Representation-Aware Prompting for Zero-Shot Marathi Text Classification: IPA, Romanization, Repetition

Van-Hien Tran, Huy Hien Vu, Hideki Tanaka, Masao Utiyama
National Institute of Information and Communications Technology (NICT)
{tran.vanhien, vuhuyhien, hideki.tanaka, mutiyama}@nict.go.jp

Abstract

Large language models (LLMs) often underperform in zero-shot text classification for low-resource, non-Latin languages due to script and tokenization mismatches. We propose *representation-aware prompting* for Marathi that augments the original script with International Phonetic Alphabet (IPA) transcriptions, romanization, or a repetition-based fallback when external converters are unavailable. Experiments with two instruction-tuned LLMs on Marathi sentiment analysis and hate detection show consistent gains over script-only prompting (up to +2.6 accuracy points). We further find that the most effective augmentation is model-dependent, and that combining all variants is not consistently beneficial, suggesting that concise, targeted cues are preferable in zero-shot settings.

1 Introduction

Large language models (LLMs) have recently demonstrated strong in-context learning (ICL) capabilities, enabling them to perform downstream tasks via natural-language prompts without updating model parameters. In many practical scenarios, however, collecting labeled demonstrations for few-shot prompting is costly and brittle: performance can vary substantially with the choice of examples, and constructing even a small, representative set may be infeasible for newly introduced domains or languages. These limitations motivate zero-shot ICL, where the model must solve a task using only an instruction and the input text.

Zero-shot text classification (ZTC) is a particularly useful setting for real-world deployment, but its effectiveness is highly uneven across languages. While instruction-tuned LLMs often perform competitively on high-resource languages (e.g., English), performance degrades noticeably on low-resource languages (Enomoto et al., 2025) such as Marathi. A key reason is *representation mismatch*: many non-Latin scripts are under-represented in pretraining data, leading to suboptimal subword segmentation and unstable token-level representations. As a result, the model may fail to reliably connect task instructions with the semantics of the input sentence, even when the underlying linguistic phenomenon is familiar. A natural remedy is to reduce this representation mismatch at the input level, without changing model parameters, by providing the same sentence in an additional form that the model can tokenize and align more reliably.

In this work, we study whether simple *input-level* interventions can narrow this gap without additional training. We propose *representation-aware prompting*, which augments the original Marathi script with alternative textual representations that are often easier for LLMs to process. Specifically, we investigate two widely used transformations: International Phonetic Alphabet (IPA) (International Phonetic Association, 1999) transcriptions and Romanization. IPA provides a script-independent phonetic rendering that can surface pronunciation-level regularities, whereas Romanization maps the sentence into Latin characters that are typically better covered by standard tokenizers. Importantly, we keep the original script in the prompt and treat the transformed string as complementary context, allowing the model to leverage both surface and transformed cues.

We further consider a practical constraint: IPA or Romanization tools may be unavailable or unreliable for some languages and domains. To address this, we introduce a lightweight *repeated-script* variant that requires no external resources. Inspired by recent findings that repetition can improve language model representations (Springer et al., 2025), we simply repeat the original script sentence within the prompt, encouraging the model to reprocess the same tokens before predicting a label.

We evaluate these prompting variants on two Marathi benchmarks—sentiment analysis and hate speech detection—using two strong instruction-tuned LLMs. Results show that representation-aware augmentations consistently improve accuracy over script-only prompting, with gains up to 2.6 points. We also observe that the most effective augmentation is model-dependent, and that combining multiple variants is not consistently beneficial, suggesting that zero-shot prompting often prefers concise, targeted cues over redundant context.

Our main contributions are: (i) a focused evaluation of IPA, Romanization, and repetition as input-level augmentations for ZTC in Marathi; (ii) results on two tasks with two instruction-tuned LLMs showing consistent gains without training; and (iii) an analysis indicating that more augmentation is not always better, underscoring the value of complementary, targeted representations in zero-shot settings.

2 Our Approach

In the baseline ZTC setup, the prompt includes only the task instruction and the Marathi-script input sentence. Figure 1 (upper part) shows an example template for Marathi sentiment analysis.

However, for low-resource languages, particularly for non-Latin languages (e.g, Marathi), providing only the input of the textual script as a traditional approach may not provide enough information and may be difficult for LLMs to understand the input. Thus, we investigate using three different data augmentation strategies as below.

International Phonetic Alphabet The International Phonetic Alphabet (IPA) provides a standardized, script-independent representation of pronunciation. We do not assume that LLMs are explicitly trained to interpret IPA; rather, we use it as an input-level cue to mitigate representation mismatch for low-resource, non-Latin scripts. By presenting the input in a pronunciation-aligned symbolic form, IPA can increase overlap with Latin/character-level patterns that are typically better covered during pre-training. Even when tokenized into fine-grained units, IPA strings retain consistent character-level regularities that can act as an auxiliary alignment signal alongside the original script.

A simple intuition comes from loanwords in Marathi: for example, the loanword *hospital* (typically written in Devanagari) may be difficult to process when presented only in its native script, whereas a pronunciation-aligned representation in a Latin-based phonetic notation (e.g., *hospital* or /hospital/) is closer to the familiar English form "hospital," facilitating retrieval of the intended lexical and semantic association.

```
You are an expert in Marathi sentiment analysis.

Task:
Classify the overall sentiment of the given Marathi
sentence into exactly one of the following labels:
- positive
- neutral
- negative

Return ONLY one label from the list above.
Do NOT include explanations, punctuation, or
additional text.

Marathi Sentence: {script_sent}
=> Label:
```

```
You are an expert in Marathi sentiment analysis.

Task:
Classify the overall sentiment of the given Marathi
sentence into exactly one of the following labels:
- positive
- neutral
- negative

Return ONLY one label from the list above.
Do NOT include explanations, punctuation, or
additional text.

Marathi Sentence: {script_sent}
{data_augmentation}
=> Label:
```

Figure 1: Zero-shot prompt templates for Marathi sentiment analysis: baseline (upper) and our augmentation-based variant (lower). `{script_sent}` denotes the Marathi-script input sentence, and `{data_augmentation}` denotes its augmented representation.

In our prompting template (the lower part of Figure 1), we include the IPA transcription as complementary input to the original script by substituting the placeholder `{data_augmentation}` with "IPA: `{ipa_sent}`", where `{ipa_sent}` denotes the IPA transcription of the original sentence `{script_sent}`.

Romanization Romanization maps non-Latin text into Latin characters using standardized transliteration schemes. When provided alongside the Marathi-script input, it can mitigate script and subword tokenization mismatches by increasing lexical overlap with high-resource Latin-script text seen during pretraining, thereby supporting cross-lingual transfer in zero-shot classification. In our prompting template (Figure 1), we incorporate the romanized sentence as complementary input to the original script by replacing `{data_augmentation}` with "Romanization: {roman_sent}", where {roman_sent} denotes the romanized form of `{script_sent}`.

Repetition Beyond IPA and Romanization, we also consider whether ZTC performance can be improved using only the original textual script when such augmented representations are unavailable. Our approach is motivated by recent findings that repetition can improve language model representations (Springer et al., 2025), particularly for autoregressive models that process input sequentially from left to right. Repeating the input sentence allows the model to revisit all original tokens within a single prompt, which may facilitate more thorough contextual processing and improve input understanding in low-resource settings. In our template (Figure 1), we implement this strategy by replacing `{data_augmentation}` with "Read Marathi Sentence again: `{script_sent}`", encouraging the model to reprocess the original script before making a prediction.

Finally, we test whether combining the proposed augmentations above yields additional gains in ZTC. Specifically, we construct a combined prompt that concatenates the IPA transcription, the Romanized form, and a repeated copy of the Marathi-script sentence in the `{data_augmentation}` slot (Figure 1). This setting assesses whether these complementary strategies yield additive benefits when applied together.

3 Experiments

3.1 Dataset and Settings

Dataset. We conduct experiments on the test sets of two publicly available benchmarks for Marathi text classification. L3CubeMahaSent (Kulkarni et al., 2021) is a sentiment analysis dataset annotated with three labels: *positive*, *neutral*, and *negative*. L3Cube-MahaHate (Patil et al., 2022) is a tweet-based hate speech detection dataset formulated as a binary classification with labels *HOF* and *NOT_HOF*. The detailed test-set statistics are reported in Table 1.

Settings. We conduct experiments using two LLMs: LLaMA-3.3-70B-Instruct and Qwen2.5-72B-Instruct, both publicly available on Hugging Face[1]. All model parameters are kept frozen throughout the experiments. To obtain augmented input representations, we use Indic-Transliteration[2] for Romanization and Epitran[3] (Mortensen et al., 2018) for IPA transcriptions of Marathi text.

For text generation, we employ greedy decoding without sampling, with a maximum of 100 newly generated tokens and FP16 precision. For the sentiment analysis task, we use the prompting template illustrated in Figure 1. For the hate detection task, we adapt this template to be suitable for the task (as shown in Appendix A). All experiments are performed on a machine equipped with eight NVIDIA A100 80GB GPUs, and classification performance is evaluated using accuracy. We report accuracy as the primary evaluation metric since the test sets of both benchmarks exhibit balanced class distributions, making accuracy an appropriate and interpretable measure of overall classification performance.

3.2 Results and Analysis

Main results. Table 2 summarizes zero-shot classification accuracy under different prompting variants. *Script* denotes the standard zero-shot setting that provides only the original Marathi sentence script. The variants +*IPA*, +*Roman*, and +*Repetition* augment this input by adding, respectively, the IPA transcription, the Romanized form, or a repeated copy of the original script sentence through `{data_augmentation}` in our template (Section 2 and Figure 1). Finally, +*All* concatenates all three augmented representations within a single prompt alongside the original script.

Across both tasks and models, augmenting the script generally improves performance over the script-only baseline, but the most effective augmentation is model-dependent. For LLaMA-3.3-70B-Instruct, +*IPA* yields the largest gains on both sentiment analysis (70.47→73.07, +2.60) and hate detection (92.51→94.69, +2.18), suggesting that phonetic cues provide particularly useful cross-lingual signals for this model. Romanization also improves LLaMA's performance (e.g., +2.13 on sentiment analysis; +1.54 on hate detection), while +*Repetition* provides smaller gains, indicating that for LLaMA, phonetic or transliteration-based signals are more informative than simply re-exposing the model to the same script.

In contrast, Qwen2.5-72B-Instruct benefits most from +*Repetition* on both benchmarks, achieving the best scores on sentiment analysis (71.80→73.40, +1.60) and hate detection (89.89→90.99, +1.10). IPA and Romanization remain helpful but yield smaller improvements for Qwen. Notably, this pattern is consistent across

[1]https://huggingface.co
[2]https://pypi.org/project/indic-transliteration/
[3]https://pypi.org/project/epitran/

Task	#Classes	#Test samples	Class distribution (test)
Sentiment Analysis	3	1,500	*Positive*: 500 (33.3%), *Neutral*: 500 (33.3%), *Negative*: 500 (33.3%)
Hate Detection	2	3,750	*HOF*: 1,875 (50%), *NOT_HOF*: 1,875 (50%)

Table 1: Statistics of the test sets used in our zero-shot Marathi text classification experiments.

Task	Model	Prompting Variants				
		Script	*+Roman*	*+IPA*	*+Repetition*	*+All*
Sentiment Analysis	LLaMA-3.3-70B-Instruct	70.47	72.6	**73.07**	72.27	70.67
	Qwen2.5-72B-Instruct	71.80	71.93	72.47	**73.40**	71.60
Hate Detection	LLaMA-3.3-70B-Instruct	92.51	94.05	**94.69**	92.61	94.08
	Qwen2.5-72B-Instruct	89.89	90.45	90.88	**90.99**	90.17

Table 2: Zero-shot text classification accuracy (%) on Marathi sentiment analysis and hate detection tasks using different prompting variants. Best results for each model and task are shown in bold.

both datasets: *+IPA* is the strongest single augmentation for LLaMA, whereas *+Repetition* is the strongest for Qwen. This consistency suggests that the relative utility of phonetic/transliteration cues versus repeated exposure is tied to model-specific pretraining and tokenization behavior rather than being an artifact of a particular task.

Interestingly, combining all augmentations (*+All*) does not yield further improvements and often underperforms the best single-augmentation setting. A plausible explanation is that zero-shot prompting benefits from concise and targeted contextual information: concatenating multiple representations can introduce redundancy, increase prompt length, and dilute the most informative cues, especially when the augmented forms are partially correlated with the original script. In addition, Romanization and IPA may carry conversion noise or ambiguities, which could become more salient when multiple augmented strings are provided simultaneously. Overall, these results suggest that selecting a single, complementary augmentation is often more effective than aggregating all available representations in a single prompt.

4 Related Work

LLMs have demonstrated strong ICL capabilities for zero-shot (ZS) and few-shot text classification (Brown et al., 2020; Milios et al., 2023; Edwards and Camacho-Collados, 2024; Wang et al., 2023). For low-resource languages, Wang et al. (2023) and Edwards and Camacho-Collados (2024) proposed methods of extracting knowledge from pretrained models to improve performance on this task. Meanwhile, Anikina et al. (2025) and Patwa et al. (2024) focused on enhancing synthetic datasets generated by LLMs, while Nie et al. (2023) and Fazili et al. (2024) extracted semantically similar sentences from higher-resource languages to augment low-resource data. Additionally, recent works (Kanjirangat et al., 2025; Petrov et al., 2023; Dewangan et al., 2025; Teklehaymanot and Nejdl, 2025; Pattnayak et al., 2025) have highlighted that tokenization itself remains a major bottleneck for low-resource languages, as subword segmentation disparities and suboptimal schemes have been shown to correlate strongly with degraded downstream performance, especially for morphologically rich languages like Marathi.

In order to mitigate the above problems on downstream tasks, other works have proposed using additional information such as romanization or IPA-based clues (Purkayastha et al., 2023; Nguyen et al., 2023; Sohn et al., 2024; Jung et al., 2024; J et al., 2024; Shurtz et al., 2025; Springer et al., 2025). These studies consistently report gains from incorporating phonological or transliteration-based signals into the modeling pipeline. However, these methods mainly focus on parameter-efficient fine-tuning or on different downstream tasks, and do not target zero-shot settings for text classification.

For Marathi specifically, only a handful of benchmark datasets have been introduced for text classification, most notably the L3CubeMahaSent (Kulkarni et al., 2021) and L3Cube-MahaSent-MD (Pingle et al., 2023) sentiment analysis corpora and the L3Cube-MahaHate (Patil et al., 2022) hate speech detection dataset. Existing work on these benchmarks has almost exclusively relied on supervised fine-tuning of monolingual and multilingual transformer models such as MahaBERT, IndicBERT, and multilingual BERT, or on cross-lingual transfer pipelines that translate Marathi inputs into high-resource languages before classification (Velankar et al., 2022). To our knowledge, there is still very limited work that systematically studies zero-shot LLM prompting for Marathi text classification.

5 Conclusion

In this work, we explored simple input-level augmentation strategies to improve ZTC for low-resource, non-Latin languages, focusing on Marathi. By augmenting the original script with IPA transcriptions, Romanized representations, or a repeated-script variant, we achieved consistent performance gains across two benchmark datasets and two LLMs without additional training or supervision. Our results show that the effectiveness of these augmentations is model- and task-dependent, and that combining all representations does not necessarily yield further improvements, suggesting that zero-shot prompting benefits from concise and targeted contextual information. Overall, this work highlights the potential of lightweight, representation-aware prompting strategies for improving ZTC in low-resource settings.

Limitations

Our experiments are limited to zero-shot text classification in Marathi using two very large instruction-tuned LLMs, and the observed improvements may not generalize to smaller models or to other low-resource languages. In addition, we focus on prompting strategies that retain the original Marathi script and treat alternative representations as complementary context; augmentation-only settings that remove the original script are not explored. Future work will evaluate the proposed representation-aware prompting strategies across a broader range of languages, scripts, and model scales to better assess their generality.

References

Tatiana Anikina, Jan Cegin, Jakub Simko, and Simon Ostermann. 2025. A rigorous evaluation of LLM data generation strategies for low-resource languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8282–8303, Suzhou, China. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Vikrant Dewangan, Bharath Raj S, Garvit Suri, and Raghav Sonavane. 2025. When every token counts: Optimal segmentation for low-resource language models. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025 - Workshops, Abu Dhabi, UAE, January 19-24, 2025*, pages 294–308. Association for Computational Linguistics.

Aleksandra Edwards and José Camacho-Collados. 2024. Language models for text classification: Is in-context learning enough? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 10058–10072. ELRA and ICCL.

Taisei Enomoto, Hwichan Kim, Zhousi Chen, and Mamoru Komachi. 2025. A fair comparison without translationese: English vs. target-language instructions for multilingual LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 649–670, Albuquerque, New Mexico. Association for Computational Linguistics.

Barah Fazili, Ashish Agrawal, and Preethi Jyothi. 2024. Boosting zero-shot crosslingual performance using llm-based augmentations with effective data selection. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 13406–13422. Association for Computational Linguistics.

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge, UK.

Jaavid J, Raj Dabre, Aswanth M, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. RomanSetu: Efficiently unlocking multilingual capabilities of large language models via Romanization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15593–15615, Bangkok, Thailand. Association for Computational Linguistics.

Haeji Jung, Changdae Oh, Jooeon Kang, Jimin Sohn, Kyungwoo Song, Jinkyu Kim, and David R Mortensen. 2024. Mitigating the linguistic gap with phonemic representations for robust cross-lingual transfer. In *Proceedings of the Fourth Workshop on*

Multilingual Representation Learning (MRL 2024), pages 200–211, Miami, Florida, USA. Association for Computational Linguistics.

Vani Kanjirangat, Tanja Samardzic, Ljiljana Dolamic, and Fabio Rinaldi. 2025. Tokenization and representation biases in multilingual models on dialectal NLP tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23992–24010, Suzhou, China. Association for Computational Linguistics.

Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. 2021. L3CubeMahaSent: A Marathi tweet-based sentiment analysis dataset. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 213–220, Online. Association for Computational Linguistics.

Aristides Milios, Siva Reddy, and Dzmitry Bahdanau. 2023. In-context learning for text classification with many labels. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 173–184, Singapore. Association for Computational Linguistics.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Hoang Nguyen, Chenwei Zhang, Tao Zhang, Eugene Rohrbaugh, and Philip Yu. 2023. Enhancing cross-lingual transfer via phonemic transcription integration. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9163–9175, Toronto, Canada. Association for Computational Linguistics.

Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. 2023. Cross-lingual retrieval augmented prompt for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8320–8340, Toronto, Canada. Association for Computational Linguistics.

Hrushikesh Patil, Abhishek Velankar, and Raviraj Joshi. 2022. L3Cube-MahaHate: A tweet-based Marathi hate speech detection dataset and BERT models. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 1–9, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Priyaranjan Pattnayak, Hitesh Laxmichand Patel, and Amit Agarwal. 2025. Tokenization matters: Improving zero-shot NER for indic languages. *CoRR*, abs/2504.16977.

Parth Patwa, Simone Filice, Zhiyu Chen, Giuseppe Castellucci, Oleg Rokhlenko, and Shervin Malmasi. 2024. Enhancing low-resource LLMs classification with PEFT and synthetic data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6017–6023, Torino, Italia. ELRA and ICCL.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Aabha Pingle, Aditya Vyawahare, Isha Joshi, Rahul Tangsali, and Raviraj Joshi. 2023. L3Cube-MahaSent-MD: A multi-domain Marathi sentiment analysis dataset and transformer models. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 274–281, Hong Kong, China. Association for Computational Linguistics.

Sukannya Purkayastha, Sebastian Ruder, Jonas Pfeiffer, Iryna Gurevych, and Ivan Vulić. 2023. Romanization-based large-scale adaptation of multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7996–8005, Singapore. Association for Computational Linguistics.

Ammon Shurtz, Christian Richardson, and Stephen D. Richardson. 2025. When scripts diverge: Strengthening low-resource neural machine translation through phonetic cross-lingual transfer. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 336–346, Suzhuo, China. Association for Computational Linguistics.

Jimin Sohn, Haeji Jung, Alex Cheng, Jooeon Kang, Yilin Du, and David R Mortensen. 2024. Zero-shot cross-lingual NER using phonemic representations for low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13595–13602, Miami, Florida, USA. Association for Computational Linguistics.

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2025. Repetition improves language model embeddings. In *The Thirteenth International Conference on Learning Representations*.

Hailay Kidu Teklehaymanot and Wolfgang Nejdl. 2025. Tokenization disparities as infrastructure bias: How subword systems create inequities in LLM access and efficiency. *CoRR*, abs/2510.12389.

Abhishek Velankar, Hrushikesh Patil, and Raviraj Joshi. 2022. Mono vs multilingual BERT for hate speech detection and text classification: A case study in marathi. In *Artificial Neural Networks in Pattern Recognition - 10th IAPR TC3 Workshop, ANNPR 2022, Dubai, United Arab Emirates, November 24-26,* *2022, Proceedings*, volume 13739 of *Lecture Notes in Computer Science*, pages 121–128. Springer.

Yuqi Wang, Wei Wang, Qi Chen, Kaizhu Huang, Anh Nguyen, and Suparna De. 2023. Prompt-based zero-shot text classification with conceptual knowledge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 30–38, Toronto, Canada. Association for Computational Linguistics.

A Appendix

We present the prompting template used for the hate detection task in Figure 2. In this setting, LLMs are instructed to perform binary classification on Marathi input sentences, assigning each instance to either the *HOF* or *NOT_HOF* label.

```
You are a Marathi Content Moderator. Your goal is to detect
Hate Speech and Offensive content (HOF).

Task:
Classify the given Marathi sentence into exactly one of the
following labels:
- HOF: Contains insults, hate speech, or offensive language.
- NOT_HOF: Neutral, respectful, or non-offensive criticism.

Return ONLY one label from the list above.
Do NOT include explanations, punctuation, or additional
text.

Marathi Sentence: {script_sent}
=> Label:
```

```
You are a Marathi Content Moderator. Your goal is to detect
Hate Speech and Offensive content (HOF).

Task:
Classify the given Marathi sentence into exactly one of the
following labels:
- HOF: Contains insults, hate speech, or offensive language.
- NOT_HOF: Neutral, respectful, or non-offensive criticism.

Return ONLY one label from the list above.
Do NOT include explanations, punctuation, or additional
text.

Marathi Sentence: {script_sent}
{data_augmentation}
=> Label:
```

Figure 2: Zero-shot text classification prompt templates for Marathi hate detection: baseline (upper) and our augmentation-based variant (lower).