

Beyond Surface Form: Probing Script and Discourse-Marker Sensitivity in Marathi LLM Reasoning

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 expanded rapidly, yet aggregate benchmark scores often conceal how performance shifts once the surface form of a language changes. We examine this question through Marathi discourse relation classification. Five instruction-following models, Llama-3.1-8B, Llama-3.3-70B, Qwen3.6-27B, GPT-OSS-20B, and GPT-OSS-120B, are evaluated on a manually curated Marathi language dataset spanning six discourse relations, two scripts, eighteen domains, and three difficulty levels, producing 2500 model-level predictions. Qwen3.6-27B obtains the highest overall accuracy at 82.4%, followed by GPT-OSS-120B at 80.8%. The central finding is a large script effect as mean accuracy falls from 78.1% on Devanagari script to 49.7% on Romanized Marathi, a decline of 28.5 percentage for content that is otherwise unchanged. Relation-level results trace to marker form rather than presence; every instance carries an explicit marker, but Temporal and Contrastive relations use a single canonical connective and are handled reliably, while Inferential and Reformulative relations lean on multi-word phrasal markers and are frequently confused with Causal, a 20.4-point accuracy gap that holds across all five models. Llama-3.1-8B shows a more severe failure, defaulting to Contrastive across most gold relations. These results indicate that strong multilingual performance on Marathi is neither script-invariant nor uniform across discourse relations.

1 Introduction

Discourse understanding requires more than recognizing the words in a sentence. A language model needs to know how one proposition relates to another, whether the second adds information, states a cause, marks a contrast, follows an earlier event in time, supports an inference, or restates something already said. This is the task of classifying

discourse relationships, which has been studied for decades through resources such as the Penn Discourse TreeBank and the shared multilingual DIS-RPT tasks (Prasad et al., 2008; Webber et al., 2019; Braud et al., 2023). Results continue to vary across languages, domains, and relation inventories, and the variation is sharpest for relations that carry no explicit connective (Kim et al., 2020; Metheniti et al., 2024).

Marathi is a useful setting in order to examine this variation. It is spoken by more than 83 million people and has gained growing NLP support through IndicNLPsuite, Samanantar, and the L3Cube Marathi resources (Kakwani et al., 2020; Ramesh et al., 2022; Joshi, 2022). It nonetheless remains underrepresented in multilingual evaluation relative to high-resource languages. Earlier similar studies examined Marathi script variation, code-mixing, and syntax at the sentence level and found substantial degradation on Romanized input alongside systematic confusion between Marathi and Hindi morphology (Patil and Chetia, 2026). The present study asks whether a comparable sensitivity appears once the task moves from sentence-level probing to discourse-level classification, where a model must reason about two propositions rather than judge a single sentence in isolation.

Script proves to be a crucial site of probe here as Marathi is conventionally written in Devanagari; however, just as often in Roman script during informal digital communication. The linguistic content can stay constant while the token sequence a model changes considerably. Multilingual models are known to transfer unevenly across languages and scripts (Conneau et al., 2020; Muller et al., 2021), and recent work shows that script match can affect knowledge transfer even between closely related languages (Bandarkar et al., 2026). These findings motivate a controlled comparison between Devanagari and Romanized Marathi under an iden-

tical relation inventory, so that script is the only variable that changes.

We organize the study around four questions.

- **RQ1:** How well do multilingual LLMs classify Marathi discourse relations overall?
- **RQ2:** How much does performance change when the same content is written in Romanized Marathi?
- **RQ3:** Which relations and difficulty levels are most challenging?
- **RQ4:** Do model errors show systematic relation confusions?

This study contributes a controlled evaluation rather than a new discourse parser. The results show a 28.5-point average script gap and a recurring confusion between Inferential, Reformulative, and Causal relations. The analysis also shows why a single aggregate accuracy figure cannot describe Marathi discourse competence on its own: relations whose markers are single canonical connectives behave very differently from relations whose markers are multi-word phrasal expressions, a distinction we return to in Section 3.3.

2 Related Work

2.1 Discourse Relation Classification

The Penn Discourse TreeBank established a widely used framework for explicit and implicit discourse relations (Prasad et al., 2008), later expanded and refined in PDTB 3.0 (Webber et al., 2019). Implicit relations remain difficult because the relation must be inferred from the two arguments rather than read off an overt connective (Kim et al., 2020). DISRPT provides a common multilingual evaluation setting across discourse frameworks (Braud et al., 2023, 2024). Relation classification remains sensitive to language, domain, and framework choice (Metheniti et al., 2024; Bourgonje and Demberg, 2024), and LLM-based probing finds that semantically adjacent relations account for most model errors (Eichin et al., 2025).

2.2 Multilingual and Indic NLP

XLM-R demonstrated the value of large-scale multilingual pretraining for cross-lingual transfer (Conneau et al., 2020). Subsequent probing work showed that linguistic information is not encoded uniformly across languages, and that transfer depends on a language’s representation in the pretraining mixture (Tenney et al., 2019;

Chi et al., 2020; Muller et al., 2021). Indian-language NLP has developed rapidly through IndicNLPsuite, Samanantar, IndicBART, and MahaBERT (Kakwani et al., 2020; Ramesh et al., 2022; Dabre et al., 2022; Joshi, 2022). Broad multilingual evaluations continue to report gaps for lower-resource languages (Ahuja et al., 2023; Bang et al., 2023). Patil and Chetia (2026) provide a closely related Marathi evaluation at the sentence level; the present study extends that line of work to discourse relations. Recent work highlights that multilingual model performance can vary depending on how information is represented and contextualized across languages. Ranaldi et al. (Ranaldi et al., 2026) examine cross-lingual retrieval and representation, while Ranaldi and Pucci (Ranaldi and Pucci, 2026) show that multilingual model behavior can be sensitive to contextual signals. These findings complement our investigation of script and discourse-marker sensitivity in Marathi.

2.3 Script Variation

Code-switched benchmarks such as GLUECoS and LinCE show that models struggle when language identity and surface form mix (Khanuja et al., 2020; Aguilar et al., 2020). Translation benchmarks such as FLORES-200 widened language coverage but retain conventional scripts (NLLB Team et al., 2022), leaving Romanized Marathi poorly represented. We treat script as an experimental variable and hold content constant across conditions.

3 Dataset Construction

We construct a manually curated dataset of 500 Marathi discourse instances. Each instance is assigned one of six relations: Additive, Causal, Contrastive, Temporal, Inferential, or Reformulative. The dataset spans both Devanagari and Romanized Marathi, eighteen topical domains, and Easy, Medium, and Hard difficulty labels. All instances were written and checked by native Marathi speakers, and no machine translation was used during construction.

3.1 Relation Inventory

The six labels capture broad discourse functions. Additive connects compatible information. Causal links a cause with its consequence. Contrastive places two propositions in opposition. Temporal relates events through time. Inferential presents

a conclusion supported by the preceding proposition. Reformulative restates or reframes information already given. This inventory is intentionally smaller than PDTB or RST so that the evaluation stays consistent across all five models.

3.2 Experimental Conditions

The dataset is balanced across the two script conditions at the content level. Romanization follows conventions used in informal Marathi digital communication rather than a formal transliteration standard. Each instance also receives a domain label and a difficulty level based on the amount of contextual information required to identify the intended relation. This design allows the same relation inventory to be examined under different surface and contextual conditions without changing the underlying task.

3.3 Discourse Markers

A discourse marker is a lexical or phrasal cue, typically a connective, that signals how two propositions relate to each other. English realizes this with words such as *so*, *but*, and *meanwhile*. Marathi realizes the same function through markers such as म्हणून and परिणामी for Causal relations, पण and उलट for Contrastive relations, त्यामुळे for Inferential relations, and शेवटी or त्यानंतर for Temporal relations. Every instance in the dataset is annotated with its governing marker, where one is present, and with the marker’s position in the sentence, whether clause-initial, medial, or final.

Every instance in the dataset carries an explicit governing marker; construction never leaves a relation unmarked. What varies systematically is the *form* the marker takes. Causal, Contrastive, and Temporal relations are always signaled by a single canonical connective, such as त्यामुळे, परंतु, or शेवटी. Additive, Inferential, and Reformulative relations draw more often on multi-word phrasal expressions, such as दुसऱ्या शब्दांत (“in other words”) or स्पष्टच आहे की (“it is clear that”). Of the 500 evaluated instances, 437 use a single-word marker and 63 use a phrasal marker; Causal, Contrastive, and Temporal contain no phrasal markers at all.

Averaged across all five models and 2,500 predictions, accuracy on single-word-marker instances is 73.4%, against 53.0% on phrasal-marker instances, a 20.4-point gap that holds for every model, from Qwen3.6-27B (83.3% vs. 76.2%) to Llama-3.1-8B, which collapses from 46.0% to 11.1%. A single canonical connective is a com-

Model	Single-word	Phrasal
Qwen3.6-27B	.833	.762
GPT-OSS-120B	.833	.635
Llama-3.3-70B	.787	.556
GPT-OSS-20B	.757	.587
Llama-3.1-8B	.460	.111

Table 1: Accuracy by marker form (437 single-word vs. 63 phrasal instances). The gap holds for every model.

pact, familiar cue; a multi-word phrase does not compress the same way, and a model that fails to resolve it appears to default to the relation it has seen signaled most often, which is consistently Causal (Section 6.1).

4 Experiments

4.1 Models

We evaluate five instruction-following models: Llama-3.1-8B, Llama-3.3-70B, Qwen3.6-27B, GPT-OSS-20B, and GPT-OSS-120B. Llama provides small and large open-weight baselines (Grattafiori et al., 2024); Qwen3 and GPT-OSS add families with different training and instruction-tuning choices (Yang et al., 2025; OpenAI, 2025). No model is fine-tuned on the evaluation instances.

All prompts require a single relation label from the six-class inventory. Generation settings are fixed across models. Raw responses are normalized to the permitted labels only when the intended label is unambiguous; otherwise the prediction is counted as incorrect.

4.2 Evaluation Metrics

We report accuracy, macro F1, and weighted F1. Accuracy measures the proportion of correct predictions. Macro F1 weights each relation equally and therefore makes relation collapse visible even when a model achieves reasonable overall accuracy. We also report performance broken down by script, relation, difficulty, and domain. The script gap is defined as

$$\Delta_{\text{script}} = \text{Acc}_{\text{Devanagari}} - \text{Acc}_{\text{Romanized}}$$

5 Results

5.1 Overall Performance

Qwen3.6-27B achieves the highest overall accuracy at 82.4%, followed by GPT-OSS-120B at 80.8%. Llama-3.3-70B and GPT-OSS-20B reach 75.8% and 73.6%, while Llama-3.1-8B reaches 41.6%. The ordering does not follow parameter count alone: the 27B Qwen model exceeds the

Model	Accuracy	Macro F1	Weighted F1
Qwen3.6-27B	.824	.822	.822
GPT-OSS-120B	.808	.799	.799
Llama-3.3-70B	.758	.728	.729
GPT-OSS-20B	.736	.716	.716
Llama-3.1-8B	.416	.375	.375

Table 2: Overall Marathi discourse relation classification performance.

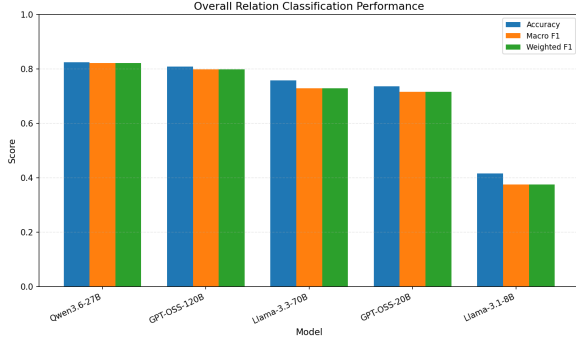


Figure 1: Overall performance across the five models.

70B Llama model, while the 20B GPT-OSS model stays close to that same larger model.

5.2 Performance by Relation

The relation-level results show a consistent separation between easier and harder relations. Qwen3.6-27B reaches 0.95 F1 on Temporal and 0.94 on Contrastive, while Inferential and Reformulative fall to 0.69 and 0.66. GPT-OSS-120B shows the same pattern, with 0.96 and 0.95 on Temporal and Contrastive but only 0.62 on both Inferential and Reformulative. The weaker models follow the same ordering. These results indicate that difficulty is tied to the relation distinction itself, and in particular to whether a relation’s marker is a single canonical connective or a multi-word phrasal expression (Table 1), rather than to any single model’s weaknesses.

5.3 Script Variation

Script is the strongest source of variation in the evaluation. The mean accuracy falls from 78.1% on Devanagari to 49.7% on Romanized Marathi. Qwen3.6-27B drops from 89.8% to 60.9%, while GPT-OSS-120B drops from 85.8% to 66.4%. Llama-3.3-70B and GPT-OSS-20B show declines of 30.5 and 28.6 points, respectively. Llama-3.1-8B falls from 50.5% to 15.6%.

The result is notable because the underlying discourse content is held constant. A model is therefore not being asked to solve a different linguistic problem. The change is primarily orthographic,

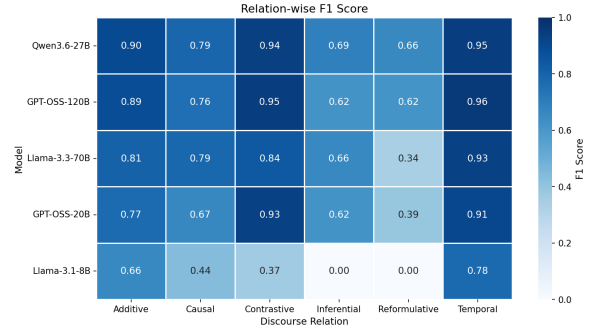


Figure 2: F1 by discourse relation. Temporal and Contrastive, both signaled by single canonical connectives, are consistently easier than Inferential and Reformulative.

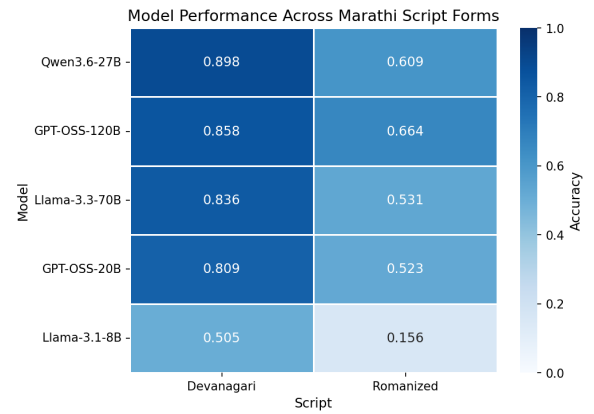


Figure 3: Accuracy on Devanagari and Romanized Marathi. All models show a substantial script gap.

yet it produces a large difference in classification accuracy.

5.4 Difficulty and Domain

Performance decreases as difficulty increases for all five models. Qwen3.6-27B falls from about 0.93 on Easy to 0.69 on Hard, GPT-OSS-120B from about 0.93 to 0.72, and Llama-3.1-8B shows the largest decline, from 0.61 to 0.27. The difficulty labels capture meaningful differences in how hard a relation is to identify.

Domain performance is also uneven. The strongest models stay relatively stable across most domains but drop sharply in a few. Qwen3.6-27B reaches 1.00 in Technology and 0.93 in Culture but only 0.57 in Environment; GPT-OSS-120B ranges from 0.57 in Environment to 0.95 in Social Work. Llama-3.1-8B is more variable still, reaching 0.59 in Academic but falling to 0.21 in Agriculture, consistent with known domain effects in discourse classification (Bourgonje and Demberg, 2024).

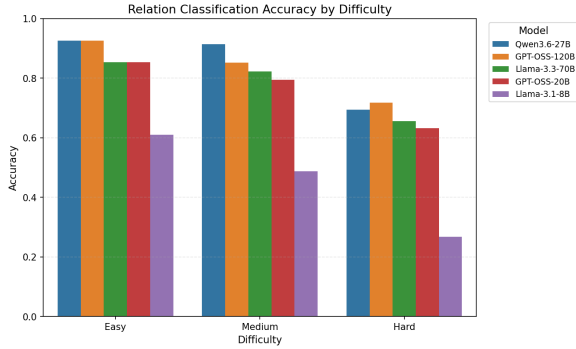


Figure 4: Accuracy by difficulty. Performance declines as the relation becomes harder to identify.

6 Error Analysis

6.1 Causal as an Attractor

The GPT-OSS-120B confusion matrix shows the clearest recurring error. Among Inferential instances, 27 are predicted as Causal and 12 as Reformulative. Among Reformulative instances, 23 are predicted as Causal. Llama-3.3-70B and GPT-OSS-20B show the same direction of confusion, though at different rates. The models therefore recognize that the two propositions are related but often choose a causal interpretation instead of the intended inferential or reformulative one. This is the failure the marker-form asymmetry in Section 3.3 predicts: Inferential and Reformulative instances lean on multi-word phrasal markers that are harder to resolve than the single canonical connectives used for Causal, so a model that fails to resolve the cue falls back on the relation it has seen signaled most often.

6.2 Relation Collapse in Llama-3.1-8B

Llama-3.1-8B exhibits a more severe failure. It predicts Contrastive for 39 Additive instances, 56 Causal instances, 78 Inferential instances, 79 Reformulative instances, and 25 Temporal instances, effectively collapsing five of the six gold relations onto one label. Its macro F1 of 0.375 captures this behavior more clearly than its accuracy of 0.416 does.

The remaining models show intermediate behavior. Llama-3.3-70B performs well on Additive, Causal, Contrastive, and Temporal but records only 0.34 F1 on Reformulative. GPT-OSS-20B similarly maps many Inferential and Reformulative instances to Causal while retaining strong Contrastive and Temporal performance. The errors are structured rather than random across every model in the study.

7 Discussion

Three findings stand out. First, multilingual competence does not imply script invariance. The average 28.5-point gap between Devanagari and Romanized Marathi is too large to treat as ordinary variation. This result extends the script sensitivity observed in our earlier Marathi probing study (Patil and Chetia, 2026) from sentence-level tasks to discourse relation classification.

Second, relation difficulty is structured, and the structure tracks the form of the discourse marker rather than its presence: every instance in the dataset carries an explicit marker. Temporal and Contrastive relations are consistently easier because their markers are always single canonical connectives, while Inferential and Reformulative relations remain difficult even for the strongest models because their markers are more often multi-word phrasal expressions. This is compatible with earlier work showing that implicit and fine-grained discourse relations are harder to classify (Kim et al., 2020; Metheniti et al., 2024; Eichin et al., 2025), though here the difficulty traces to marker form rather than marker absence. The pattern also shows that high overall accuracy can coexist with weak performance on specific relations, particularly the ones a benchmark average would never surface.

Furthermore, parameter count does not explain the ranking. Qwen3.6-27B is the strongest model, the 70B Llama model finishes third, and the 120B GPT-OSS model finishes second. Architecture, training data, tokenizer behavior, and the amount of Marathi seen during pretraining may matter alongside scale. The result argues for evaluations that report relation-level and script-level performance rather than a single benchmark average.

The findings carry a direct implication for Marathi NLP. Evaluation should include Romanized input because it is a common form of informal written Marathi. A model that performs well only on clean Devanagari text does not offer the same level of discourse understanding to users who type in Roman script. The same principle applies to relation inventories: a single accuracy number cannot show whether a model preserves distinctions among semantically adjacent discourse functions.

Limitations

The study is restricted to relation classification and does not evaluate translation, generation, or

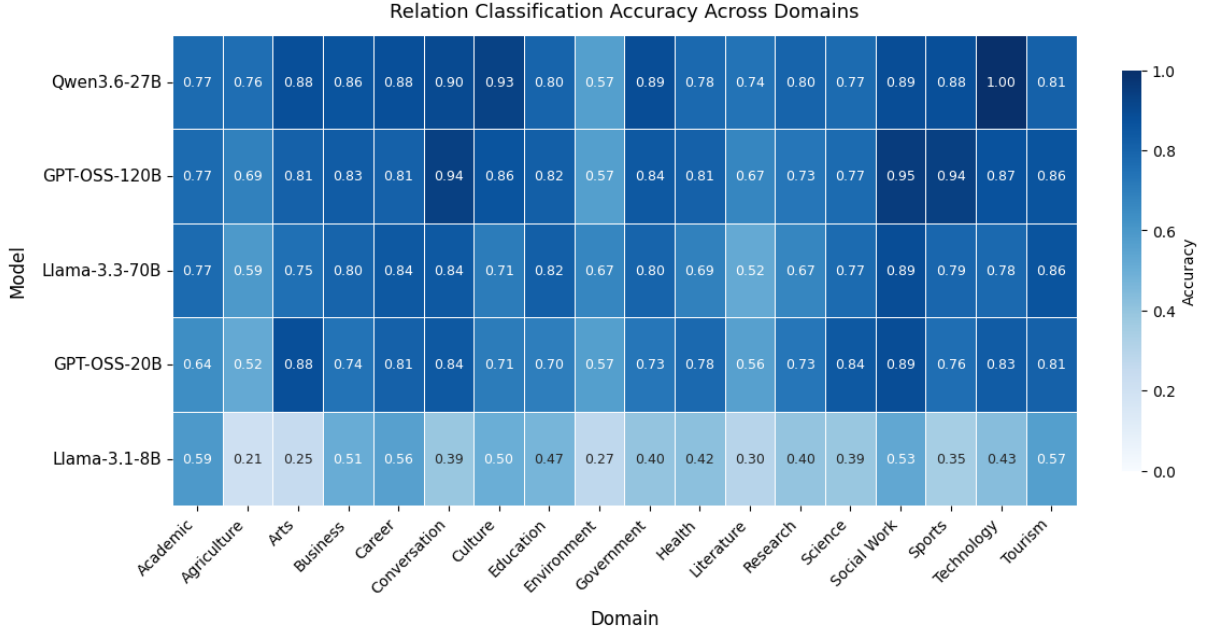


Figure 5: Accuracy across eighteen domains. Domain variation remains visible even among the stronger models.

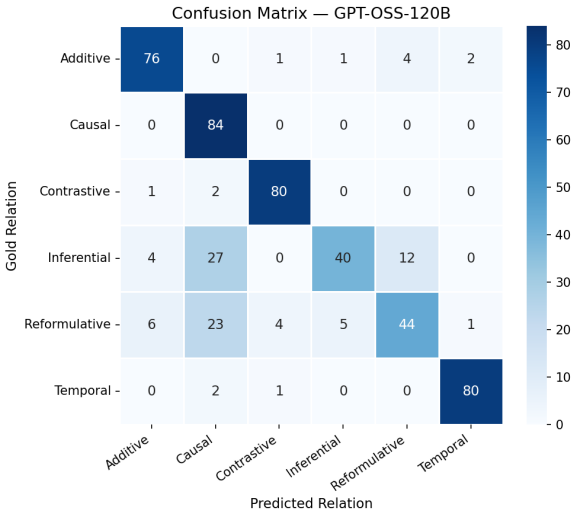


Figure 6: Confusion matrix for GPT-OSS-120B. Causal attracts a substantial share of Inferential and Reformulative errors.

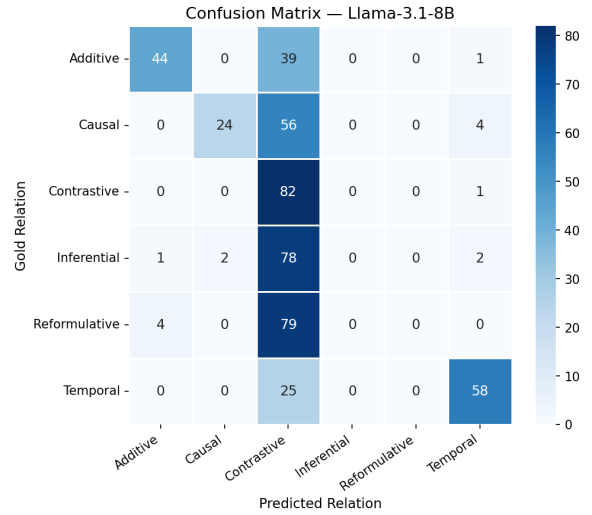


Figure 7: Confusion matrix for Llama-3.1-8B. Contrastive is over-predicted across most gold relations.

full discourse parsing. The six-label inventory is coarser than PDTB, RST, or discourse dependency schemes. The dataset contains 500 instances, so domain-level cells are small once partitioned across eighteen domains. Five open-weight models are evaluated; different prompts or decoding settings could shift absolute scores. A natural extension is testing whether Romanized Marathi instruction tuning narrows the script gap without degrading Devanagari performance. The reported comparisons should be read as a controlled probe rather than a complete benchmark of Marathi discourse competence.

Conclusion

We evaluated five multilingual LLMs on Marathi discourse relation classification under controlled variation in script, relation type, difficulty, and domain. Qwen3.6-27B achieves the highest overall accuracy at 82.4%, however crucial result is the script gap. The mean accuracy falls from 78.1% on Devanagari to 49.7% on Romanized Marathi. Inferential and Reformulative relations are also consistently harder than Temporal and Contrastive relations, with Causal acting as a frequent attractor in model errors, a pattern traceable not to marker presence, every instance carries one, but to

marker form: single canonical connective versus multi-word phrase. Llama-3.1-8B shows an even stronger collapse toward Contrastive predictions. Future work should enlarge the dataset, test additional Indic languages, and examine whether Romanized Marathi training data can improve script robustness without reducing performance on Devanagari input.

Acknowledgments

We thank the native Marathi speakers who wrote and verified the dataset instances. We also acknowledge the computational resources used to obtain the model predictions.

References

- Gustavo Aguilar, Sudipta Kar, and Tamar 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.
- Lucas Bandarkar, Alan Ansell, and Trevor Cohn. 2026. [Large reasoning models struggle to transfer parametric knowledge across scripts](#). *arXiv preprint arXiv:2603.17070*.
- 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](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*.
- Peter Bourgonje and Vera Demberg. 2024. [Generalizing across languages and domains for discourse relation classification](#). In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 554–565, Kyoto, Japan. Association for Computational Linguistics.
- Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. [The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification](#). In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. Association for Computational Linguistics.
- Chloé Braud, Amir Zeldes, Laura Rivière, Yang Janet Liu, Philippe Muller, Damien Sileo, and Tatsuya Aoyama. 2024. [DISRPT: A multilingual, multi-domain, cross-framework benchmark for discourse processing](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4990–5005, Torino, Italia. ELRA and ICCL.
- 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. 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. Association for Computational Linguistics.
- Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh 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.
- Florian Eichin, Yang Janet Liu, Barbara Plank, and Michael A. Hedderich. 2025. [Probing LLMs for multilingual discourse generalization through a unified label set](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18665–18684, Vienna, Austria. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. 2024. [The Llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Raviraj Joshi. 2022. [L3Cube-MahaCorpus and MahaBERT: Marathi monolingual corpus, Marathi BERT language models, and resources](#). In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 97–101, Marseille, France. European Language Resources Association.
- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, N. C. Gokul, Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. [IndicNLP Suite: 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. Association for Computational Linguistics.
- Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. [GLUE-CoS: An evaluation benchmark for code-switched NLP](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585. Association for Computational Linguistics.
- Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Las-tras. 2020. [Implicit discourse relation classification: We need to talk about evaluation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414. Association for Computational Linguistics.

Eleni Metheniti, Chloé Braud, and Philippe Muller. 2024. [Feature-augmented model for multilingual discourse relation classification](#). In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 91–104, St. Julians, Malta. Association for Computational Linguistics.

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*, pages 448–462. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *arXiv preprint arXiv:2207.04672*.

OpenAI. 2025. [gpt-oss-120b and gpt-oss-20b model card](#). *arXiv preprint arXiv:2508.10925*.

Tejas Patil and Barnali Chetia. 2026. [Beyond accuracy: A structured error analysis of multilingual llms on marathi script variation and syntax](#). In *Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026)*, pages 119–126, San Diego, United States. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. [The Penn Discourse TreeBank 2.0](#). In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan A K, 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.

Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. 2026. [Multilingual retrieval-augmented generation for knowledge-intensive question answering task](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 697–716, Rabat, Morocco. Association for Computational Linguistics.

Leonardo Ranaldi and Giulia Pucci. 2026. [Learning multilingual agentic policy to control sycophancy](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3664–3681, Rabat, Morocco. 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. Association for Computational Linguistics.

Relation	Devanagari example	Interpretation
Additive	तो दररोज मराठी वाचतो आणि तो नियमितपणे लिहितो.	Adds compatible information.
Causal	जोरदार पाऊस पडला, म्हणून सामना रद्द झाला.	Gives the cause of an event.
Contrastive	तो बैठकीला आला, पण त्याने कोणतेही मत मांडले नाही.	Marks a contrast.
Temporal	जेवण झाल्यावर आम्ही बाहेर फिरायला गेलो.	Relates events in time.
Inferential	रस्ता ओला आहे, त्यामुळे रात्री पाऊस पडला असावा.	Supports an inference.
Reformulative	त्याने काम पूर्ण केले नाही, म्हणजेच ते अजून बाकी आहे.	Restates or reframes information.

Table 3: Representative Devanagari examples for the six relation labels.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. [The Penn Discourse Treebank 3.0 annotation manual](#). Technical report, University of Pennsylvania.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.

A Dataset Statistics and Examples

The Causal and Inferential examples in Table 3 illustrate the relation boundary behind most recurring Causal errors. The causal example states that rain caused the match to be canceled, while the inferential example reasons backward from a wet road to a likely earlier rain. Both use resultative connectives, but the direction of reasoning differs, one moves forward from cause to effect, the other moves backward from an observed state to a probable earlier cause. This is precisely the boundary that GPT-OSS-120B and the other mid-sized models blur most often, since both relations can share a surface-level resultative marker while requiring opposite inferential direction. Romanized Marathi is treated as a surface variant, not a separate condition. Romanization follows informal online usage, not a formal transliteration standard, to measure robustness to realistic input rather than to a standardized orthography that few Marathi speakers actually type in practice.

Model	Add.	Caus.	Contr.	Infer.	Reform.	Temp.
Qwen3.6-27B	.90	.79	.94	.69	.66	.95
GPT-OSS-120B	.89	.76	.95	.62	.62	.96
Llama-3.3-70B	.81	.79	.84	.66	.34	.93
GPT-OSS-20B	.77	.67	.93	.62	.39	.91
Llama-3.1-8B	.66	.44	.37	.00	.00	.78

Table 4: Relation-wise F1 scores.