Beyond Accuracy: A Structured Error Analysis of Multilingual LLMs on Marathi Script Variation and Syntax

Tejas Patil, Barnali Chetia
Department of Humanities and Social Sciences
Indian Institute of Information Technology Vadodara
{tejas_patil, barnali}@iiitvadodara.ac.in

Abstract

Evaluation of multilingual large language models has grown rapidly in recent years, yet Marathi, spoken by over 83 million people across India, has received almost no systematic probing beyond surface-level benchmark tests. Most existing multilingual evaluations either omit Marathi entirely or rely on machine-translated test sets that fail to capture the morphological complexity that defines the language. We evaluate four models, namely Llama-3.1-8B, Llama-3.3-70B, Mistral-7B, and Qwen3-32B, on our manually curated Marathi dataset across three probing dimensions: Devanagari versus Romanized script, Marathi-English code-mixing, and syntactic structures including SOV word order, vibhakti case markers, verb gender agreement, and postpositions. Models are tested under English and Marathi instruction conditions across translation, similarity, grammaticality, and case marker tasks. Translation quality is evaluated using both token-level F1 and BERTScore to capture paraphrase equivalence beyond surface word overlap. All models drop between 7.9% and 20.5% on Romanized input. The negative subjunctive marker *nasta* is ignored by every model. Vibhakti case markers are consistently replaced with Hindi equivalents, revealing that multilingual training has not produced separate internal representations for Hindi and Marathi despite their distinct morphological systems. These findings reveal structural gaps in how current multilingual LLMs handle morphologically rich, low-resource Indic languages and point to specific areas where dedicated Marathi pretraining data would most benefit future work.

1 Introduction

Marathi is a scheduled language of India with over 83 million speakers, making it the third most spoken language in the country. Despite this, it sits at the margins of large language model evaluation. Most multilingual benchmarks either skip Marathi entirely or rely on machine-translated test sets that strip away the morphological complexity that defines the language.

The challenge goes beyond vocabulary. Marathi follows Subject-Object-Verb word order and uses a rich system of postpositions. It encodes three-gender verb agreement and marks grammatical roles through vibhakti suffixes. These suffixes carry information that English conveys through prepositions and word order alone. For instance, the dative role in English expressed as "give *to* him" is marked in Marathi by attaching the suffix *-la* to the noun, yielding *tyaala*, with no preposition needed.

There is another layer of complexity. Marathi follows Balbodh, a form of Devanagari script, but speakers regularly switch to Roman script on mobile devices. Many users also mix English words into Marathi sentences at varying depths. Models struggle to distinguish Marathi from Hindi when both appear in Devanagari, often conflating the two in ways that cause systematic errors. None of this is unusual. It is how Marathi is actually used every day.

Prior work has built useful foundations through resources like IndicNLPSuite (Kakwani et al., 2020), Samanantar (Ramesh et al., 2022), and MahaBERT (Joshi, 2022). But LLM-specific probing for Marathi morphology and syntax remains almost entirely absent. We address this with a controlled probing study organized around three questions.

RQ1. Do LLMs perform differently on Devanagari versus Romanized Marathi, and how large is the gap?

RQ2. How well do models handle Marathi-English code-mixed text compared to monolingual Marathi?

RQ3. Which Marathi syntactic features are most challenging under realistic input conditions?

Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026), pages 119–126

We build a dataset of 1,020 sentences and evaluate four models. All results, code, and data are released to support follow-up work.

2 Related Work

2.1 Indic NLP Resources

Building NLP resources for Indian languages has accelerated since 2020. Kakwani et al. (2020) released IndicNLPSuite with monolingual corpora and the IndicGLUE benchmark for eleven Indian languages, including over 34 million Marathi sentences in their IndicCorp corpus. Kunchukuttan et al. (2020) released the IndicNLP library, a standard preprocessing tool for Indic pipelines. The Samanantar corpus (Ramesh et al., 2022) provides more than 49.7 million English-Indic sentence pairs. Despite this scale, Marathi lags significantly behind Hindi in task coverage. Lahoti et al. (2022) document persistent gaps in Marathi evaluation benchmarks. Dani and Sathe (2024) identify syntactic analysis and case marking as the weakest areas for current models, which directly motivated our probe design.

2.2 Multilingual Language Models

Conneau et al. (2020) introduced XLM-R covering 100 languages and substantially outperforming mBERT on cross-lingual tasks. Joshi (2022) released MahaBERT, MahaAlBERT, and MahaRoBERTa on 752 million Marathi tokens, showing monolingual models outperform multilingual ones on Marathi classification. Dabre et al. (2022) presented IndicBART for eleven Indic languages. Gala et al. (2023) extended coverage to all 22 scheduled Indian languages with IndicTrans2. BigScience Workshop (2022) released BLOOM across 46 languages, though Marathi coverage is limited.

2.3 LLM Evaluation and Code-Mixing

Ahuja et al. (2023) evaluated models on 70 languages and found sharp drops for underrepresented ones. Lai et al. (2023) extended MMLU to 26 languages and showed that performance gaps persist at scale. Bang et al. (2023) found consistent weaknesses for low-resource languages in ChatGPT. Khanuja et al. (2020) introduced GLUECoS for Hindi-English and Spanish-English code-switching. Most prior code-mixing work focuses on Hindi-English, leaving Marathi-English largely absent. Winata et al. (2021) showed that multilingual models struggle on code-mixed inputs even when they handle the constituent languages individually. Aguilar et al. (2020) proposed LinCE as a consolidated code-switching benchmark.

2.4 Script Variation and Syntactic Probing

FLORES-200 (Costa-jussà et al., 2022) includes Marathi only in Devanagari, leaving Romanized input entirely unevaluated. Tenney et al. (2019) established that BERT encodes syntactic structure across layers. Muller et al. (2021) showed multilingual models degrade for typologically distant languages during probing. Chi et al. (2020) found that cross-lingual syntactic transfer depends on morphological similarity. None of these studies target Marathi-specific features such as vibhakti or three-gender verb agreement.

2.5 Evaluation Metrics for Text Generation

Token-level F1 and BLEU have long served as standard metrics for translation evaluation, but both penalize correct paraphrases that do not match the reference surface form. Zhang et al. (2020) introduced BERTScore, which computes token similarity using contextual embeddings from pre-trained models, enabling semantic matching beyond exact word overlap. Rei et al. (2020) further showed that learned metrics correlate more strongly with human judgements than surface-level metrics on low-resource language pairs. We adopt BERTScore alongside token F1 in our evaluation to address the paraphrase equivalence limitation that affects Marathi translation assessment. syntactic transfer depends on morphological similarity. None of these studies target Marathi-specific features such as vibhakti or three-gender verb agreement.

3 Dataset Construction

3.1 Overview

We construct a probing dataset of 1,020 Marathi sentences across three probe types, two surface forms, four task types, and three difficulty levels. The original 500 sentences were written and verified by native Marathi speakers. An additional 520 sentences follow the same schema and are pending full native speaker verification. No machine translation is used at any stage.

3.2 Probe Types

Transliteration probe (367 sentences). Each sentence appears in Devanagari and Romanized form. Romanization follows informal conventions used by Marathi speakers online, reflecting realistic mobile input conditions.

Code-mixing probe (307 sentences). Sentences contain English words inserted into Marathi at the noun, verb, adjective, and clause levels. Patterns reflect genuine Marathi-English usage across technology, workplace, and social media domains in urban Maharashtra.

Syntax probe (346 sentences). Four syntactic features are targeted: SOV word order, vibhakti case markers, verb gender agreement, and postpositions. Each sentence is paired with a distractor containing a systematic error such as a Hindi case marker replacing a Marathi one, a verb gender mismatch, or English preposition order replacing Marathi postposition order.

3.3 Task Types and Conditions

Translation asks models to produce English output from Marathi input, evaluated using token-level F1 and BERTScore (Zhang et al., 2020). **Semantic similarity** asks whether two sentence variants share meaning, using binary accuracy. **Grammaticality judgement** presents two variants and asks which is correct Marathi. **Case marker selection** presents a sentence with one Marathi and one Hindi distractor marker. Sentences span easy (302), medium (323), and hard (395) difficulty levels. Each sentence is tested under English and Marathi instruction conditions, yielding 4,080 unique prompt conditions across the full dataset.

4 Experiments

4.1 Models

We evaluate four models. **Llama-3.1-8B** (Grattafiori et al., 2024) and **Mistral-7B** (Jiang et al., 2023) represent the small model regime. **Llama-3.3-70B** covers the medium-large range. **Qwen3-32B** (Qwen Team, 2025) rounds out the evaluation. All models use temperature 0 and a maximum of 50 output tokens. Models are accessed via the Groq and Mistral APIs.

4.2 Evaluation Metrics

Translation quality is measured using two metrics. Token-level F1 computes word overlap between predicted and reference translations. BERTScore (Zhang et al., 2020) uses contextual embeddings from RoBERTa-large to capture semantic similarity, addressing the paraphrase equivalence limitation of token F1. All other tasks use binary or multiple-choice accuracy.

Model	Parameters	Score
Qwen3-32B	32B	0.790
Llama-3.3-70B	70B	0.640
Mistral-7B	7B	0.624
Llama-3.1-8B	8B	0.526

Table 1: Overall performance averaged across all probe types, surface forms, and instruction conditions.

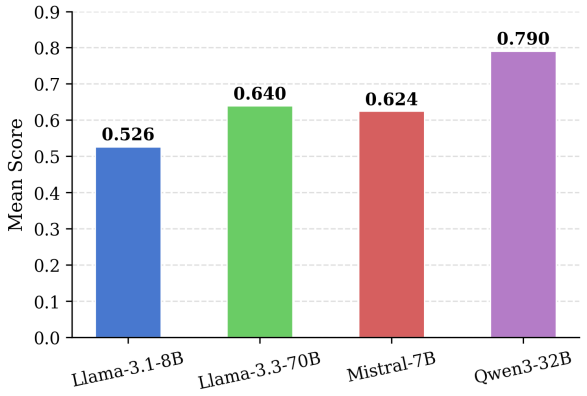

Figure 1: Overall mean performance per model. Qwen3-32B leads at 0.790.

5 Results

5.1 Overall Performance

Table 1 and Figure 1 report mean scores across all valid outputs. Qwen3-32B leads at 0.790, followed by Llama-3.3-70B at 0.640, Mistral-7B at 0.624, and Llama-3.1-8B at 0.526. Scale alone does not explain the ordering. Mistral-7B nearly matches the much larger Llama-3.3-70B, pointing to architecture and training data as important factors beyond parameter count.

5.2 Performance by Probe Type

Table 2 breaks results down by probe type. Transliteration is the hardest probe across all four models. Llama-3.1-8B scores only 0.257 here, the lowest single probe score in the study. Even Qwen3-32B reaches just 0.629 on transliteration, well below its scores on the other probes. Syntax probing shows the widest model gap: Qwen3-32B scores 0.919 while Mistral-7B scores 0.676 on identical sentences. Code-mixed text provides

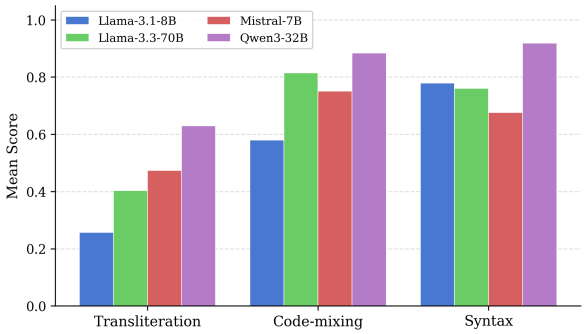

Figure 2: Performance by probe type. Transliteration is the hardest probe across all models.

Probe	L-8B	L-70B	M-7B	Q-32B
Transliteration	0.257	0.403	0.473	0.629
Code-mixing	0.579	0.815	0.751	0.884
Syntax	0.778	0.761	0.676	0.919

Table 2: Performance by probe type. L-8B = Llama-3.1-8B, L-70B = Llama-3.3-70B, M-7B = Mistral-7B, Q-32B = Qwen3-32B.

enough English vocabulary for models to recover partial meaning. Romanized Marathi removes those anchors entirely.

5.3 Script Variation Effect

Table 3 and Figure 3 show performance across script forms. All four models degrade when Marathi switches from Devanagari to Roman script. Llama-3.1-8B and Llama-3.3-70B both drop around 20%. Mistral-7B drops 17.5%. Qwen3-32B shows the smallest decline at 7.9%. Even this best case is not reassuring. A model that loses nearly 8% accuracy because a user typed in Roman script is unreliable for a large portion of real Marathi users. The drop is steepest for syntax probes, where performance falls from 0.461 on Devanagari to 0.338 on Romanized input across all models combined.

5.4 Instruction Language Effect

Table 4 shows the effect of instruction language on model performance. Qwen3-32B shows the most striking pattern, scoring 0.709 under English instructions but jumping to 0.873 under Marathi instructions, a gap of 16.4 points. No other model shows this magnitude. Llama-3.1-8B actually degrades under Marathi instructions, dropping from 0.598 to 0.454. This asymmetry shows that models differ substantially in how well they can interpret task instructions written in the target language.

Model	Deva.	Roman.	Drop
Llama-3.1-8B	0.586	0.466	20.5%
Llama-3.3-70B	0.714	0.569	20.4%
Mistral-7B	0.684	0.564	17.5%
Qwen3-32B	0.823	0.758	7.9%

Table 3: Devanagari versus Romanized performance and percentage drop per model. Qwen3-32B is the most script-robust model.

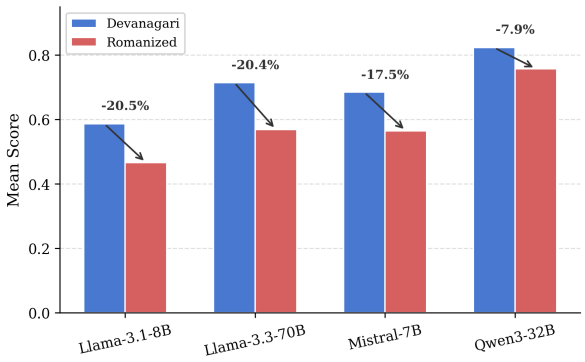

Figure 3: Performance drop from Devanagari to Romanized input per model. Qwen3-32B is most robust at 7.9%.

5.5 Syntax Feature Breakdown

Table 5 and Figure 5 present results by syntactic feature. SOV word order is handled best across all models. Qwen3-32B scores 0.981 and Llama-3.3-70B reaches 0.943 on this feature. Verb agreement is similarly strong. Vibhakti case markers tell a different story. Qwen3-32B leads at 0.880 but Mistral-7B scores only 0.467, barely above chance on a binary task. Vibhakti is the feature where model capability varies most. Postpositions are uniformly difficult, with scores between 0.709 and 0.889.

5.6 BERTScore Analysis

Table 6 compares token-level F1 and BERTScore on translation tasks. BERTScore is consistently higher than token F1 for all models. The gap is largest for Llama-3.1-8B, which scores 0.448 on token F1 but 0.921 on BERTScore. This means the model's translations are often correct paraphrases rather than exact matches, which token F1 unfairly penalizes. Script sensitivity persists in BERTScore too. Qwen3-32B scores 0.991 on Devanagari translations but 0.973 on Romanized ones. For Llama-3.1-8B the gap is wider: 0.930 versus 0.912. Even when measured semantically,

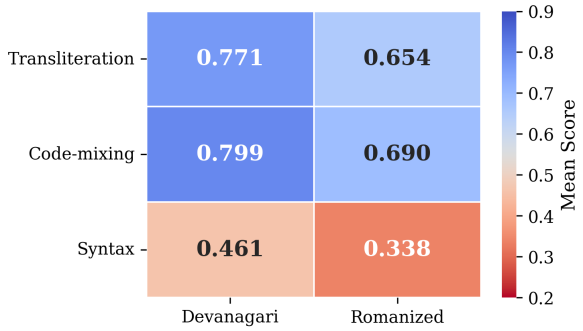

Figure 4: Mean scores by probe type and script form. Syntax on Romanized input is the hardest condition.

Instr.	L-8B	L-70B	M-7B	Q-32B
English	0.598	0.656	0.699	0.709
Marathi	0.454	0.623	0.550	0.873

Table 4: Performance by instruction language. Qwen3-32B benefits strongly from Marathi instructions while Llama-3.1-8B degrades under the same condition.

the Romanized penalty does not disappear.

6 Error Analysis

Manual inspection of zero-score outputs reveals five recurring failure patterns.

Hallucination under Romanization. A large share of outputs have zero lexical overlap with the ground truth, and 91.5% of these occur on Romanized input. Llama-3.1-8B translates *Mulga shaalet gela* (The boy went to school) as "The sun is setting." Mistral-7B renders *Baajaarat bhaaji milte* (Vegetables are available in the market) as "The treasure is found." Models appear to treat Romanized Marathi tokens as phonetically similar but semantically unrelated English words.

Hindi confusion. Llama-3.1-8B produces Hindi translations in several instances under Marathi instructions. It outputs *mujhe bhuk lagi hai* in Hindi instead of the English translation of "I am hungry." This occurs only under Marathi instruction conditions, pointing to the absence of a reliable internal boundary between Hindi and Marathi when the model is prompted in Devanagari.

Vibhakti marker substitution. All models prefer Hindi case markers over Marathi equivalents. For the dative case, models consistently choose *ko* (Hindi) over *la* (Marathi). This is not random error. It reflects a systematic default to Hindi morphological patterns for any Devanagari-script Indic input.

Feature	L-8B	L-70B	M-7B	Q-32B
SOV drift	0.831	0.943	0.820	0.981
Vibhakti	0.712	0.583	0.467	0.880
Verb agreement	0.860	0.727	0.706	0.750
Postposition	0.709	0.778	0.760	0.889

Table 5: Syntax probe performance by feature. SOV drift is easiest. Vibhakti shows the largest variance across models.

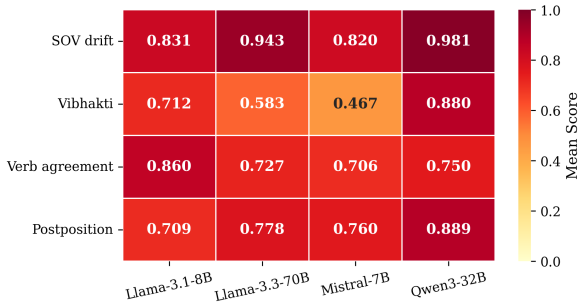

Figure 5: Syntax feature breakdown by model. SOV and verb agreement are best. Vibhakti varies most across models.

Negative subjunctive deletion. Every model fails to preserve negation in counterfactual constructions. The sentence *Jar paaus aala nasta tar aamhi baher gelo asto* (If it had not rained we would have gone outside) is rendered as affirmative by all four models. Mistral-7B produces "If the rain stops, we go out." The marker *nasta* disappears every time, across both surface forms and both instruction languages.

Subject and person shift. Multiple models change the grammatical person of the subject during translation. "Mother cooks food" becomes "I cook" across three models. "She likes mango" becomes "You like mangoes." When Marathi subject agreement morphology is unfamiliar, models default to first or second person, changing the meaning entirely.

7 Discussion

Script sensitivity is the most consistent finding. A drop of 8 to 21% for semantically identical input in a different script is not a minor artifact. Marathi speakers who type in Roman script on mobile devices encounter a systematically weaker model than those who type in Devanagari. The practical consequence is a two-tier experience divided by script choice alone.

BERTScore adds a useful nuance. Models produce translations that are semantically closer to

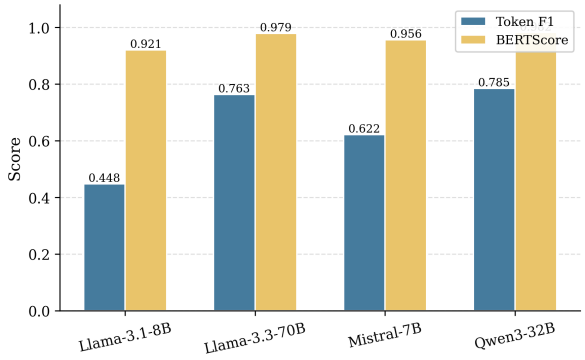

Figure 6: Token F1 versus BERTScore on translation tasks. BERTScore is consistently higher across all models.

Model	Token F1	BERTScore
Qwen3-32B	0.785	0.982
Llama-3.3-70B	0.763	0.979
Mistral-7B	0.622	0.956
Llama-3.1-8B	0.448	0.921

Table 6: Token F1 versus BERTScore on translation tasks. The gap is largest for Llama-3.1-8B, indicating correct paraphrases penalized by surface-level F1.

the reference than token F1 suggests. The large gap between token F1 and BERTScore for Llama-3.1-8B shows that this model often captures the right meaning but expresses it differently. That is a meaningfully different failure mode than hallucination.

The vibhakti substitution pattern goes beyond Marathi. When models default to Hindi case markers for all Devanagari-script input, they reveal that multilingual training has not produced separate internal representations for the two languages. Hindi and Marathi share a script but differ deeply in morphology. A model that cannot distinguish them at the morphological level will fail Marathi users on tasks that depend on case marking.

Qwen3-32B's strong response to Marathi instructions is the most practically interesting finding in the instruction language analysis. It scores 0.873 under Marathi instructions versus 0.709 under English, a gap of 16.4 points. No other model shows this. It suggests the model has internalized enough Marathi to parse task-level instructions in the language. Whether this transfers to other morphologically rich Indic languages is an open question.

Limitations

We report both token-level F1 and BERTScore for translation tasks following reviewer feedback. Our translation metric is token-level F1, which does not capture paraphrase equivalence. A correct translation with different wording may receive a low score. Our dataset is manually curated rather than crowd-sourced, which limits scale while improving linguistic precision. Evaluation of frontier closed models such as GPT-4 and Gemini is left for future work due to API cost constraints.

Conclusion

We presented a probing study of four multilingual LLMs on Marathi across script variation, code-mixing, and four syntactic features. The failure patterns are not random. Models lose between 8 and 21% performance when Marathi switches from Devanagari to Roman script. The negative subjunctive marker *nasta* is ignored by every model. Vibhakti case markers are replaced with Hindi equivalents across the board. BERTScore reveals that some failures are paraphrase mismatches rather than semantic errors, but the script sensitivity and morphological gaps remain real and practically significant. Future work should examine the effect of dedicated Marathi pretraining data on vibhakti and postposition handling. Building larger Romanized Marathi corpora for instruction tuning is a natural next step.

Acknowledgments

We thank the native Marathi speakers who verified the dataset sentences. All experiments were conducted using free inference APIs from Groq and Mistral.

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages

4232–4267, Singapore. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

BigScience Workshop. 2022. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M. Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for Indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Asang Dani and Shailesh R. Sathe. 2024. A review of the Marathi natural language processing. *arXiv preprint arXiv:2412.15471*.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages. *arXiv preprint arXiv:2305.16307*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Raviraj Joshi. 2022. L3Cube-MahaCorpus and MahaBERT: Marathi monolingual corpus, Marathi BERT language models, and resources. *arXiv preprint arXiv:2202.01159*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. AI4Bharat-IndicNLP corpus: Monolingual corpora and word embeddings for Indic languages. *arXiv preprint arXiv:2005.00085*.

Pawan Lahoti, Namita Mittal, and Girdhari Singh. 2022. A survey on NLP resources, tools, and techniques for Marathi language processing. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2).

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv preprint arXiv:2307.16039*.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mbert is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 448462, Online. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar,

Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6984–6996. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. *CoRR*, abs/2109.07684.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A Dataset Statistics and Examples

Table 7 shows the dataset distribution. Tables 8–10 give representative examples from each probe type.

Probe	Difficulty	N
Transliteration	Easy	107
	Medium	90
	Hard	170
Code-mixing	Easy	90
	Medium	107
	Hard	110
Syntax	Easy	105
	Medium	126
	Hard	115
Total		**1,020**

Table 7: Dataset distribution. Full dataset = 1,020 sentences.

Task	Romanized Input	Ground Truth
Translation	*Majhe naav Raam aahe.*	My name is Ram.
Translation	*Aaj khup ukaada aahe.*	It is very hot today.
Similarity	*To shaalet gela.*	YES
Translation	*Jar paaus aala nasta tar aamhi baher gelo asto.*	If it had not rained, we would have gone outside.

Table 8: Transliteration examples. Row 4 tests *nasta*, which all models fail to preserve.

Task	Code-Mixed Input	Ground Truth
Translation	*Majha **project** aaj **submit** karaaycha aahe.*	I have to submit my project today.
Translation	*Aamhi **meeting** madhye **presentation** dili.*	We gave a presentation in the meeting.
Similarity	*Tyaane naveen **phone** **buy** kela.*	YES
Translation	***Traffic** muLe mala **office** la usheer zhaala.*	I was late to the office because of traffic.

Table 9: Code-mixing examples. English words in bold.

Feature	Correct	Distractor	T
SOV	*Raam shaalet jaato.* (SOV)	*Raam jaato shaalet.* (SVO)	Gr
Vibhakti	*Tyaa-**la** bhook laagali.* (*la*=Marathi)	*Tyaa-**ko** bhook laagali.* (*ko*=Hindi)	CM
Verb agr.	*Ti shaalet **geli**.* (fem.)	*Ti shaalet **gela**.* (masc.)	Gr
Postpos.	*Pustak **teblavar** aahe.* (postpos.)	*Pustak **on table** aahe.* (Eng. order)	Gr

Table 10: Syntax examples. T = task type. Gr = grammaticality judgement. CM = case marker selection.

All 1,020 sentences were written and verified by native Marathi speakers with at least graduate-level education. Difficulty annotation was done by two independent annotators with a Cohen's kappa of 0.76, indicating substantial agreement. Romanization follows conventions observed in real Marathi social media text and does not conform to any formal transliteration standard, reflecting actual user behavior on mobile platforms. Code-mixing patterns were drawn from genuine Marathi-English usage across technology, workplace, and social media domains in urban Maharashtra, India.