Can Prompting LLMs Unlock Hate Speech Detection across Languages? A Zero-shot and Few-shot Study

Faeze Ghorbanpour[1,2,3] **Daryna Dementieva**[1] **Alexander Fraser**[1,3]

[1]School of Computation, Information and Technology, TU Munich
[2]Center for Information and Language Processing, LMU Munich
[3]Munich Center for Machine Learning (MCML)

faeze.ghorbanpour@tum.de, daryna.dementieva@tum.de

Abstract

Despite growing interest in automated hate speech detection, most existing approaches overlook the linguistic diversity of online content. Multilingual instruction-tuned large language models such as LLaMA, Aya, Qwen, and BloomZ offer promising capabilities across languages, but their effectiveness in identifying hate speech through zero-shot and few-shot prompting remains underexplored. This work evaluates LLM prompting-based detection across eight non-English languages, utilizing several prompting techniques and comparing them to fine-tuned encoder models. We show that while zero-shot and few-shot prompting lag behind fine-tuned encoder models on most of the real-world evaluation sets, they achieve better generalization on functional tests for hate speech detection. Our study also reveals that prompt design plays a critical role, with each language often requiring customized prompting techniques to maximize performance. [1]

1 Introduction

Hate speech is a worldwide issue that undermines the safety of social media platforms, no matter the language (Thomas et al., 2021). It can violate platform rules, damage user trust, influence opinions, and reinforce harmful biases against individuals or groups targeted (MacAvaney et al., 2019; Vedeler et al., 2019; Stockmann et al., 2023). However, most recent advancements in hate speech detection have focused primarily on English, as the majority of datasets and language models are centered on English content. This has led to limited—but not negligible—attention to other languages (Huang et al., 2023; Peng et al., 2023). Since users on social media write and engage with content in many languages—not just English—it is crucial to find tools that can detect hate speech across various languages.

Instruct-tuned Large language models (LLMs) have demonstrated exceptional performance across a wide range of text-related tasks (Skibicki, 2025; Zhang et al., 2024). Many of these models possess multilingual capabilities, enabling them to process and understand text in various languages (Pedrazzini, 2025; Shaham et al., 2024). This makes them suitable for tasks like hate speech detection, even without additional fine-tuning. Although fine-tuning is possible, it requires computational and resource costs, which leads many users to rely on prompt-based use instead (Min et al., 2022; Zhao et al., 2023). While their effectiveness in detecting hate speech in English has been studied extensively (Roy et al., 2023; Guo et al., 2023; Zhang et al., 2025), their performance on non-English datasets remains underexplored.

To evaluate the capabilities of multilingual instruction-tuned LLMs in detecting hate speech in various languages, we conduct a study using several prompting techniques, including zero-shot prompting (e.g., vanilla, chain-of-thought, role-play), few-shot prompting, and combinations of these prompts. We evaluate performance across eight non-English hate speech detection tasks, covering Spanish, Portuguese, German, French, Italian, Turkish, Hindi, and Arabic, using real-world[2] and hate speech functional test sets. This study seeks to address the following research questions: (1) How well do LLMs perform on hate speech detection across various non-English languages? (2) Does few-shot prompting improve performance compared to zero-shot prompting? (3) How does LLM performance compare to that of traditional fine-tuned models?

Our findings highlight the importance of prompt design in multilingual hate speech detection. While performance varies by the prompting strategy, ex-

[1]The code and prompts are publicly available at: https://github.com/FaezeGhorbanpour/MultilingualHateSpeechPrompting

[2]By real-world test sets, we meant datasets collected from actual conversations, which better reflect real-world scenarios.

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 413–425

perimenting with different techniques leads to reasonably strong results. In most languages, few-shot prompting combined with other techniques outperforms zero-shot prompting, suggesting that providing a few task-specific examples is beneficial.

Compared to fine-tuned encoder models, prompting LLMs shows lower performance on real-world test sets. However, in functional test cases, prompting often performs better. Further analysis of languages where prompting underperforms on real-world data suggests that prompting can still be a practical option when only limited training data is available. Nonetheless, with access to larger training sets, fine-tuning encoder models remains the more effective approach. Overall, instruction-tuned LLMs demonstrate stronger generalization in controlled functional benchmarks, without the need for additional training.

2 Related Work

The ability of instruction-tuned LLMs to perform a wide range of NLP tasks without the need for fine-tuning or training data has drawn growing interest, particularly in applications like hate speech detection. Recent studies have explored LLM-based hate speech detection, primarily in English. Zhu et al. (2025) reports low agreement between LLM predictions and human annotations, while Li et al. (2024) finds that LLMs are more effective at identifying non-hateful content. Huang et al. (2023) examines the use of LLMs for generating explanations of implicit hate, and Roy et al. (2023) shows that including target-specific information in prompts improves performance.

Another study examines how in-context learning, combined with few-shot examples and task descriptions, boosts the performance of hate speech detection by LLMs (Han and Tang, 2022). Guo et al. (2023) investigates using LLMs for real-world hate speech detection using four diverse prompting strategies and finds that few-shot and chain-of-thought prompts help. While these works have explored prompting techniques, they primarily assess the capabilities of LLMs for hate speech detection in English and do not examine a broad range of prompting strategies across languages.

There have been efforts to investigate the capabilities of LLMs for non-English hate speech. Guo et al. (2023) and Faria et al. (2024) tested prompt strategies only in Chinese and Bangla, respectively. Ahmad et al. (2025) used an LLM for hate speech detection in Urdu, outperforming BERT in detecting both explicit and implicit hate. Moving beyond isolated languages, Masud et al. (2024) assesses LLMs' sensitivity to geographical priming and persona attributes in five languages, showing that geographical cues can improve regional alignment in hate speech detection. Similarly, Zahid et al. (2025) uses geographical contextualization into prompts for five languages. These motivate our use of culture-aware prompts; However, these studies do not explore a wide range of prompting strategies, such as few-shot, chain-of-thought, etc.

He et al. (2024) introduced a multilingual benchmark for offensive language detection in eight languages, focusing on offensive language and model alignment over prompt design. Tonneau et al. (2024) evaluate hate speech detection in eight languages using real-world and functional test sets, but rely solely on vanilla prompting. Similarly, Dey et al. (2024) applied prompting LLMs to three low-resource South Asian languages, finding that translating inputs to English outperformed prompting in the original language. This motivated us to prompt the LLM to translate before classifying. In contrast to these efforts, our work covers eight languages and evaluates a broader range of prompt designs on real-world and functional test sets.

3 Datasets

We selected datasets with explicit hate speech labels that adhere to definitions commonly used in social science and by social media platforms: *abusive language that targets a protected group or individuals for being part of that group*.

The datasets, along with their overall sizes and the percentage of hateful instances, are summarized as follows: **OUS19_AR** (Ousidhoum et al., 2019): Contains 3,353 Arabic tweets, with 22.5% labeled as hateful. **OUS19_FR** (Ousidhoum et al., 2019): Consist of 4,014 French tweets, with 11.0% labeled as hateful. **BAS19_ES** (Basile et al., 2019): Compiled for SemEval 2019, it includes 4,950 Spanish tweets, 41.5% of which are labeled as hateful. **HAS21_HI** (Modha et al., 2021): Collected for HASOC 2021, it contains 4,594 Hindi tweets, with 12.3% labeled hateful. **SAN20_IT** (Sanguinetti et al., 2020): Created for Evalita 2020, it includes 8,100 Italian tweets, 41.8% of which are hateful. **FOR19_PT** (Fortuna et al., 2019): Consists of 5,670 Portuguese tweets, with 31.5% labeled as hateful. **Gahd24_DE** (Goldzycher et al., 2024):

A German adversarial dataset consisting of 10,996 tweets, 42.4% of which are labeled as hateful. **Xdomain_TR** (Toraman et al., 2022): A large-scale, multi-domain Turkish dataset consisting of 38K samples, with a class imbalance rate of 74.4%.

For functional hate speech evaluation, we used the **HateCheck benchmark** (Röttger et al., 2021), a benchmark for evaluating the robustness of hate speech detection systems across languages. It includes functional test cases—controlled examples designed to test specific capabilities such as handling implicit hate, negation, and non-hateful slurs. Originally developed for English, it has been extended by Röttger et al. (2022) to multiple languages to support cross-lingual evaluation and reveal systematic model weaknesses not captured by standard datasets.

4 Models

We evaluate four instruction-tuned multilingual LLMs for hate speech detection across eight languages: **LLaMA-3.1-8B-Instruct** (Grattafiori et al., 2024): Meta's instruction-tuned decoder model, optimized for reasoning tasks and primarily designed for English, with multilingual support. **Qwen2.5-7B-Instruct** (Qwen et al., 2025; Yang et al., 2024): A multilingual decoder model by Alibaba Cloud, supporting 30+ languages with strong instruction-following capabilities. **Aya-101** (Üstün et al., 2024): Cohere's multilingual model trained on 100+ languages, tuned for equitable cross-lingual NLP, including hate speech detection. **BloomZ-7B1** (Muennighoff et al., 2023a): A decoder model by BigScience, fine-tuned via multitask instruction tuning on 46 languages for cross-lingual instruction following.

For the encoder-based baseline, we fine-tuned two multilingual models with strong performance on classification tasks: **XLM-T** (Barbieri et al., 2022; Conneau et al., 2020): An XLM-R extension pre-trained on 198M Twitter posts in 30+ languages. **mDeBERTa** (He et al., 2021): A multilingual encoder covering 100+ languages, effective in zero-shot and low-resource settings. See Appendix A for model versions and additional details.

5 Experimental Setup

For each dataset, we randomly sampled 2,000 samples to serve as the test set for evaluating both prompting-based and fine-tuned models. Due to limited dataset sizes, the test sets for Arabic and French were restricted to 1,000 and 1,500 samples, respectively. Instruction-tuned multilingual LLMs were evaluated in inference-only mode, without additional fine-tuning, on both real-world and functional test sets. The models were prompted such that they responded with `yes` if the input text was hateful and `no` otherwise. Each experiment was repeated with three random seeds, and we alternated the order of `yes` and `no` in the prompt to reduce positional bias.

For the encoder-based models, after setting aside the test set, we held out 500 samples for validation and used the rest for training. After training, we evaluated the models on both their respective test sets, representing real-world evaluation, and on their language-specific subsets of the HateCheck benchmark, representing functional test evaluation. Model outputs and labels were mapped to binary values: 0 for non-hateful and 1 for hateful. Each experiment was run with five different random seeds, and the final results were averaged across these runs. Moreover, since several of the datasets are imbalanced, we report **F1-macro** as the primary evaluation metric to ensure fair assessment across classes. Further implementation details and hyperparameters are provided in Appendix A.

6 Prompts

We assess instruction-tuned multilingual LLMs using a range of prompting strategies for hate speech detection, such as: directly asking whether a comment is hateful (vanilla); prompting the model to act as a classifier (classification); chain-of-thought prompting for step-by-step reasoning (CoT); natural language inference-inspired prompts (NLI); language-aware prompts that consider linguistic and cultural context (cultural); assigning the LLM the role of a community moderator (role-play); translate then classify prompts (translation); definition-based prompts that explain hate speech (definition); and defining related forms of abusive content to help the model differentiate them from hate speech (distinction), etc. We also include few-shot prompting, where we retrieve and insert example instances from the training set into the prompt. We also explore combinations of these strategies. For full prompt texts and implementation details, see Appendix B.

		BloomZ		Aya101		Llama3				Qwen			
		zero-shot		zero-shot		zero-shot		Few Shot		zero-shot		Few Shot	
		prompt	f1	prompt	f1	prompt	f1	prompt	f1	prompt	f1	prompt	f1
Real World Tests	es	Classification	54.50	Definition	63.68	Classification	63.13	5 shot + CoT	**68.89**	Translation	64.79	5 shot + CoT	**68.90**
	pt	Definition	63.92	Definition	71.51	Role Play	70.79	5 shot + Cultural	**73.70**	Role Play + CoT	**73.44**	5 shot + Role Play	72.56
	hi	Cultural	51.33	Classification	47.33	CoT	52.09	5 shot + Role Play	**55.55**	Distinction	53.76	1 shot + CoT	49.57
	ar	NLI	58.67	Distinction	64.67	Classification	62.66	5 shot + Cultural	66.93	NLI	**70.61**	5 shot	65.88
	fr	NLI	**55.63**	Translation	53.44	CoT	55.22	5 shot + Definition	51.53	NLI	**55.59**	5 shot	51.78
	it	CoT	55.50	Vanilla	74.82	Distinction	75.86	5 shot + CoT	76.18	Cultural	73.34	5 shot + Cultural	**79.00**
	de	CoT	38.36	Vanilla	67.51	Role Play	50.16	5 shot + Cultural	**78.14**	Target	50.19	5 shot + Definition	77.55
	tr	Role Play	55.20	-	-	Classification	76.16	5 shot + CoT	**81.76**	Translation	75.89	5 shot + CoT	77.03
Functional Tests	es	Definition	64.88	Distinction	73.19	Vanilla	86.37	5 shot	**86.45**	Vanilla	84.39	5 shot + Definition	**86.43**
	pt	Definition	66.04	Distinction	72.39	Classification	83.37	3 shot	**86.59**	CoT	82.15	5 shot + Definition	84.08
	hi	Role Play	51.99	Distinction	65.95	Classification	65.31	1 shot + Cultural	65.36	Definition	65.41	1 shot + Definition	**66.61**
	ar	Definition	62.08	Vanilla	62.99	Impact	64.00	1 shot	67.95	Vanilla	70.42	3 shot + Definition	**71.88**
	fr	CoT	63.34	Distinction	71.94	Vanilla	84.61	5 shot + Role Play	84.37	Vanilla	82.06	5 shot + Definition	**86.08**
	it	Role Play	55.15	Distinction	71.25	Role Play	79.72	5 shot + COT	**87.08**	Target	78.35	5 shot + Definition	84.17
	de	Role Play	51.75	Distinction	72.64	Classification	85.86	5 shot + Cultural	**89.65**	Impact	82.64	5 shot + Definition	86.62

Table 1: Zero-shot and few-shot prompting results for instruction-tuned multilingual LLMs. The best or near-best results for each language for both evaluation setups are highlighted in **bold**. F1 refers to F1-macro.

7 Results

We evaluate instruction-tuned LLMs with various prompt types over three runs in the inference mode and report the average F1-macro scores. Table 1 summarizes the performance of zero-/few-shot results for four instruction-tuned models across eight languages. We observe that prompt design significantly affects performance. *Aya101* performs best with definition- and distinction-based prompts, suggesting that explicit definitions improve its accuracy. In contrast, *Qwen* excels with NLI and role-play prompts, indicating sensitivity to context and conversational cues.

In zero-shot settings, Qwen and LLaMA3 generally outperform the other models, with similar overall performance. However, Qwen performs better in most real-world test cases, whereas LLaMA3 leads on functional benchmarks. Few-shot prompting (typically five-shot) improves performance, especially on functional tests, as examples help the model apply contextual distinctions more effectively. On real-world tests, improvement is less consistent—even with examples from the same training data. This suggests that few-shot effectiveness depends not only on data quality but also on prompt clarity and structure. Overall, instruction-tuned LLMs perform notably well on functional tests and reasonably well on real-world tests in different languages. However, their effectiveness depends heavily on prompt design and the inclusion of few-shot examples. Appendix D contains detailed performance results.

For comparison, we fine-tune two encoder models for binary hate speech classification on train

		fine-tuned mDeBERTa	fine-tuned XLM-T	zero-shot prompting	few-shot prompting
Real World Tests	es	81.45	**82.78**	64.79	68.90
	pt	**73.22**	72.62	**73.44**	**73.70**
	hi	51.34	**59.18**	53.76	55.55
	ar	68.34	**70.31**	**70.61**	67.36
	fr	51.56	51.42	**55.63**	51.78
	it	**79.71**	78.82	75.86	79.00
	de	**80.39**	79.18	67.51	78.14
	tr	**92.72**	88.32	76.16	81.76
Functional Tests	es	60.94	67.93	**86.37**	**86.45**
	pt	58.94	57.28	83.37	**86.59**
	hi	24.91	23.26	65.95	**66.61**
	ar	23.93	25.47	70.42	**71.88**
	fr	25.89	26.61	84.61	**86.08**
	it	54.07	52.05	78.54	**87.08**
	de	74.36	70.60	83.27	**89.65**

Table 2: Results (f1-macro) of fine-tuned encoder models vs. best zero-/few-shot prompting LLMs. The best or near-best results for each language for both evaluation setups are highlighted in **bold**.

sets of datasets using five random seeds and report the average macro F1 scores. Table 2 summarizes the performance of encoder models alongside the best zero- and few-shot prompting results. On real-world datasets, encoder models generally outperform LLM prompting across most languages, benefiting from fine-tuning on task-specific data. However, the trend reverses on functional tests, where few-shot prompting often yields better results—highlighting the stronger generalization ability of large LLMs in controlled evaluation settings.

To understand when prompting is preferable, we conducted additional experiments comparing encoder model performance at varying training set sizes to that of prompting. Figure 1 presents results for three languages where prompting underperforms compared to fine-tuned models. Depending on the language, prompting becomes compet-

Figure 1: Performance of zero-/few-shot prompted LLMs vs. fine-tuned XLM-T across varying training sizes.

itive when training data is limited—for example, with 100–200 examples in Spanish, 300–400 in Hindi, or 600-700 in German. Beyond that, fine-tuning generally yields better performance. See Appendix C for more results across other languages.

8 Conclusion

In this study, we explore the capabilities of multilingual instruction-tuned LLMs in detecting hate speech across eight non-English languages. The findings suggest that different prompting techniques work better for different languages, indicating that it is beneficial to experiment with various prompt designs when addressing a new language. In real-world scenarios, where the data is more culturally dependent, prompting LLMs is less effective than training encoder models with task-specific data. However, in functional hate speech tests, LLMs tend to perform better and offer more flexibility. Incorporating few-shot examples into prompts in such cases may further enhance the LLMs' performance.

Limitations

One unavoidable limitation of our work is the number of multilingual instruction-tuned LLMs we were able to include. Given the rapid growth and proliferation of generative AI models, new LLMs are continually emerging. However, due to resource and time constraints, we were unable to include more models in our evaluation. We also did not fine-tune the instruction-tuned LLMs to better adapt them to our datasets.

A second limitation concerns the additional contextual information available for prompt construction. Most of our datasets included only the text, label, and language, but lacked richer metadata. Incorporating information such as the targeted group of the hate speech, the context in which it occurred, or the domain of the text could potentially improve model performance (Roy et al., 2023).

Finally, while we incorporated a wide range of carefully designed prompt variations to probe model behavior, our set of prompt configurations is not exhaustive. Alternative formulations or edge cases may exist that we have not explored. Therefore, our findings should be interpreted as indicative rather than definitive.

Acknowledgements

The work was supported by the European Research Council (ERC) through the European Union's Horizon Europe research and innovation programme (grant agreement No. 101113091) and the German Research Foundation (DFG; grant FR 2829/7-1). Daryna Dementieva's work was additionally supported by Friedrich Schiedel TUM Think Tank Fellowship.

References

Muhammad Ahmad, Muhammad Usman, Sulaiman Khan, Muhammad Muzamil, Ameer Hamza, Muhammad Jalal, Ildar Batyrshin, Usman Sardar, and Carlos Aguilar-Ibañez. 2025. Hate speech detection using social media discourse: A multilingual approach with large language model. *African Journal of Biomedical Research*, 28(2S):321–328.

Mistral AI. 2024. Un ministral, des ministraux. https://mistral.ai/news/ministraux. Accessed: 2025-04-19.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Jan Ebert, Alexander Arno Weber, Richard Rutmann, Charvi Jain, Max Lübbering, Daniel Steinigen, Johannes Leveling, Katrin Klug, Jasper Schulze Buschhoff, Lena Jurkschat, Hammam Abdelwahab, Benny Jörg Stein, Karl-Heinz Sylla, Pavel Denisov, Nicolo' Brandizzi, Qasid Saleem, Anirban Bhowmick, Lennard Helmer, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Alex Jude, Lalith Manjunath, Samuel Weinbach, Carolin Penke, Oleg Filatov, Shima Asaadi, Fabio Barth, Rafet Sifa, Fabian

Küch, Andreas Herten, René Jäkel, Georg Rehm, Stefan Kesselheim, Joachim Köhler, and Nicolas Flores-Herr. 2024. Teuken-7b-base & teuken-7b-instruct: Towards european llms.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of LREC 13th*, pages 258–266. ELRA.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63. ACL.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 58th*, pages 8440–8451. ACL.

Krishno Dey, Prerona Tarannum, Md. Arid Hasan, Imran Razzak, and Usman Naseem. 2024. Better to ask in english: Evaluation of large language models on english, low-resource and cross-lingual settings.

Fatema Tuj Johora Faria, Laith H. Baniata, and Sangwoo Kang. 2024. Investigating the predominance of large language models in low-resource bangla language over transformer models for hate speech detection: A comparative analysis. *Mathematics*, 12(23).

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104. ACL.

Janis Goldzycher, Paul Röttger, and Gerold Schneider. 2024. Improving adversarial data collection by supporting annotators: Lessons from GAHD, a German hate speech dataset. In *Proceedings of NAACL 2024*, pages 4405–4424. ACL.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, and et al. 2024. The llama 3 herd of models.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2023. An investigation of large language models for real-world hate speech detection. In *ICMLA 2023*, pages 1568–1573. IEEE.

Lawrence Han and Hao Tang. 2022. Designing of prompts for hate speech recognition with in-context learning. In *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 319–320. IEEE.

Jianfei He, Lilin Wang, Jiaying Wang, Zhenyu Liu, Hongbin Na, Zimu Wang, Wei Wang, and Qi Chen. 2024. Guardians of discourse: Evaluating llms on multilingual offensive language detection. In *Proceedings of IEEE Smart World Congress*, pages 1603–1608. IEEE.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023*, pages 294–297.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18(2):1–36.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Sarah Masud, Sahajpreet Singh, Viktor Hangya, Alexander Fraser, and Tanmoy Chakraborty. 2024. Hate personified: Investigating the role of LLMs in content moderation. In *Proceedings of EMNLP 2024*, pages 15847–15863. ACL.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of EMNLP 2022*, pages 11048–11064. ACL.

Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech. In *Proceedings of the 13th annual meeting of the forum for information retrieval evaluation*, pages 1–3.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie,

Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023a. Crosslingual generalization through multitask finetuning. In *Proceedings of ACL 61st*, pages 15991–16111. ACL.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023b. Crosslingual generalization through multitask finetuning.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of EMNLP 2019 and IJCNLP 9th*, pages 4675–4684. ACL.

Filippo Pedrazzini. 2025. Multilingual llms: Progress, challenges, and future directions. Accessed: 2025-04-14.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2023. When does in-context learning fall short and why? a study on specification-heavy tasks.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169. ACL.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of ACL 59th and IJCNLP 11th*, pages 41–58. ACL.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of EMNLP 2023*, pages 6116–6128. ACL.

Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task. *Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of ACL 2024*, pages 2304–2317. ACL.

Michał Skibicki. 2025. Large language models: Functionality and impact on everyday applications. Accessed: 2025-04-14.

Daniela Stockmann, Sophia Schlosser, and Paxia Ksatryo. 2023. Social media governance and strategies to combat online hatespeech in germany. *Policy & Internet*, 15(4):627–645.

Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. 2021. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 247–267.

Manuel Tonneau, Diyi Liu, Niyati Malhotra, Scott A. Hale, Samuel P. Fraiberger, Victor Orozco-Olvera, and Paul Röttger. 2024. Hateday: Insights from a global hate speech dataset representative of a day on twitter.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of LREC 13th*, pages 2215–2225. ELRA.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of ACL 62nd*, pages 15894–15939. ACL.

Janikke Solstad Vedeler, Terje Olsen, and John Eriksen. 2019. Hate speech harms: A social justice discussion of disabled norwegians' experiences. *Disability & Society*, 34(3):368–383.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng,

Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Anwar Hossain Zahid, Monoshi Kumar Roy, and Swarna Das. 2025. Evaluation of hate speech detection using large language models and geographical contextualization.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction tuning for large language models: A survey.

Yaqi Zhang, Viktor Hangya, and Alexander Fraser. 2025. LLM sensitivity challenges in abusive language detection: Instruction-tuned vs. human feedback. In *Proceedings of COLING 31st*, pages 2765–2780. ACL.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2025. Exploring the capability of chatgpt to reproduce human labels for social computing tasks. In *Social Networks Analysis and Mining*, pages 13–22. Springer Nature Switzerland.

A Model and Training Details

A.1 Instruction-tuned LLM Setup

To select suitable instruction-tuned multilingual LLMs, we first conducted a brief experiment to ensure that their safety tuning would not interfere with hate speech classification. Our goal was to evaluate detection capabilities, not robustness to jailbreak attempts. We excluded models such as mT0-large (Muennighoff et al., 2023b), Ministral-8B-Instruct (AI, 2024), and Teuken-7B-instruct (Ali et al., 2024) that failed to follow instructions reliably. We used the `transformers` library to load and run models in inference mode, generating binary outputs (`yes` or `no`). We set `max_new_tokens=10`, `do_sample=False`, and left temperature/top-k/top-p unset. Batch size and max sequence length varied depending on the prompt and model.

A.2 Encoder Model Training

For training the encoder-based models, in addition to the previously mentioned 2,000-sample test set, we randomly held out 500 samples for validation and used the remaining data for training. Models were fine-tuned for 10 epochs using the `transformers Trainer`, with a batch size of 16 and max sequence length of 128. Default settings were used for the learning rate, optimizer, and scheduler.

A.3 Data Formatting

Most datasets used were binary hate vs. non-hate classification tasks. Any remaining datasets, such as German and Turkish ones, were also converted to this binary format to ensure consistency. The datasets we used in this study are legally licensed and permitted for use in research projects.

A.4 Model Size and Budget

Experiments with instruction-tuned LLMs—LLaMA3[3], Qwen2.5[4], Aya101[5], and BloomZ[6]—were primarily conducted on NVIDIA RTX A6000 servers in inference mode, with no parameter updates during prompting. In contrast, fine-tuning of encoder models was performed on NVIDIA GeForce GTX 1080 Ti GPUs, where all model parameters were updated during training. The mDeBERTa[7] has approximately 86 million parameters, while XLM-T[8] consists of around 279 million parameters. All models used in this study were sourced from Hugging Face and are licensed for legal use in academic research.

[3] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[4] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[5] https://huggingface.co/CohereLabs/aya-101

[6] https://huggingface.co/bigscience/bloomz-7b1

[7] https://huggingface.co/microsoft/mdeberta-v3-base

[8] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base

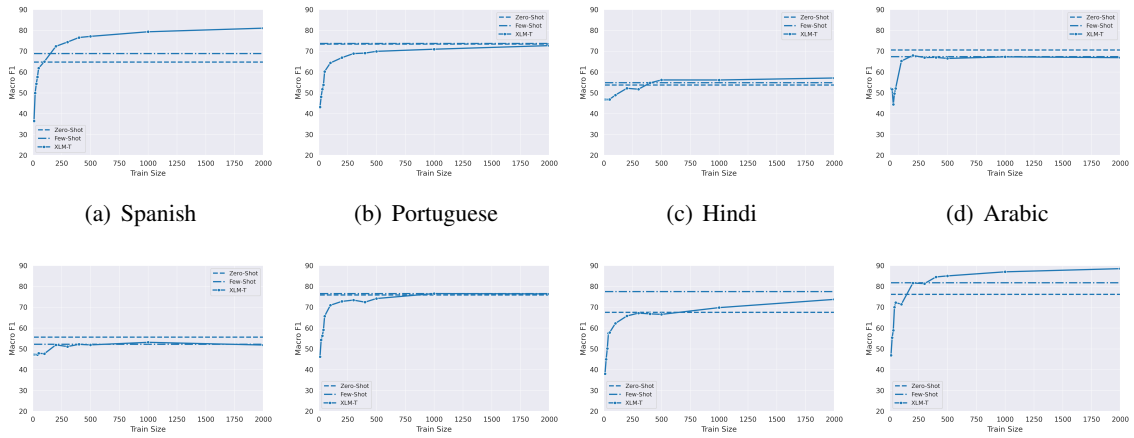
B Prompts Details

You can find the zero-shot prompt texts in Table 3and Table 4 and the few-shot prompt texts in Table 5. In these tables, "CoT" refers to chain-of-thought prompting, and "NLI" refers to prompts guided by natural language inference. The "+" symbol indicates a combination of the two prompt types. Dataset instances are enclosed in `{text}`. The placeholders `{yn1}` and `{yn2}` correspond to the expected outputs ("yes" and "no"), which were randomly swapped to reduce positional bias. The input language is represented as `{language}`, and if the prompt requires few-shot examples, they are inserted in `{examples}`.

In our few-shot experiments, we retrieve 1, 3, or 5 examples per class from the training set and include them in the prompt to guide the model's predictions. For example, a 5-shot setting includes five hate and five non-hate examples, interleaved by class (e.g., one hate, one non-hate), resulting in a total of 10 examples. While we experimented with using more than five examples per class, context length limitations and computational constraints prevented us from applying this to all prompts and models.

C Comparing Prompting and Fine-tuning Under Varying Data Conditions

Figure 2 illustrates the performance of the XLM-T model fine-tuned on training sets ranging from 10 to 2,000 instances across various languages, alongside the best zero-/few-shot results from instruction-tuned LLMs. Notably, in Portuguese, Arabic, French, and Italian, zero- or few-shot prompting matches or exceeds the performance of XLM-T even when trained on 2,000 labeled examples. In other languages, prompting performs competitively when training data is limited, offering a strong alternative in low-resource settings. As expected, fine-tuning generally surpasses prompting when sufficient labeled data is available, highlighting a practical trade-off between data availability and model adaptation strategy.



(a) Spanish (b) Portuguese (c) Hindi (d) Arabic

(e) French (f) Italian (g) German (h) Turkish

Figure 2: Performance of zero-/few-shot prompted LLMs vs. fine-tuned XLM-T across varying training sizes.

D Full Prompting Results

The complete results for Spanish and Hindi are shown in Table 6. Italian and German results are in Table 7, Arabic and French in Table 8, and Portuguese in Table 9. The zero-shot and few-shot results for each language in both evaluation setups for each LLM are highlighted in **bold**. The symbol '—' indicates failure due to excessive prompt length or runtime issues.

Prompt Type	Prompt Text
Vanilla	Is the following comment hateful? Respond only with {yn1} or {yn2}. Don't provide explanations or any other information. Comment: "{text}". Answer:
Classification	Classify the following comment as either hateful or non-hateful. Respond with yes for hateful and no for non-hateful. Answer only {yn1} or {yn2}. Don't provide explanations or any other information. Comment: "{text}". Answer:
Definition	Hate speech is language that expresses discrimination, hostility, or violence against individuals or groups based on race, religion, ethnicity, nationality, gender, sexual orientation, or disability. Does the following comment contain hate speech? Answer only {yn1} or {yn2}. Don't provide explanations or any other information. Comment: "{text}". Answer:
CoT	Analyze the following comment step by step to determine if it contains hate speech. Step 1: Identify if the comment targets a specific person or group based on race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity. If no target is identified, the comment is not hate speech. Step 2: Assess whether the language expresses hostility, discrimination, incitement to violence, dehumanization, or derogatory stereotypes toward the target. Step 3: Based on the above steps, conclude whether this is hate speech. Answer only {yn1} or {yn2}. Don't provide explanations or any other information. Comment: "{text}". Answer:
NLI	Premise: "{text}" Hypothesis: "This statement contains hate speech as defined by social and legal standards." Does the premise entail the hypothesis? Don't provide explanations or any other information. Answer only {yn1} or {yn2}. Answer:
Role Play	You are a community moderator responsible for ensuring compliance with hate speech policies. Analyze the following text and determine if it violates the guidelines. Text: "{text}" Does this text violate the community's hate speech policy? Don't provide explanations or any other information. Answer only {yn1} or {yn2}. Answer:
Cultural	Hate speech is defined as language that expresses discrimination, hostility, or violence against individuals or groups based on race, religion, ethnicity, nationality, gender, sexual orientation, or disability. Analyze the following text written in {language} considering cultural context, idioms, and societal norms of {language}-speaking communities. Does this text contain hate speech? Don't provide explanations or any other information. Answer only {yn1} or {yn2}. Text: "{text}". Answer:
Cultural + Definition	Hate speech is defined as language that expresses discrimination, hostility, or violence against individuals or groups based on race, religion, ethnicity, nationality, gender, sexual orientation, or disability. Analyze the following text written in {language} considering cultural context, idioms, and societal norms of {language}-speaking communities. Does this text contain hate speech? Don't provide explanations or any other information. Answer only {yn1} or {yn2}. Text: "{text}". Answer:
Role Play + CoT	You are a community moderator responsible for enforcing hate speech policies. Analyze the following comment step by step to determine if it violates these policies. Step 1: Identify any offensive or harmful language targeting a person or group. Step 2: Assess if the language attacks are based on attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity. Step 3: Determine if the comment violates the community's hate speech policy. Does this comment violate the community guidelines on hate speech? Answer only {yn1} or {yn2}. Don't provide explanations or any other information. Comment: "{text}". Answer:
Cultural + CoT	Analyze the following text written in {language} step by step, considering cultural context, idioms, and societal norms in {language}-speaking communities. Step 1: Identify any offensive or harmful language targeting a person or group. Step 2: Determine if the language attacks are based on attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity. Step 3: Conclude whether it is hate speech. Does this text contain hate speech? Don't provide explanations or any other information. Answer only {yn1} or {yn2}. Text: "{text}". Answer:

Table 3: Zero-shot prompt templates for hate speech detection in eight non-English languages (part 1).

Prompt Type	Prompt Text
Target	Identify the target group, if any, in the following comment. Then, determine if it contains hate speech. Does it contain hate speech? Answer only {yn1} or {yn2}. Comment: "{text}". Answer:
Impact	Consider the context of the following comment, including the target audience and potential impact. Does this comment contain hate speech? Answer only {yn1} or {yn2}. Don't provide explanations or any other information. Comment: "{text}". Answer:
Translation	Translate the following text from {language} to English: Text: "{text}" Now, classify the text as either hateful or non-hateful. Don't provide explanations or any other information. Respond with yes for hateful and no for non-hateful. Answer:
Distinction	Considering the following definitions, tell me if the given comment contains hate speech? - Hate Speech: Attacks, dehumanizes, incites violence, or promotes discrimination against a group based on identity (e.g., race, religion, gender, etc.). - Toxic Language: Aggressive, harassing, or harmful speech, but not necessarily identity-based. - Offensive Language: Rude or impolite but does not incite harm or target a group. Answer only {yn1} or {yn2}. Don't provide explanations or any other information. Comment: "{text}". Answer:

Table 4: Zero-shot prompt templates for hate speech detection in eight non-English languages (part 2).

Prompt Type	Prompt Text
Few Shot	Consider following examples: {examples} Now, classify this comment as either hateful or non-hateful: Comment: "{text}" Respond with yes for hateful and no for non-hateful. Answer only {yn1} or {yn2}. Don't provide explanations or any other information. Answer:
Few Shot + CoT	Here are some examples of comments and their classifications as yes if it is hateful or no if not: {examples} Now, analyze the following comment step by step and determine if it is hateful. Step 1: Identify any offensive or harmful language targeting a person or group. Step 2: Determine if the language attacks are based on attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity. Step 3: Conclude whether it is hate speech. Don't provide explanations or any other information. Answer only {yn1} or {yn2}. Don't provide explanations or any other information. Comment: "{text}". Answer:
Few Shot + Role Play	As a community moderator, you have previously handled the following cases they including hate speech or not: {examples} Based on these examples, analyze the following comment and determine if it violates the community's hate speech policy. Don't provide explanations or any other information. Answer only {yn1} or {yn2}. Comment: "{text}". Answer:
Few Shot + Cultural	Here are examples of comments in {language} and their classifications as yes if it is hateful or no if not: {examples} Now, analyze the following comment written in {language} and determine if it is hateful. Is this comment hateful? Don't provide explanations or any other information. Answer only {yn1} or {yn2}. Comment: "{text}". Answer:
Few Shot + Definition	Hate speech is defined as language that expresses discrimination, hostility, or violence against individuals or groups based on race, religion, ethnicity, nationality, gender, sexual orientation, or disability. Here are some examples: {examples} Is this comment hateful? Comment: "{text}" Don't provide explanations or any other information. Answer only {yn1} or {yn2}. Answer:

Table 5: Few-shot prompt templates for hate speech detection in non-English languages.

		Bas19_es								Has21_hi							
		Real-world test sets				Functional test sets				Real-world test sets				Functional test sets			
		llama3	aya101	bloomz	qwen	llama3	aya101	bloomz	qwen	llama3	aya101	bloomz	qwen	llama3	aya101	bloomz	qwen
Vanilla		40.14	62.56	44.47	62.26	**86.37**	67.33	36.30	**84.39**	25.00	36.17	48.06	33.25	40.56	42.13	23.89	64.11
Classification		**63.13**	60.16	**54.50**	62.64	83.92	39.06	38.53	81.16	32.48	**47.33**	35.13	31.87	**65.31**	55.08	33.41	**65.27**
Definition		62.21	**63.68**	55.98	42.35	82.79	58.71	**64.88**	80.46	31.67	33.48	44.11	31.98	**65.27**	60.33	50.94	**65.41**
CoT		50.65	37.19	50.26	42.58	33.59	28.08	60.44	55.58	**52.09**	20.41	21.60	36.10	39.86	27.02	50.56	64.35
Cultural		57.54	60.62	48.66	63.08	69.72	59.04	29.33	48.79	23.67	34.50	**51.33**	32.49	57.04	59.43	30.26	52.11
NLI		47.68	35.74	58.66	18.50	25.18	37.76	57.90	28.44	33.79	34.84	50.93	32.36	23.18	31.51	47.25	30.41
Role Play		58.99	60.68	55.90	43.79	78.03	44.25	56.81	55.39	29.63	30.72	49.59	32.54	59.60	58.19	**51.99**	56.57
Cultural + Definition		57.54	60.65	54.15	42.43	76.73	51.08	53.11	47.35	26.25	42.59	36.01	28.30	64.31	61.72	37.59	64.08
Role Play + CoT		55.42	26.90	29.58	59.92	74.25	41.94	41.27	73.87	35.42	9.09	10.55	46.55	59.69	46.06	41.10	63.30
Cultural + CoT		59.06	47.15	30.61	60.57	61.37	42.20	41.27	72.68	34.30	24.72	19.39	42.38	39.13	55.39	41.10	57.53
Target		41.03	37.41	36.82	**64.77**	47.91	30.61	23.38	81.28	29.23	11.00	47.28	32.53	40.79	17.42	23.20	64.20
Impact		60.83	40.49	46.25	61.74	80.34	42.69	44.90	82.52	29.30	30.40	47.00	32.40	62.88	38.29	25.39	55.29
Translation		55.91	39.05	35.89	**64.79**	71.84	44.42	40.50	76.16	18.36	33.86	14.72	34.20	**65.19**	40.75	30.65	51.55
Distinction		62.80	62.94	36.80	63.42	78.06	**73.19**	23.53	78.38	49.00	-	47.17	**53.76**	61.48	**65.95**	23.15	62.02
Few Shot	1	44.76	23.42	44.76	42.11	86.44	28.46	34.41	53.74	34.49	9.56	43.64	27.47	43.83	25.41	23.72	41.41
	5	47.85	-	50.78	45.33	58.63	-	58.81	56.90	37.02	-	47.52	30.77	44.10	-	23.20	63.40
Few Shot + CoT	1	64.44	28.09	54.27	64.34	82.93	32.59	35.54	80.68	**54.90**	-	45.49	**49.57**	64.33	18.92	24.66	63.65
	5	**68.89**	-	55.15	**68.90**	**86.45**	-	38.36	84.03	-	-	48.27	49.61	44.35	-	24.67	63.19
Few Shot + Role Play	1	43.61	37.55	57.06	66.44	55.86	37.88	36.83	81.60	50.63	19.46	46.45	29.35	64.46	35.63	23.72	42.36
	5	46.01	-	53.25	45.88	56.29	-	29.77	83.45	55.55	-	48.33	33.39	60.44	-	23.20	**66.61**
Few Shot + Cultural	1	65.25	42.51	**58.71**	64.41	84.89	42.88	43.10	81.03	34.33	25.95	40.83	29.99	**65.36**	34.10	24.84	42.36
	5	**68.35**	-	57.82	45.28	84.94	-	45.03	82.99	36.01	-	45.98	28.26	62.78	39.32	27.30	42.48
Few Shot + Definition	1	64.25	41.09	52.40	64.83	82.53	41.99	29.18	80.06	34.75	24.03	47.22	28.75	44.14	40.86	23.41	65.93
	5	66.85	-	48.85	66.94	83.59	-	25.00	**86.43**	47.81	-	**48.61**	32.69	42.40	-	23.20	65.22

Table 6: Complete Zero- and Few-shot Prompting Results for Spanish & Hindi.

		San20_it								Gahd24_de							
		Real-world test sets				Functional test sets				Real-world test sets				Functional test sets			
		llama3	aya101	bloomz	qwen	llama3	aya101	bloomz	qwen	llama3	aya101	bloomz	qwen	llama3	aya101	bloomz	qwen
Vanilla		48.31	**74.82**	43.65	46.18	57.40	69.28	24.06	53.50	48.30	**67.51**	26.43	48.41	82.38	66.35	24.58	54.50
Classification		50.91	50.58	39.35	49.43	74.94	40.21	27.47	77.69	**50.04**	43.71	31.53	49.68	**85.86**	38.50	28.89	49.53
Definition		49.03	**74.51**	37.87	47.04	55.43	58.70	49.45	76.97	48.38	45.75	35.67	49.19	52.03	61.34	42.60	75.28
CoT		54.55	-	**55.50**	45.45	30.15	27.66	50.09	77.93	44.08	-	**38.36**	48.52	75.53	28.61	49.21	52.80
Cultural		46.77	48.64	31.73	50.67	68.23	62.41	25.68	48.10	43.25	44.58	27.63	45.99	61.48	58.14	25.42	50.15
NLI		32.90	45.21	40.04	19.29	23.29	35.87	37.56	28.73	26.09	34.09	30.88	10.56	25.37	40.35	36.26	30.81
Role Play		48.94	70.55	35.64	50.48	**78.54**	44.00	**55.15**	51.08	**50.16**	43.81	37.14	48.43	80.03	47.56	**51.75**	53.02
Cultural + Definition		45.94	44.09	35.84	51.15	79.72	52.20	42.32	77.31	46.44	43.55	32.20	48.08	66.55	65.15	33.57	77.42
Role Play + CoT		65.68	-	29.45	72.55	77.60	42.79	41.18	73.48	43.81	-	19.95	49.54	71.10	44.82	41.16	70.47
Cultural + CoT		49.33	51.09	33.99	**73.34**	40.61	43.04	41.18	71.42	26.78	-	20.30	48.09	71.80	47.64	41.16	70.43
Target		46.96	35.00	24.51	49.85	45.23	30.49	23.08	**78.35**	39.32	34.32	24.25	**50.19**	37.06	31.22	23.15	75.62
Impact		47.51	49.37	25.36	44.80	72.63	42.09	31.85	73.63	48.33	43.55	25.25	45.55	83.27	38.38	24.99	**82.64**
Translation		43.44	45.30	25.80	50.36	74.45	44.91	35.15	71.87	48.36	35.33	27.39	45.65	71.02	43.71	32.96	75.45
Distinction		**75.86**	-	36.81	71.52	75.31	**71.25**	23.16	75.16	-	-	24.38	49.82	75.94	**72.64**	23.16	70.20
Few Shot	1	73.97	28.40	51.31	48.58	56.12	28.76	37.03	47.82	74.67	24.05	55.92	43.98	84.92	28.50	42.51	36.61
	5	48.71	32.08	**58.74**	52.05	82.95	31.40	41.13	53.11	77.07	28.70	55.92	49.85	88.37	29.07	47.02	55.25
Few Shot + CoT	1	73.16	15.36	52.89	73.54	82.10	22.09	33.91	79.35	76.16	-	52.44	76.26	84.12	28.56	35.69	81.43
	5	**76.18**	-	51.03	76.57	**87.08**	-	45.90	83.00	-	-	58.66	**77.04**	88.39	-	48.26	85.19
Few Shot + Role Play	1	48.58	29.21	47.66	73.91	54.33	33.40	34.91	80.15	50.69	27.52	58.60	76.24	58.16	27.61	47.70	80.10
	5	48.35	-	56.00	50.76	80.08	33.97	35.65	81.45	51.78	31.54	57.33	78.42	**89.54**	27.54	46.00	85.28
Few Shot + Cultural	1	74.83	38.24	55.57	76.36	80.56	36.69	44.18	78.45	75.16	40.11	61.34	49.36	86.67	34.58	54.07	80.87
	5	71.43	-	55.14	**79.00**	77.65	37.93	49.17	80.50	**78.14**	43.16	**64.40**	51.28	**89.65**	-	**57.25**	83.44
Few Shot + Definition	1	74.44	33.00	42.38	71.16	82.79	37.03	28.80	80.12	47.96	36.84	50.91	**77.22**	82.53	31.36	37.05	81.14
	5	47.04	-	46.55	77.91	78.32	37.55	29.32	**84.17**	73.21	37.92	57.24	**77.55**	86.43	30.77	48.27	**86.62**

Table 7: Complete Zero- and Few-shot Prompting Results for Italian & German.

		Ous19_ar								Ous19_fr							
		Real-world test sets				Functional test sets				Real-world test sets				Functional test sets			
		llama3	aya101	bloomz	qwen	llama3	aya101	bloomz	qwen	llama3	aya101	bloomz	qwen	llama3	aya101	bloomz	qwen
Vanilla		27.63	51.61	51.27	60.41	42.87	**62.99**	33.89	**70.42**	28.73	46.15	48.28	49.75	**84.61**	66.78	38.11	**82.06**
Classification		**62.66**	44.90	56.33	56.34	61.22	38.01	50.29	68.51	50.51	39.37	49.34	43.81	83.31	42.57	47.40	52.05
Definition		50.72	52.63	50.34	61.87	61.71	**62.50**	**62.08**	68.98	29.85	47.88	48.87	46.73	54.99	59.58	**63.10**	77.59
CoT		60.70	29.24	31.17	64.36	39.20	27.64	57.66	68.65	**55.22**	24.18	38.36	53.74	61.34	29.79	**63.34**	81.25
Cultural		51.12	47.55	44.48	67.21	57.79	57.68	36.62	60.16	43.07	36.21	50.75	49.73	72.65	59.52	35.39	72.51
NLI		47.24	40.03	**58.67**	**70.61**	23.40	27.66	47.06	47.01	49.36	36.39	**55.63**	**55.59**	24.20	41.15	60.28	61.88
Role Play		47.74	48.35	50.99	64.71	62.05	47.67	56.32	65.85	47.78	45.38	48.84	47.89	78.62	46.26	59.68	80.03
Cultural + Definition		56.76	47.06	49.81	59.66	61.43	61.16	42.50	68.81	39.94	44.52	52.84	44.68	75.05	57.10	51.19	77.32
Role Play + CoT		42.40	13.76	18.96	49.56	61.35	41.82	41.13	64.75	38.75	16.31	9.42	35.67	76.83	42.39	41.15	72.19
Cultural + CoT		43.44	36.26	20.62	58.22	27.92	59.06	41.13	64.57	38.87	35.89	11.82	40.07	45.52	43.43	41.15	70.81
Target		34.42	18.31	43.37	61.86	34.46	23.39	23.52	**70.02**	27.95	16.47	47.26	45.13	41.39	31.56	24.00	77.76
Impact		34.28	28.70	43.86	64.23	**64.00**	38.05	36.90	63.18	32.01	23.79	51.75	51.69	78.52	42.45	49.28	79.08
Translation		34.46	43.33	28.84	66.46	58.25	39.04	39.59	59.01	29.78	**53.44**	24.31	48.15	70.65	44.63	34.99	49.06
Distinction		60.45	**64.67**	43.72	60.47	60.52	**62.54**	24.52	69.48	51.62	53.09	47.26	50.09	76.92	**71.94**	27.98	74.19
Few Shot	1	60.26	17.79	57.82	59.47	**67.95**	26.84	34.66	68.74	47.07	8.44	46.31	45.19	82.97	28.55	37.70	77.41
	3	61.56	18.78	**59.29**	64.62	64.13	26.44	35.53	65.89	47.93	8.22	46.33	48.71	82.17	27.84	40.62	80.82
	5	61.62	19.89	55.89	**65.88**	63.39	26.59	32.57	64.89	48.90	8.52	49.61	**51.78**	82.83	27.48	36.55	81.60
Few Shot + CoT	1	64.55	15.33	53.46	60.73	62.66	20.92	29.12	68.22	49.09	7.32	48.85	43.31	81.95	29.28	35.59	78.13
	3	65.79	-	53.24	62.01	63.21	16.11	37.23	68.04	50.30	7.47	50.95	45.27	83.32	29.12	43.76	80.68
	5	65.06	15.19	53.96	62.33	63.55	10.62	35.96	69.57	51.26	6.75	51.13	46.41	83.81	27.97	43.69	81.36
Few Shot + Role Play	1	59.63	14.76	55.15	62.30	43.62	22.15	40.11	69.26	44.87	9.75	48.91	46.39	55.35	30.94	41.37	80.28
	3	62.06	14.69	55.66	61.60	65.16	23.92	36.41	68.79	49.36	8.10	50.55	46.90	83.50	30.67	36.53	81.62
	5	63.05	15.47	55.63	63.54	62.74	25.67	33.05	68.26	49.57	8.86	50.71	47.14	**84.37**	31.12	32.79	82.61
Few Shot + Cultural	1	62.36	**23.20**	56.74	63.37	63.00	31.17	42.09	66.01	46.42	**19.10**	45.63	47.46	55.25	37.90	53.59	52.19
	3	65.82	23.12	57.42	64.06	59.52	38.55	**48.68**	66.48	47.79	17.26	44.61	32.29	82.68	**41.68**	59.97	79.99
	5	**66.93**	22.68	55.74	64.62	57.88	**40.59**	47.80	66.14	49.69	15.79	48.51	49.23	55.47	41.57	**60.27**	80.88
Few Shot + Definition	1	63.17	14.66	52.14	54.41	65.44	30.95	29.66	**71.40**	50.33	9.76	**52.10**	40.10	79.91	34.42	33.17	77.51
	3	64.71	14.17	54.24	58.82	61.50	31.03	30.11	**71.88**	50.07	14.65	50.28	42.29	82.15	34.30	37.70	82.37
	5	63.78	14.30	55.00	60.68	58.96	31.60	28.31	**71.59**	**51.53**	14.76	50.74	43.30	82.44	33.38	38.04	**86.08**

Table 8: Complete Zero- and Few-shot Prompting Results for Arabic & French.

		Real-world test sets				Functional test sets			
		llama3	aya101	bloomz	qwen	llama3	aya101	bloomz	qwen
Vanilla		44.98	**71.08**	45.00	42.72	82.71	64.93	42.98	56.25
Classification		67.05	45.70	39.12	70.63	**83.37**	34.35	28.83	79.50
Definition		45.83	**71.51**	**63.92**	66.00	79.25	55.36	**66.04**	79.50
CoT		50.22	41.54	56.89	36.01	48.04	28.44	61.43	**82.15**
Cultural		67.64	67.68	49.39	66.97	76.88	56.38	37.26	75.78
NLI		49.54	41.36	53.68	8.51	35.47	40.45	58.81	37.60
Role Play		**70.79**	65.33	58.76	41.59	**82.22**	43.78	56.03	55.16
Cultural + Definition		63.29	60.29	48.28	69.49	69.82	48.01	50.48	78.71
Role Play + CoT		67.14	40.57	24.01	**73.44**	72.82	42.30	41.15	73.05
Cultural + CoT		45.91	55.23	35.90	72.06	78.14	41.91	41.25	73.22
Target		44.27	38.33	40.62	67.79	46.38	27.97	24.57	80.58
Impact		66.05	46.03	30.90	57.51	81.22	40.76	52.24	79.48
Translation		69.40	63.96	43.74	60.41	76.44	45.83	41.49	75.05
Distinction		69.82	67.11	40.62	59.75	78.30	**72.39**	25.36	77.36
Few Shot	1	46.16	21.18	49.97	70.60	85.80	28.94	45.59	51.28
	3	46.13	21.07	55.92	70.19	**86.59**	29.30	65.08	53.12
	5	47.00	21.19	59.69	70.39	57.48	29.44	**68.70**	53.59
Few Shot + CoT	1	70.63	26.28	53.76	72.55	83.78	27.66	33.75	78.71
	3	72.35	28.78	55.27	72.34	84.25	26.09	42.63	79.99
	5	72.98	25.80	55.23	72.53	84.22	22.55	44.59	80.57
Few Shot + Role Play	1	46.35	32.91	55.36	70.74	56.07	30.08	40.13	53.55
	3	69.63	33.24	54.29	72.36	83.32	31.57	38.86	82.38
	5	69.78	31.40	52.23	**72.56**	83.82	31.31	35.93	83.29
Few Shot + Cultural	1	71.81	44.30	60.42	70.64	85.14	42.58	49.75	80.05
	3	72.24	**44.60**	**60.78**	70.47	84.72	**42.70**	55.90	80.84
	5	**73.70**	43.88	60.54	71.13	84.78	42.30	55.55	82.16
Few Shot + Definition	1	71.06	33.39	50.79	71.74	83.19	45.67	37.20	80.40
	3	70.55	32.86	52.60	71.42	84.42	33.07	37.52	82.82
	5	71.42	32.41	51.50	71.59	84.72	32.54	36.51	**84.08**

Table 9: Complete Zero- and Few-shot Prompting Results for Portuguese.