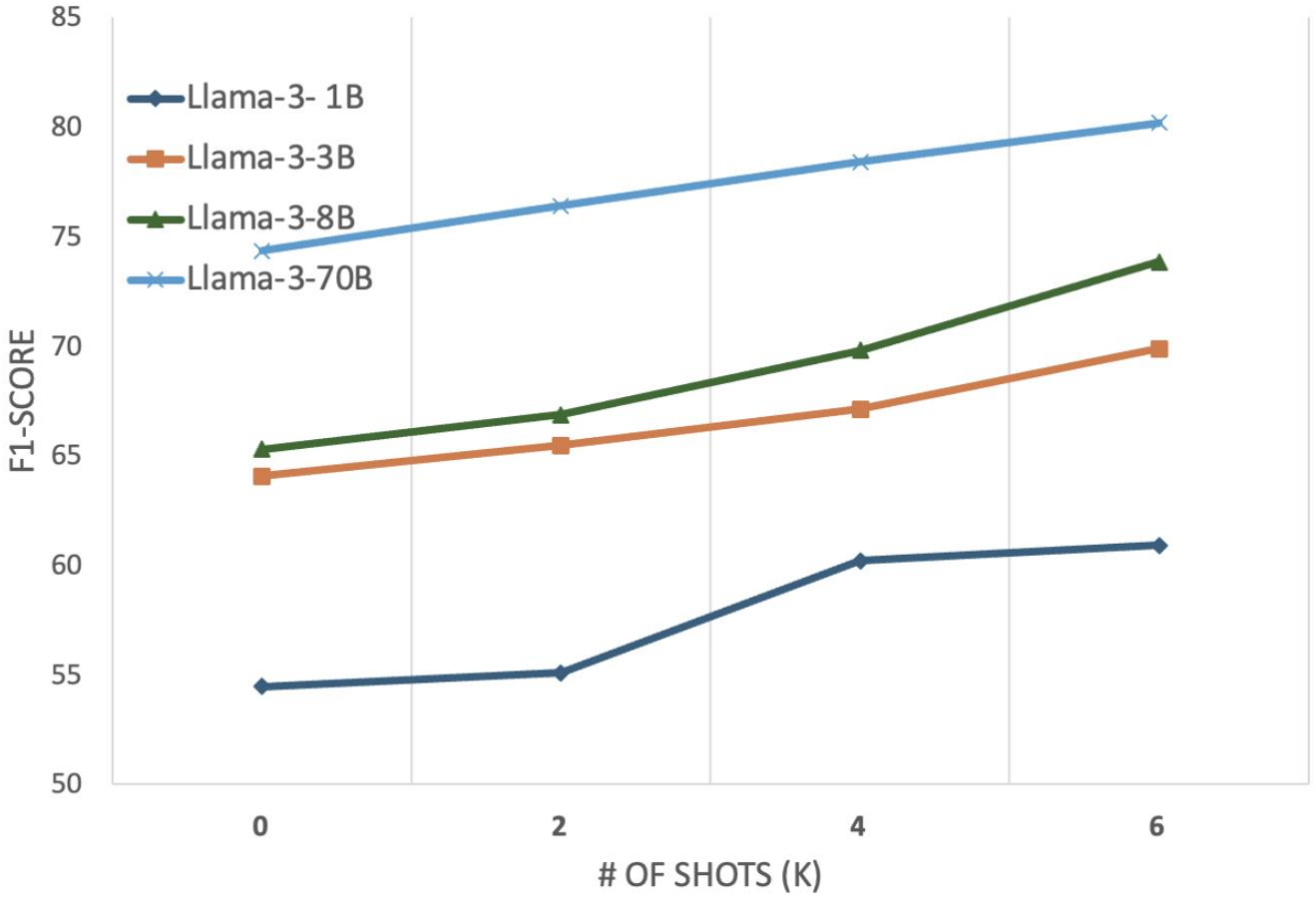


Figure: Influence of `k' in few-shot setting with English prompts for zero-, 2-, 4-, and 6-shot experiments on Llama-3.2 and Llama-3 models of various sizes.

Conclusion

- This work highlights the importance of **script choice, prompt language, and low-resource realism** in evaluating LLMs.
- Our experiments reveal both the **potential and limitations** of LLMs for Urdu hate/offensive content classification **without training**.
- **English prompts** consistently give the **best performance** across models.
- **Urdu-script prompts** tend to be less reliable, producing weaker prompt quality and lower accuracy.

Key Takeaways

- **Roman Urdu prompts** often improve results and provide a **strong middle ground** for prompt design.
- **Prompt-based LLM approaches** are effective for **coarse-grained (binary)** offensive language detection, while **supervised fine-tuned models** remain reliable for **fine-grained classification tasks.**

Future Work

- Extend evaluation to broader abuse categories (e.g., **cultural, racial, age-related** harmful content).
- Improve robustness for **Urdu-script prompting** and **code-mixed settings**.

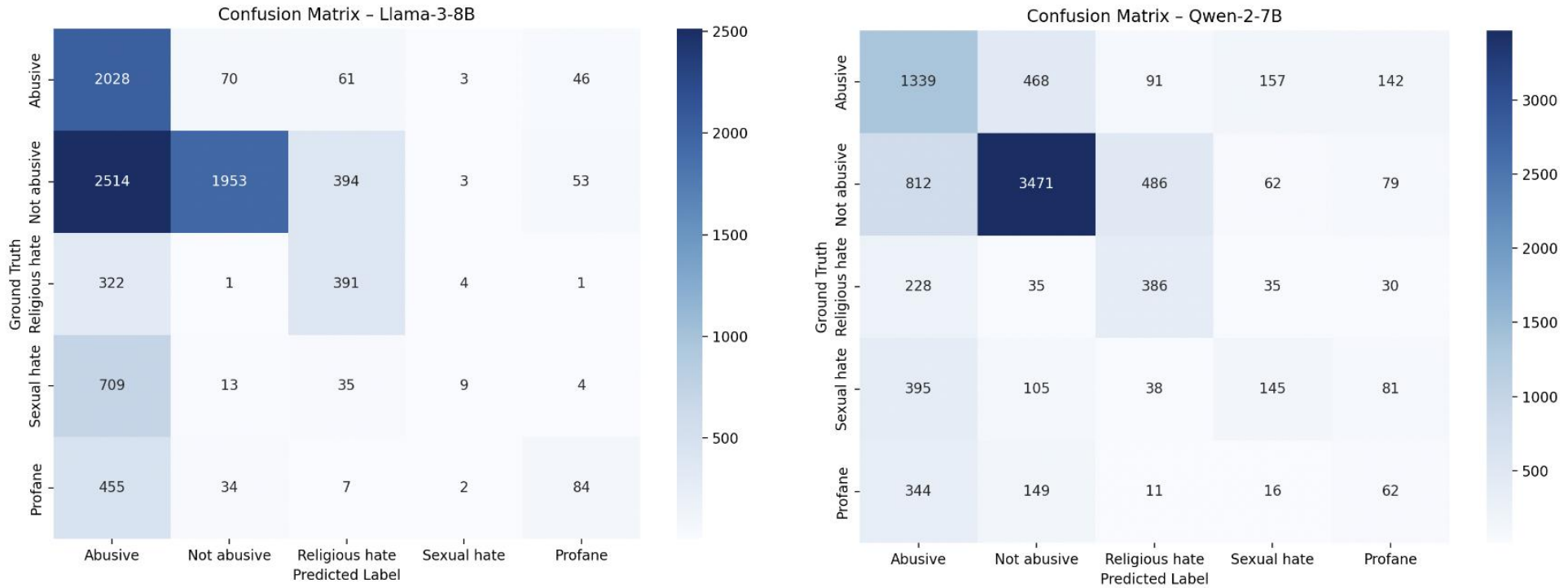
Acknowledgements

The authors wish to express gratitude to the funding organization, as this work is supported by the JST CREST Grant (JPMJCR20D3), Japan, and the TSUBAME 4.0 supercomputer at the Institute of Science, Tokyo, whose computational resources are gratefully acknowledged.

Thank you very much!

Appendix

Error Analysis (Fine-grained RUHSOLD)



(a) Llama-3-8B (b) Qwen-2-7B

Figure 2: Comparison of model performance across coarse-grained RUHSOLD using Roman Urdu prompt with few-shot settings for (a) Llama-3-8B and (b) Qwen-2-7B