Quantifying Cross-Lingual Interference: Algorithmic Standardization of Kamtapuri in Large Language Models

Roumak Das
Independent Researcher
roumak.das@gmail.com

Abstract

Multilingual Large Language Models (LLMs) often demonstrate impressive zero-shot capabilities on low-resource languages. However, for languages that share a script and significant lexical overlap with a high-resource language (HRL), models may exhibit negative transfer. Focusing on **Kamtapuri** (Rajbanshi), a distinct low-resource language of North Bengal, we investigate the extent to which SOTA models (e.g., GPT-5.1, Gemini 2.5) preserve distinct dialectal features versus reverting to the dominant language's norms. We introduce the **Kamta-Shibboleth-100**[1], a diagnostic benchmark derived from a curated 400k-token corpus. Our evaluation reveals a significant discrepancy: while models show high receptive understanding (up to 88% translation accuracy), they exhibit a **0% Syntactic Competence Rate** in zero-shot generation of distinct Kamtapuri morphology, compared to **96%+ accuracy** on a Standard Bengali control set. Even with 5-shot prompting, syntactic accuracy improves only to 10%, while the **Substitution Erasure Rate (SER)** reaches 71%, systematically replacing Kamtapuri vocabulary with Bengali cognates. We characterize this behavior not as a lack of knowledge, but as a strong alignment bias toward high-resource standards.

1 Introduction

The development of massively multilingual Large Language Models (LLMs) aims to democratize access to language technology, theoretically covering hundreds of languages (NLLB Team et al., 2022; Magueresse et al., 2020). However, the performance of these models on **closely related languages** remains an open research question. When a low-resource language (LRL) shares a script and lexical roots with a high-resource language (HRL), models often struggle to distinguish between the two (Joshi et al., 2020). This leads to a phenomenon we term **Algorithmic Standardization**: the tendency of a model to normalize the LRL input into the syntax and vocabulary of the HRL.

This study analyzes **Kamtapuri** (also known as Rajbanshi), a language spoken across North Bengal (India) and parts of Nepal (Toulmin, 2006; Grierson, 1903). Kamtapuri shares the Eastern Nagari script with Standard Bengali but possesses distinct morphosyntax, such as unique verb inflections and case markers (Wilde, 2008). Due to the scarcity of digital Kamtapuri data and the overwhelming abundance of Standard Bengali data in pre-training corpora, LLMs risk treating Kamtapuri as "noisy" or "non-standard" Bengali, a known failure mode in cross-lingual transfer (Blodgett et al., 2020).

To quantify this effect, we leverage a privately curated dataset of Kamtapuri folklore to evaluate late-2025 SOTA models. We employ a dual-phase evaluation strategy: a **Zero-Shot** baseline to simulate default model behavior, followed by a **Few-Shot** probe to assess whether in-context learning can override standardization bias (Hershcovich et al., 2022).

1.1 Contributions

1. **The Kamta-Shibboleth-100 Benchmark:** We introduce a diagnostic challenge set of 100 constraint-locked sentence pairs designed to isolate specific morphological and lexical differences between Kamtapuri and Bengali.

2. **Comparative Control Study:** We validate model capabilities using a control set of 30 Standard Bengali sentences. Models achieved **>96% accuracy** on the control, contrasting with near-zero accuracy on Kamtapuri syntax, isolating the failure to language-specific features.

3. **The SER Metric:** We propose the *Substitu-*

[1] Benchmark available at: https://github.com/kamtapuri-research/Kamta-Shibboleth-100-BENCHMARK

tion Erasure Rate (SER) to quantify the frequency with which valid low-resource terms are replaced by high-resource cognates during generation.

2 Methodology

2.1 Data Source: Curated Corpus

This study utilizes a proprietary corpus of approximately 400,000 Kamtapuri tokens (*Kamta-400k*). The data was digitized from printed sources, including 15 volumes of folklore (c. 1970–1995) and transcripts from community radio broadcasts. This ensures the evaluation is grounded in authentic, high-quality text rather than potentially noisy web-scraped data (Bird, 2020).

Script Note: All experiments were conducted using the native Eastern Nagari script. Examples in this paper are Romanized for readability.

2.2 Benchmark Construction

To distinguish between general generation failure and specific standardization bias, we constructed the **Kamta-Shibboleth-100** using a dual-prompt strategy.

Subset A: Context-Locked Grammar ($N = 30$). These examples rely on *Syntactic Constraint Locking*. We selected sentences where the grammatical subject forces a specific morphological inflection unique to Kamtapuri.

- *Prompt:* `Mui sungsar-ta [MASK].`
- *Target (Kamtapuri):* `Cholang` (unique inflection).
- *Standardized (Bengali):* `Chalai`.

A correct generation requires the model to adhere to Kamtapuri grammar rules despite the high probability of the Bengali token.

Subset B: Meaning-Locked Vocabulary ($N = 100$). These examples test lexical availability. The model is prompted with a sentence and an English concept (to resolve ambiguity) and must generate the correct word.

- *Prompt:* `[MASK] (Now) mor matha ...`
- *Target:* `Elay` (Kamtapuri).
- *Standardized:* `Ekhon` (Bengali).

2.3 Metric: Substitution Erasure Rate

We define SER to quantify the conditional probability of a model reverting to the HRL given that it understood the semantic intent.

$$SER = \frac{N_{standardized}}{N_{correct\_meaning}} \times 100 \qquad (1)$$

Where $N_{standardized}$ is the count of responses where the model generates the Standard Bengali cognate despite the prompt being in Kamtapuri, and $N_{correct\_meaning}$ is the subset of responses where the model correctly understood the semantic context.

3 Experimental Setup

We evaluated four State-of-the-Art models: **GPT-5.1**, **GPT-4o**, **Gemini 2.5 Pro**, and **Gemini 2.5 Flash**.

Zero-Shot Setting: We primarily utilize a Zero-Shot strategy with implicit language cues (i.e., the prompt is written in the target language, but the language name is not explicitly invoked in the system instruction). This tests the model's ability to identify and adhere to the dialectal context based on linguistic features alone.

Few-Shot Probe: To address limitations regarding in-context learning, we perform a secondary evaluation on Gemini 2.5 Pro using 1, 3, and 5-shot examples.

4 Results

4.1 Syntactic Accuracy (Subset A)

Table 1 presents the zero-shot accuracy on grammatical constraints.

Table 1: Zero-Shot Syntactic Accuracy comparison. Models successfully model the grammatical constraints of the high-resource language (Bengali) but fail to generalize to the low-resource sister language.

Model Family	Kamta Acc.	Bengali Control
GPT Family	**0.00%**	96.6%
Gemini Family	**0.00%**	100.0%

The **0% Syntactic Competence** on Kamtapuri indicates that the distinct grammatical rules of the language are not represented in the models' top-k output distribution, despite the models performing near-perfectly on the Bengali control set.

4.2 Lexical Standardization (Subset B)

Table 2 shows the results for zero-shot vocabulary generation.

Table 2: Substitution Erasure Rate (SER) on Vocabulary ($N = 100$). A high SER indicates the model replaced the Kamtapuri target with a Bengali equivalent.

Model	Success	SER (Standardized)	Hallucination
GPT-5.1	6.00%	52.00%	42.00%
GPT-4o	3.00%	39.00%	58.00%
Gemini 2.5 Pro	**11.00%**	**62.00%**	27.00%
Gemini 2.5 Flash	7.00%	58.00%	35.00%

4.3 Impact of Few-Shot Prompting

To determine if the observed standardization bias could be corrected via in-context learning, we evaluated **Gemini 2.5 Pro** with varying numbers of exemplars ($k = \{1, 3, 5\}$). The results are summarized in Table 3.

Table 3: Few-Shot Performance on Gemini 2.5 Pro. While accuracy improves marginally, the Substitution Erasure Rate (SER) remains persistently high.

Shots (k)	Vocab Accuracy	Vocab SER
0-Shot	11.0%	62.0%
1-Shot	16.0%	68.0%
3-Shot	19.0%	67.0%
5-Shot	**23.0%**	**71.0%**

Vocabulary Task: Providing examples improved accuracy from 11% to 23%. However, the SER actually *increased* to 71% in the 5-shot setting. This counter-intuitive finding suggests that as the model becomes more confident in the task format, it retrieves Standard Bengali synonyms with higher probability, effectively "overwriting" the lower-probability Kamtapuri terms.

Grammar Task: Syntactic constraints remained highly resistant to few-shot correction.

- **1-Shot:** 0% Accuracy.
- **3-Shot:** 3.3% Accuracy (1/30).
- **5-Shot:** 10.0% Accuracy (3/30).

Even with 5 clear examples of Kamtapuri verb morphology, the model failed 90% of the time, reverting to Bengali verbal inflections.

4.4 Receptive Competence

To distinguish between generation failure and lack of understanding, we evaluated translation accuracy (Kamtapuri to English). SOTA models achieved up to **88% accuracy** (Gemini 2.5 Pro). This high receptive competence contrasts with the low productive competence.

5 Discussion

5.1 The Receptive-Productive Gap

The disparity between Receptive Competence (88% translation accuracy) and Productive Competence (<23% generation accuracy) is the central finding of this study. It raises the question: *If the model understands the language enough to translate it, why cannot it generate it?*

We hypothesize this is a result of **Probability Mass Alignment**. During pre-training and instruction tuning, the model updates its weights to maximize the likelihood of high-quality, standard text. Since Standard Bengali is the dominant representation for the Eastern Nagari script in the training corpus, the model's decoder likely assigns near-zero probability to Kamtapuri-specific morphemes (e.g., *-ng* endings), treating them as "errors" or "noise" to be corrected to the nearest Standard Bengali equivalent (e.g., *-i* endings). Thus, the model possesses the *latent representation* to decode the input (Receptive), but the *generation head* is aligned to suppress the low-resource dialect (Productive) (Hovy and Spruit, 2016).

5.2 Implications for Low-Resource NLP

These findings suggest that "scale" alone is insufficient for dialect preservation. In fact, larger models (like Gemini 2.5 Pro) exhibited higher rates of standardization (62-71% SER) compared to smaller models. This implies that stronger alignment makes the model *more* rigid in enforcing standard language norms, to the detriment of minority dialects.

6 Limitations

Sample Size and Scope: The *Kamta-Shibboleth-100* is a diagnostic probe ($N = 130$) rather than a comprehensive corpus. The 0-10% syntactic accuracy represents a failure on specific, high-frequency constraints, not necessarily a total lack of linguistic knowledge. A larger, more diverse test set covering various tenses and cases would be required to map the full extent of the models' grammatical deficits.

Prompting Strategy: We evaluated models primarily in a zero-shot setting to simulate natural user interaction. While we included a limited few-shot probe, we did not employ explicit "persona" prompting (e.g., "You are a Kamtapuri speaker"). It is possible that explicit instructions could shift the probability distribution more effectively than examples alone.

Dataset Bias: Our control set focuses on Standard Bengali. We did not test for interference from other regional dialects (e.g., Sylheti, Chittagonian), which limits our ability to fully map the dialect continuum.

7 Data Availability and Ethics

Dataset Access The *Kamta-400k* corpus was constructed by digitizing copyrighted physical books. To comply with intellectual property rights regarding non-consumptive research, we do not release the full training corpus. However, to facilitate reproducibility, the **Kamta-Shibboleth-100** benchmark is released publicly: https://github.com/kamtapuri-research/Kamta-Shibboleth-100-BENCHMARK

Indigenous Data Sovereignty Much of the source material represents the cultural heritage of the Kamtapuri community. In line with Indigenous Data Sovereignty principles (Bird, 2020), we prefer a controlled release of specific, non-sensitive benchmark examples over indiscriminate open-sourcing.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback, particularly regarding the expansion of the experimental scope to include few-shot evaluations. This research was conducted independently without external funding.

References

Steven Bird. 2020. Decolonising speech and language technology. *Proceedings of COLING*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

George Abraham Grierson. 1903. *Linguistic Survey of India: Indo-Aryan Family Eastern Group*, volume V. Office of the Superintendent of Government Printing, Calcutta.

Daniel Hershcovich, Stella Frank, Jamie Lenz, and Miryam de Lhoneux. 2022. Challenges and strategies in cross-cultural nlp. *Proceedings of ACL*.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of ACL*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past trends and future challenges. *arXiv preprint arXiv:2006.07264*.

NLLB Team and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Matthew W.S. Toulmin. 2006. *Reconstructing linguistic history in a dialect continuum: The Kamta, Rajbanshi, and Northern Deshi Bangla subgroup of Indo-Aryan*. Australian National University.

Christopher P. Wilde. 2008. *A grammatical description of the Rajbanshi language*. University of Helsinki.

A Methodology Examples

Table 4: Examples illustrating how the benchmark distinguishes between valid dialectal morphology (Target) and standardization (Trap).

Type	Prompt (Romanized)	Target	Trap (Bengali)
Grammar	Mui sungsar-ta [MASK].	*Cholang*	*Chalai*
Vocab	[MASK] (Now) dekha jay na.	*Elay*	*Ekhon*
Grammar	Tui thike [MASK] ma.	*Kochis*	*Bolchis*
Vocab	[MASK] (Our) to bhale hobe.	*Hamar*	*Amader*