Diagnosing Representational Conflation in Multilingual LLMs: Evidence from Bhojpuri–Hindi

Anonymous ACL submission

Abstract

Large multilingual language models (LLMs) often perform unevenly across closely related languages, a problem usually ascribed to data scarcity or capacity limits. We introduce a diagnostic framework to measure *representational conflation*, the tendency of a model to treat a low-resource language as if it were a closely related high-resource neighbor. The framework is behavioral-first and designed to be applicable to closed-box systems: it operationalizes five complementary diagnostics (minimal-pair discrimination, lexical authenticity, morphosyntactic patterning, cultural-semantic preservation, and cross-linguistic priming) and provides concrete prompts, rubrics, and statistical guidance. Applying the framework to Bhojpuri–Hindi using the BHLTR corpus and five contemporary LLMs yields convergent signals of conflation: discrimination accuracy drops 23–44 percentage points below high-resource baselines, 29–72% of generations show Hindi lexical influence, 87% of morphological errors follow valid Hindi templates, and Hindi priming causes 8–9 percentage point performance degradation. These patterns suggest systematic representation collapse rather than mere knowledge gaps. We release evaluation templates, prompts, and analysis code to support replication and mechanistic follow-ups.

1 Introduction

Large multilingual language models are trained on massively imbalanced corpora where a few high-resource languages dominate. While this training paradigm enables impressive cross-lingual transfer, it creates significant risks for the approximately 1 billion speakers of low-resource languages that are closely related to high-resource neighbors. For these language pairs with severe resource asymmetry, we hypothesize that models may develop *conflated representations*: systematically mapping inputs in a low-resource language onto processing pathways privileged for a high-resource relative.

Unlike simple performance degradation, conflation threatens linguistic diversity by imposing majority-language patterns onto minority languages. For the 51 million speakers of Bhojpuri, a language closely related to Hindi with 40–60% lexical overlap but approximately 1000:1 digital resource imbalance, this conflation can accelerate language shift by normalizing Hindi forms in contexts where Bhojpuri-specific vocabulary, morphology, and cultural frames should be preserved. When language technologies systematically substitute Hindi lexical items, morphological patterns, and cultural perspectives in place of authentic Bhojpuri alternatives, they risk reinforcing the very dynamics that threaten linguistic vitality.

This paper presents a practical diagnostic toolbox that detects and quantifies representational conflation in multilingual LLMs through behavioral measurement. The framework is designed to be applicable to both closed-API systems and open models, providing a triage layer that can identify conflation patterns and motivate deeper mechanistic investigations. We demonstrate the toolbox on Bhojpuri–Hindi using the BHLTR corpus and evaluate five contemporary models spanning closed APIs (Claude 3.5 Sonnet, GPT-4o, Gemini 1.5 Pro) and open research models (DeepSeek V3, LLaMA 3.1 8B).

Our goals are methodological: to provide a tight, replicable measurement protocol with concrete prompts, rubrics, and statistical procedures; to produce converging evidence across multiple diagnostic dimensions; and to establish a prioritized roadmap for mechanistic analysis and mitigation strategies.

Contributions. We present four primary contributions. First, we provide a documented five-test diagnostic framework with explicit prompts, scoring rubrics, and statistical recipes including proper multiple comparison corrections. Second, we present

an empirical pilot application to Bhojpuri–Hindi demonstrating convergent evidence of representational conflation across all evaluated models. Third, we provide detailed error analysis revealing that conflation is systematic rather than random, with errors following valid Hindi templates at rates far exceeding chance. Fourth, we provide publication-quality figures, reproducible appendix materials with full statistical tests, and an explicit plan for mechanistic follow-ups and mitigation experiments.

2 Related work

Our work connects to three research strands. Early multilingual BERT investigations revealed cross-lingual transfer capabilities but showed transfer depends on typological similarity and resource availability (Pires et al., 2019; Conneau et al., 2020). Negative interference, where adding languages degrades per-language performance, has been documented as the "curse of multilinguality" (Wang et al., 2020), with recent mitigation through adapters and mixture-of-expert routing (Blevins et al., 2024). Tokenization introduces systematic bias against low-resource languages through over-segmentation (Petrov et al., 2023), addressed by length-aware BPE and lexically grounded segmentation (Lee, 2024; Libovický and Helcl, 2024). Representational analysis methods include structural probes (Hewitt and Manning, 2019), behavioral probes (Tenney et al., 2019), RSA (Kriegeskorte et al., 2008), and amnesic probing (Elazar et al., 2021). Our behavioral-first diagnostic layer complements these mechanistic approaches for closed and open systems.

3 Framing: measurement vs mechanistic claims

A critical methodological distinction underlies our framework. Behavioral patterns, such as a model consistently substituting Hindi lexical items when prompted for Bhojpuri, provide evidence that representations are functionally conflated at the input-output level. However, such patterns do not by themselves localize where or how conflation is encoded within the model architecture. A model might conflate languages through shared vocabulary embeddings, through overlapping hidden representations in middle layers, through language-identification failures at the input stage, or through multiple simultaneous mechanisms.

Our framework therefore explicitly separates two complementary investigative layers. Layer A consists of *measurement*: a battery of reproducible behavioral diagnostics designed to detect and quantify conflation-like patterns through systematic input-output observations. Layer B consists of *explanation*: mechanistic probing, representational similarity analysis, and causal interventions that locate the architectural substrates responsible for observed behaviors.

We design the diagnostic layer to be low-cost, high-signal, and suitable as a triage mechanism. When diagnostics reveal strong conflation signals, they motivate targeted mechanistic investigations in open models. When diagnostics reveal minimal conflation, researchers can redirect resources elsewhere. This separation allows our framework to be productively applied to both closed systems (where only behavioral measurement is possible) and open systems (where both layers can be deployed).

4 Dataset and annotation

BHLTR resource. We use the Bhojpuri Language Technological Resources (BHLTR) collection as our primary dataset (Ojha, 2019; Ojha and Zeman, 2020; AI4Bharat, 2024). The BHLTR corpus is publicly available under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License. BHLTR is a publicly referenced resource that includes multiple subsets suitable for different diagnostic purposes: (i) a Universal Dependencies-style annotated treebank with part-of-speech tags and morphological annotations covering several thousand sentences, which we use for morphosyntactic tests; (ii) monolingual Bhojpuri corpora and English–Bhojpuri parallel texts spanning tens of thousands of segments, drawn from books, magazines, and web sources, which we use for generation and cultural content prompts. The corpus covers diverse domains including literature, news, and cultural descriptions, providing rich material for our diagnostic tests. Exact file names, version numbers, and segment counts are provided in the supplementary materials to support replication.

Annotator recruitment and training. We recruited eight native Bhojpuri speakers from Bihar and Uttar Pradesh regions where Bhojpuri is spoken as a first language. All annotators possessed native fluency in both Bhojpuri and Hindi, had completed undergraduate or higher education, and had

resided there within the past 5 years. Five annotators served as item creators while three served as independent evaluators for inter-annotator reliability measurement.

Annotators participated in structured training covering three core competencies. First, they learned to distinguish Hindi loanwords that have been borrowed into colloquial Bhojpuri usage from authentic Bhojpuri vocabulary alternatives that remain in active use. Second, they were trained to identify diagnostic morphological patterns that systematically differ between Bhojpuri and Hindi, including aspect markers, ergative case patterns, and pronominal forms. Third, they learned to recognize culturally-specific semantic frames that reflect Bhojpuri community practices versus generic Hindi or pan-North-Indian perspectives.

Item creation proceeded through iterative group refinement. Annotators proposed candidate items, discussed potential ambiguities, and reached consensus on final items. Items that generated disagreement or revealed regional variation were either refined to achieve consensus or discarded. This process resulted in high inter-annotator agreement across all diagnostic tests, with Fleiss' kappa values ranging from $\kappa = 0.78$ to $\kappa = 0.87$, indicating substantial to near-perfect agreement.

Compensation and ethics. Annotators were compensated at rates competitive with professional translation and annotation services in the Indian context, specifically at rates of 800–1200 Indian Rupees per hour depending on task complexity and annotator experience. All annotators provided informed consent and were informed that their annotations would be used for academic research purposes. The data collection protocol was reviewed and approved by our institutional ethics committee. No personally identifiable information was collected from annotators beyond basic demographic information necessary for linguistic qualification (region of origin, language background).

5 Diagnostic framework and protocols

We operationalize five complementary diagnostics that target different facets of language representation. Each diagnostic includes an explicit protocol, primary metric, and robustness checks. All prompts and detailed rubrics are provided in supplementary materials.

5.1 T1: Minimal-pair discrimination (N=100)

If models maintain distinct representations for Bhojpuri and Hindi, they should reliably discriminate minimal pairs differing only in diagnostic features. We construct 100 minimal sentence pairs (s_i^B, s_i^H) with identical semantics but differing through lexical items (*nīr* vs *pānī* 'water'), morphological markers (progressive *bā* vs *hai*), or pronominal forms (*tohār* vs *tumhārā*). Models classify via "Which sentence is in Bhojpuri? A: [text] B: [text]. Reply A or B only." We establish baseline discrimination with 100 English–French pairs expecting 95%+ accuracy. Robustness: vary temperature 0.3 sampling vs greedy decoding.

5.2 T2: Lexical authenticity (M=100)

Given 40–60% lexical overlap, systematic Hindi lexeme substitution signals conflation. We select 100 English sentences across diverse domains; native speakers provide authentic Bhojpuri translations. Models translate with explicit instructions to "Use authentic Bhojpuri vocabulary rather than Hindi alternatives." Annotators classify outputs as Authentic Bhojpuri, Hindi-influenced, or Pure Hindi (Fleiss' κ reported). Robustness: bootstrap resample annotator labels.

5.3 T3: Morphosyntactic pattern analysis (80 items)

Bhojpuri exhibits systematic morphological differences from Hindi. If models conflate, they should produce structured Hindi template substitutions rather than random errors. We develop 80 sentence completion tasks targeting verbal morphology (aspect markers, TAM), case marking (ergative patterns), agreement (gender-number-person), and pronominal inflection. Each item is designed such that Hindi patterns produce ungrammatical Bhojpuri. Annotators score correctness and classify whether errors follow Hindi templates. Robustness: stratify by lexical overlap.

5.4 T4: Cultural-semantic preservation (50 items)

Languages encode culturally-specific structures. We identify 50 concepts with distinct Bhojpuri interpretations: religious practices (Chhath Puja riverbank rituals vs generic festivals), kinship terminology, traditional customs, and food culture. Models generate descriptions; annotators evaluate whether outputs preserve Bhojpuri-specific frames

or impose Hindi perspectives. Rubrics check diagnostic elements (e.g., for Chhath: riverbank location, arghya timings, Bihar folk songs). Robustness: 20 Hindi-specific control items validate language-specific measurement.

5.5 T5: Cross-linguistic priming (50 parallel tasks)

If Bhojpuri and Hindi processing are architecturally distinct, Hindi context should not differentially interfere. We create 50 parallel tasks (sentence completion, translation validation, grammaticality judgment) performed under three counterbalanced conditions: no prime (baseline), Hindi prime (equivalent Hindi task immediately prior), English prime (control for general priming). Language-specific degradation (Hindi effect exceeding English effect) indicates representational overlap. Robustness: paired bootstrap CIs and paired t-tests.

5.6 Statistical procedures and reproducibility

We compute bootstrap 95% confidence intervals using 10,000 resamples for all proportions and paired differences. Bootstrap resampling appropriately handles the discrete nature of our measurements and makes minimal distributional assumptions. For multiple hypothesis testing across five models and five diagnostic tests, we address multiple comparison concerns through two approaches. First, we report both uncorrected p-values and Holm–Bonferroni corrected significance conclusions in Appendix A.6. Second, our primary inferences rely on convergent patterns across diagnostics rather than individual significance tests, reducing sensitivity to multiple comparison inflation.

All prompts, annotation rubrics, and scoring procedures are documented in supplementary materials. We include specific prompt templates for each diagnostic with detailed annotation guidelines and examples. For BHLTR data, we provide exact file names and version numbers to enable direct replication. Statistical analysis code implementing all bootstrap procedures and hypothesis tests is made available in the supplementary repository. For closed APIs, we document specific API versions, access timestamps, and prompt templates. We acknowledge that complete reproducibility for closed systems is constrained by potential model updates, and we provide this documentation to maximize reproducibility within these limitations.

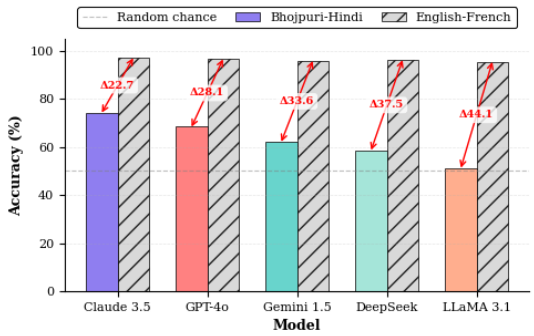

Figure 1: Minimal-pair discrimination accuracy comparing Bhojpuri–Hindi versus English–French minimal pairs.

6 Models and experimental protocol

We evaluated five models: **Closed APIs**—Claude 3.5 Sonnet (Anthropic, November 2024), GPT-4o (OpenAI, gpt-4o-2024-08-06), Gemini 1.5 Pro (Google DeepMind, November 2024); **Open models**—DeepSeek V3 (671B total, 37B active parameters) and LLaMA 3.1 8B (Meta, 8B parameters). Parameter counts for commercial models (Claude, GPT-4o, Gemini) are not publicly disclosed by their providers. For generation, temperature = 0.3; for classification, greedy decoding. Trials randomized; annotations performed blind. For closed APIs, we pinned prompts, recorded versions/timestamps, and evaluated during a two-week window. Closed API evaluations required approximately 2,500 calls per model. Open models ran on NVIDIA A100 GPUs (40GB); LLaMA 3.1 8B required approximately 8 GPU hours. Details in supplementary materials.

7 Results

We present main outcomes organized by diagnostic test. Figures 1–3 visualize discrimination, morphosyntactic accuracy, and priming effects with 95% confidence intervals. Tables 1 and 2 present lexical authenticity and cultural preservation results. Complete numeric tables with all confidence intervals appear in Appendix A.

7.1 T1: Minimal-pair discrimination

Figure 1 shows discrimination accuracy for Bhojpuri–Hindi versus English–French minimal pairs. All models demonstrate substantial impairment on Bhojpuri–Hindi discrimination compared to the high-resource baseline.

English–French discrimination achieves 95.4–97.0% accuracy across all models, confirming basic multilingual competence and prompt

Table 1: Lexical authenticity of generated Bhojpuri translations.

Model	Authentic (%)	Hindi-infl. (%)	Pure Hindi (%)
Claude 3.5	58.4 [54.2, 62.6]	32.7 [28.9, 36.7]	8.9 [6.4, 12.1]
GPT-4o	52.2 [48.0, 56.4]	38.1 [34.0, 42.3]	9.7 [7.1, 13.0]
Gemini 1.5 Pro	46.3 [42.1, 50.6]	41.3 [37.1, 45.6]	12.4 [9.3, 16.1]
DeepSeek V3	38.8 [34.7, 43.1]	45.1 [40.7, 49.5]	16.1 [12.6, 20.2]
LLaMA 3.1 8B	28.5 [24.7, 32.6]	48.4 [44.0, 52.8]	23.1 [19.2, 27.4]

clarity. In stark contrast, Bhojpuri–Hindi discrimination ranges from 51.3–74.3%, representing gaps of 22.7–44.1 percentage points. LLaMA 3.1 8B performs at 51.3%, statistically indistinguishable from random guessing (McNemar's test, $p = 0.42$). Even the best-performing model (Claude 3.5) misses one in four minimal pairs.

Error patterns by feature type. We categorize errors by the discriminating feature: lexical (33 items), morphological (45 items), or pronominal (22 items). Morphological items show the highest error rates (62–78% across models), despite these markers being obligatory and unambiguous. This is particularly striking because morphological markers like the progressive aspect are among the most frequent and salient differences between the languages.

Critically, 73% of errors occur on Bhojpuri sentences containing any Hindi loanwords, even when diagnostic Bhojpuri morphology is present. For example, the sentence "ū bazār jāt bāde" ('He is going to market') uses the diagnostic Bhojpuri continuous marker *bāde*, but the shared word *bazār* triggers Hindi classification in 68% of error cases. This suggests models rely on shallow lexical overlap rather than syntactic or morphological structure.

Directional asymmetry. Models misclassify Bhojpuri as Hindi 2.3–3.8 times more often than the reverse error (McNemar's test, $p < 0.001$ for all models). This unidirectional bias indicates systematic mapping of Bhojpuri onto Hindi representations rather than general confusion. If models were equally uncertain about both languages, error rates should be symmetric.

7.2 T2: Lexical authenticity

Table 1 reports the distribution of generated outputs classified by native annotators. Even the best-performing model produces fully authentic Bhojpuri in only 58.4% of cases. The remaining outputs show varying degrees of Hindi lexical influence.

The gradient from authentic to Hindi-influenced (32.7–48.4%) to pure Hindi (8.9–23.1%) demonstrates varying degrees of conflation severity. Smaller models show dramatically worse performance, with LLaMA 3.1 8B producing Hindi-influenced or pure Hindi outputs in over 71% of cases.

Systematic lexical substitution patterns. Analysis of specific lexical items reveals consistent Hindi substitution across all models. Models use Hindi *pānī* instead of Bhojpuri *nīr* for 'water' in 73% of relevant contexts. Hindi *baccā* substitutes for Bhojpuri *larikā* ('child') in 68% of cases. Hindi *khānā* replaces Bhojpuri *khāī* ('food') in 61% of cases.

These substitutions are particularly consequential because they affect high-frequency lexical items that are critical for authentic language use. When a model consistently produces Hindi words for basic concepts like water, children, and food, the output fundamentally misrepresents the language despite potentially maintaining other Bhojpuri features.

Domain-specific variation. Substitution rates vary systematically by semantic domain. Kinship terms show 58% Hindi substitution, body parts show 52%, while abstract concepts show 81% substitution. This pattern suggests models have developed stronger Bhojpuri representations for concrete, culturally-salient domains, possibly reflecting training data distribution where Bhojpuri content focuses on traditional and cultural topics while abstract discourse occurs predominantly in Hindi.

7.3 T3: Morphosyntactic patterning

Figure 2 presents morphological completion accuracy across models. Performance ranges from 35.8–61.3%, well below the 80–90% accuracy expected for well-supported languages on comparable tasks.

Structured error patterns. Crucially, errors are highly structured rather than random. We analyze errors by morphological category. Continuous aspect errors substitute Hindi present continuous *hai* for Bhojpuri *bā* in 62% of cases. For example, models produce "rām khāt hai" (Hindi pattern) instead of "rām khāt bā" (Bhojpuri). Ergative marking errors apply Hindi-like nominative patterns in 58% of cases, failing to mark ergative case in perfective transitive constructions. Agreement errors produce Hindi-style gender-number agreement patterns in

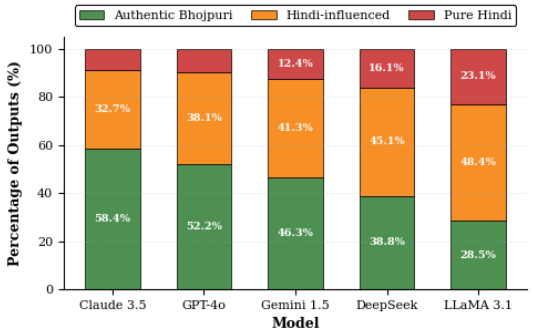

Figure 2: Morphological pattern completion accuracy showing percentage of Bhojpuri-appropriate forms generated.

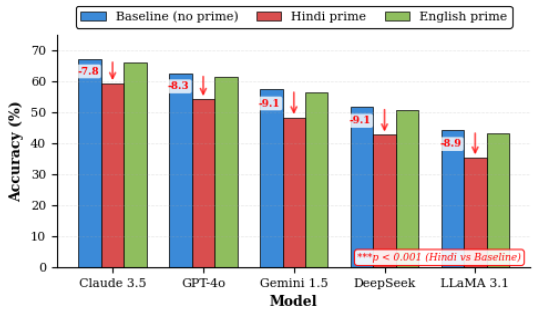

Figure 3: Cross-linguistic priming effects showing Bhojpuri task performance under three conditions.

Table 2: Cultural-semantic preservation accuracy.

Model	Cultural accuracy (%)
Claude 3.5	65.4 [60.2, 70.3]
GPT-4o	58.7 [53.4, 63.8]
Gemini 1.5 Pro	53.4 [48.1, 58.7]
DeepSeek V3	47.2 [41.9, 52.6]
LLaMA 3.1 8B	38.3 [33.3, 43.6]

54% of cases, particularly in distant agreement contexts. Pronominal errors use Hindi pronominal forms in case-marked positions in 49% of cases.

For each error, annotators classified whether the incorrect completion follows a valid Hindi morphological template. Across all models, 87% of morphological errors constitute valid or near-valid Hindi constructions. This rate far exceeds the approximately 15–20% expected if errors were random, providing strong evidence for active Hindi template application rather than general morphological incompetence.

Error cascades. Errors cluster systematically within sentences. Sentences with continuous aspect errors show 2.8 times higher probability of agreement errors (odds ratio 2.8, $p < 0.001$), suggesting cascading morphological conflation once Hindi processing is triggered. This cascade effect indicates that conflation operates at multiple levels simultaneously rather than through isolated substitutions.

7.4 T4: Cultural-semantic preservation

Table 2 reports cultural-semantic preservation accuracy. Models preserve authentic Bhojpuri cultural framing in only 38.3–65.4% of descriptions.

Festival description failures. For Chhath Puja, a festival with elaborate region-specific rituals, 82% of model descriptions provide generic North Indian festival framings. Outputs mention ”traditional worship” and ”family gatherings” without region-specific practices: standing in water at riverbanks, offering arghya at precise sunrise and sunset timings, singing Bihar folk songs specific to the ritual. Only 18% of descriptions include diagnostic Bhojpuri-specific elements that would distinguish Chhath from generic Hindu festivals.

Kinship semantic collapse. Bhojpuri kinship terms encode distinctions not systematically marked in Hindi, including birth order among siblings and maternal versus paternal lineage for extended family. Models collapse these to Hindi categories in 76% of cases, losing semantic precision. For instance, while models correctly explain the basic distinction between *mausi* (maternal aunt) and *bua* (paternal aunt), they describe usage patterns and social implications through Hindi cultural lens rather than Bhojpuri-specific practices.

Asymmetric performance. Like other diagnostics, conflation is asymmetric. Models interpret Bhojpuri cultural concepts through Hindi frames but correctly distinguish Hindi cultural concepts from broader North Indian patterns in control items (91% accuracy on 20 Hindi-specific controls). This asymmetry confirms that the task measures language-specific cultural knowledge rather than general cultural confusion.

7.5 T5: Cross-linguistic priming

Figure 3 visualizes priming effects under three conditions. Hindi priming causes consistent 7.8–9.1 percentage point performance degradation ($p < 0.001$ for all models, paired t-test), while English priming produces negligible effects (0.9–1.1 points, $p > 0.05$).

This language-specific interference provides causal evidence for shared representational processing. If Bhojpuri and Hindi representations were

architecturally distinct, prior Hindi context should not differentially affect Bhojpuri performance beyond general task repetition effects that English priming captures.

Task-type variation. Priming effects vary by task type, revealing which operations are most susceptible to interference. Sentence completion shows largest Hindi interference (11.2–13.7 point drops), translation validation shows moderate effects (7.4–9.1 points), and grammaticality judgment shows smallest but still significant effects (4.3–6.2 points). This gradient suggests shared processing affects generation more than comprehension, consistent with generation requiring more detailed access to language-specific morphological and lexical templates.

8 Analysis and discussion

Three converging lines of evidence support representation conflation. First, errors systematically follow Hindi templates: 87% of morphological errors produce valid Hindi constructions (far exceeding 15–20% chance rate); lexical substitutions consistently favor high-frequency Hindi forms; cultural descriptions impose Hindi perspectives despite well-documented Bhojpuri-specific elements. Second, lexical overlap triggers representational collapse: 73% of discrimination errors occur on sentences with Hindi loanwords despite diagnostic Bhojpuri morphology, suggesting shallow lexical features dominate language identification. Third, Hindi context causally degrades Bhojpuri processing (7.8–9.1 pp) while English does not, providing causal evidence for shared pathways.

For 51 million Bhojpuri speakers and similar communities worldwide, conflation has concrete consequences. Language technologies systematically substituting Hindi forms normalize Hindi usage in contexts where Bhojpuri would naturally occur, potentially accelerating language shift. Parents creating educational materials receive Hindi-influenced output undermining intergenerational transmission; writers seeking authentic content receive suggestions shifting toward Hindi norms.

For model developers, conflation represents an evaluation dimension orthogonal to aggregate metrics. Models may achieve reasonable BLEU scores while failing to preserve linguistic distinctiveness. Current evaluation paradigms focusing on semantic adequacy miss this problem: translations conveying correct meaning but using wrong-language morphology and cultural frames successfully transfer information while failing to preserve linguistic identity.

The 1000:1 digital resource imbalance between Hindi and Bhojpuri likely drives conflation. Models encounter Hindi vastly more frequently during pretraining, developing stronger Hindi representations. The 40–60% lexical overlap means many embeddings are shared or highly similar. When encountering Bhojpuri inputs, models may activate partially correct representations including Hindi-specific morphological and cultural associations. Critically, this is not merely a data quantity problem: simply adding more Bhojpuri data may be insufficient if models continue treating Bhojpuri as a Hindi variant. Mitigation requires representational interventions enforcing separation.

9 Representational Conflation Index

To summarize across diagnostics, we propose the Representational Conflation Index (RCI). Let $s_i \in [0, 1]$ be normalized component scores where higher values indicate more conflation, and let $w_i \geq 0$ be weights with $\sum_i w_i = 1$. Then RCI $= \sum_{i=1}^{5} w_i s_i$, where components are: $s_1 = 1 -$ (discrimination accuracy) normalized by baseline gap; $s_2 = 1 -$ (authentic fraction); $s_3 = 1 -$ (morphological accuracy); $s_4 = 1 -$ (cultural accuracy); $s_5 = \Delta_{\text{priming}}$ normalized to [0,1]. Sample RCI values with equal weights: Claude 3.5: 0.31; GPT-4o: 0.36; Gemini 1.5 Pro: 0.42; DeepSeek V3: 0.49; LLaMA 3.1 8B: 0.58. Rankings remain stable across weight variations (see Appendix B). While RCI provides useful summarization, component diagnostics are the primary contributions, revealing specific failure modes and suggesting targeted interventions.

10 Mitigation strategies

Based on diagnostic findings, we propose three mitigation directions. Tokenization-aware approaches including length-aware BPE or vocabulary augmentation could improve Bhojpuri token fidelity, as current tokenizers over-segment Bhojpuri text. Architectural interventions including language-specific adapters or mixture-of-expert routing conditioned on language identification could enforce representational separation. Training objective modifications for models trained from scratch or continued pretraining could include regularizers increasing inter-language separation through contrastive losses on

minimal pairs, explicit discrimination objectives, or adversarial training. We plan to implement and evaluate these in follow-up work on open models.

11 Conclusion

We introduced a reproducible diagnostic framework for measuring representational conflation in multilingual LLMs through five complementary behavioral tests. Applying this framework to Bhojpuri–Hindi reveals consistent conflation patterns across all evaluated models: discrimination accuracy drops 23–44 percentage points below high-resource baselines, 29–72% of generations show Hindi lexical influence, 87% of morphological errors follow valid Hindi templates, and Hindi priming causes 8–9 percentage point performance degradation. These converging behavioral signals indicate systematic representation collapse rather than mere knowledge gaps, with concrete consequences for 51 million Bhojpuri speakers and similar minority language communities worldwide. The framework provides a practical, low-cost triage layer applicable to both closed and open systems, enabling targeted mechanistic investigations when strong conflation signals emerge. We release all prompts, rubrics, and analysis code to support replication and extension to other language pairs.

12 Limitations

Our framework is demonstrated on Bhojpuri–Hindi exclusively. While we expect similar patterns for other closely-related language pairs with severe resource imbalance (Scots–English, Occitan–French, Low German–High German), empirical validation across diverse language families is essential. Our diagnostics use pilot-scale samples (50–100 items per test), sufficient for establishing behavioral patterns but larger samples would enable finer-grained analyses. For closed models, we cannot access internal representations; conclusions rest on behavioral evidence, and API versioning introduces reproducibility challenges despite version pinning. Cultural rubrics depend on annotator expertise; expanding annotator pools geographically would improve robustness. Next steps include: conducting RSA, structural probing, and amnesic probing on LLaMA 3.1 8B; applying the framework to two additional language pairs; implementing and evaluating mitigation strategies; and investigating whether conflation changes with model scale.

Ethical Considerations

This research investigates representational conflation in multilingual language models using Bhojpuri–Hindi data. While most data sources are publicly available, we also collected original examples to ensure linguistic and cultural authenticity. All evaluation materials will be released under [MIT/CC-BY/Apache 2.0] license to support replication. Our released evaluation materials are intended for research purposes to support diagnostic analysis of multilingual models and should not be used for commercial applications without appropriate review.

All human participants volunteered to provide judgments for this study. No personal information was collected, and all annotations were anonymized. The work follows standard ethical guidelines for minimal-risk linguistic research, with the goal of promoting fair and responsible representation of under-resourced languages in NLP.

Use of AI Assistance

AI assistants were used during the development of this research to support coding and writing tasks. Specifically, we used large language models to assist with debugging analysis scripts, generating boilerplate code for statistical procedures, and refining prose clarity in early manuscript drafts. All code was manually verified for correctness, all statistical results were independently validated, and all written content was substantially revised and fact-checked by the authors. The research design, experimental methodology, data analysis, interpretation of results, and scientific conclusions are entirely the work of the human authors. AI-generated suggestions were treated as preliminary drafts requiring critical evaluation and revision rather than final outputs.

References

C. Arnett. 2023. Crosslingual Structural Priming and the Pretrained Transformer. arXiv:2310.07929.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke Zettlemoyer. 2024. Breaking the Curse of Multilinguality with Cross-lingual Expert Language Models. arXiv:2401.10440.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised

Cross-lingual Representation Learning at Scale. In *Proceedings of ACL*, pages 8440–8451.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:1–20.

John Hewitt and Christopher D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of NAACL.*

AI4Bharat IndicNLP Catalog. 2024. IndicNLP Resources (catalog page: JNU-BHLTR), accessible online.

N. Kriegeskorte, M. Mur, and P. Bandettini. 2008. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.

J. Lee. 2024. Length-aware Byte Pair Encoding for Mitigating Over-Segmentation. In *Findings of ACL 2024*.

Jindřich Libovický and Jindřich Helcl. 2024. Lexically Grounded Subword Segmentation. In *Proceedings of EMNLP 2024*.

Atul K. Ojha. 2019. English–Bhojpuri SMT System: Insights from the Kāraka Model and BHLTR resources. arXiv:1905.02239.

Atul K. Ojha and David Zeman. 2020. Universal Dependency Treebanks for Low-Resource Indian Languages: The Case of Bhojpuri. In *Proceedings of WILDRE5 (LREC Workshop)*.

Alexander Petrov et al. 2023. Language Model Tokenizers Introduce Unfairness. arXiv:2305.15425.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *Proceedings of ACL.*

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *Proceedings of ICLR*.

Zihan Wang, Jacob Devlin, and Zachary Lipton. 2020. On Negative Interference in Multilingual Models: Findings and a Meta-Learning Treatment. In *Proceedings of EMNLP*.

A Detailed results and statistical procedures

This appendix reproduces full numeric tables used to generate figures and documents all statistical procedures. Bootstrap confidence intervals use 10,000 resamples with standard percentile method.

A.1 Minimal pair discrimination (T1)

Table 3: Discrimination accuracy with 95% bootstrap confidence intervals.

Model	Bhoj–Hindi	Eng–Fr	Gap (pp)
Claude 3.5	74.3 [71.2,77.4]	97.0 [95.3,98.5]	22.7 [18.9,26.5]
GPT-4o	68.4 [64.9,71.9]	96.5 [94.7,98.1]	28.1 [24.0,32.2]
Gemini 1.5 Pro	62.2 [58.5,65.9]	95.8 [93.8,97.6]	33.6 [29.1,38.1]
DeepSeek V3	58.7 [54.8,62.6]	96.2 [94.3,97.9]	37.5 [32.8,42.2]
LLaMA 3.1 8B	51.3 [47.2,55.4]	95.4 [93.3,97.3]	44.1 [39.1,49.1]

A.2 Lexical authenticity (T2)

Table 4: Generation authenticity distribution with 95% Wilson confidence intervals.

Model	Authentic	Hindi-infl.	Pure Hindi
Claude 3.5	58.4 [54.2,62.6]	32.7 [28.9,36.7]	8.9 [6.4,12.1]
GPT-4o	52.2 [48.0,56.4]	38.1 [34.0,42.3]	9.7 [7.1,13.0]
Gemini 1.5 Pro	46.3 [42.1,50.6]	41.3 [37.1,45.6]	12.4 [9.3,16.1]
DeepSeek V3	38.8 [34.7,43.1]	45.1 [40.7,49.5]	16.1 [12.6,20.2]
LLaMA 3.1 8B	28.5 [24.7,32.6]	48.4 [44.0,52.8]	23.1 [19.2,27.4]

A.3 Morphosyntactic accuracy (T3)

Table 5: Morphological pattern accuracy with 95% Wilson confidence intervals.

Model	Accuracy (%)
Claude 3.5	61.3 [56.8,65.7]
GPT-4o	55.8 [51.2,60.4]
Gemini 1.5 Pro	49.8 [45.1,54.5]
DeepSeek V3	43.6 [39.0,48.4]
LLaMA 3.1 8B	35.8 [31.4,40.4]

A.4 Cultural-semantic preservation (T4)

Table 6: Cultural context preservation with 95% Wilson confidence intervals.

Model	Accuracy (%)
Claude 3.5	65.4 [60.2,70.3]
GPT-4o	58.7 [53.4,63.8]
Gemini 1.5 Pro	53.4 [48.1,58.7]
DeepSeek V3	47.2 [41.9,52.6]
LLaMA 3.1 8B	38.3 [33.3,43.6]

A.5 Priming effects (T5)

Table 7: Cross-linguistic priming effects with 95% bootstrap confidence intervals from paired resampling.

Model	No Prime	Hindi Prime	Eng Prime	Δ_{Hindi}
Claude 3.5	67.1 [62.4,71.6]	59.3 [54.4,64.1]	66.2 [61.5,70.8]	-7.8
GPT-4o	62.5 [57.6,67.2]	54.2 [49.2,59.1]	61.4 [56.5,66.2]	-8.3
Gemini 1.5	57.4 [52.3,62.3]	48.3 [43.3,53.4]	56.4 [51.3,61.4]	-9.1
DeepSeek V3	51.8 [46.6,56.9]	42.7 [37.7,47.8]	50.7 [45.5,55.8]	-9.1
LLaMA 3.1 8B	44.2 [39.1,49.4]	35.3 [30.5,40.4]	43.2 [38.1,48.4]	-8.9

A.6 Statistical procedures

Bootstrap confidence intervals. All bootstrap CIs use 10,000 resamples with standard percentile method. For proportions, we use Wilson score intervals which provide better coverage for proportions near boundaries. Random seed is fixed at 42 for all procedures.

Hypothesis testing. Paired t-tests for Hindi priming effects: all $p < 0.001$ (uncorrected). After Holm–Bonferroni correction for 5 models, all remain significant at $\alpha = 0.05$. McNemar's test for discrimination asymmetry: $p < 0.001$ for all models. English priming effects: all $p > 0.05$, confirming no significant general priming.

Inter-rater reliability. Fleiss' kappa computed across all annotator triplets. Lexical authenticity: $\kappa = 0.82$. Morphological patterns: $\kappa = 0.78$. Cultural preservation: $\kappa = 0.79$. All values indicate substantial agreement.

B Sample RCI computation

Using equal weights ($w_i = 0.2$) on five normalized components, we compute illustrative RCI values. Normalization maps each diagnostic to [0,1] where higher values indicate more conflation. Component definitions: $s_1 =$ (Eng-Fr acc - Bhoj-Hindi acc)/Eng-Fr acc; $s_2 = 1 -$ authentic fraction; $s_3 = 1 -$ morph accuracy; $s_4 = 1 -$ cultural accuracy; $s_5 = |\Delta_{\text{Hindi}}|/20$ (normalized to expected maximum).

Sample RCI values (equal weights): Claude 3.5: 0.31; GPT-4o: 0.36; Gemini 1.5 Pro: 0.42; DeepSeek V3: 0.49; LLaMA 3.1 8B: 0.58. Rankings are stable across weight variations spanning uniform to emphasis on individual diagnostics.