Sambalpuri Sieve: Probing Dialectal Drift and Cultural Grounding in SLMs and LLMs

Sumit Barik
IIIT Naya Raipur

January 2026

Abstract

This report identifies and evaluates ”Standard Language Bias” exhibited by Small Language Models(SLMs) and Large Language Models(LLMs) when translating to Sambalpuri, a dialect of Odia language widely spoken in Western Odisha, India. This study measures ”Dialectal Drift” towards the parent language and loss of cultural and social context while translating sentences. We have used Microsoft Phi 3.5, Phi 4 mini and Google's Gemini 3 Flash model for performing the experiments. Our evaluation highlights a significant representation gap for the Sambalpuri dialect in the digital world. This reduces the local reach and accessibility of the language models and also highlights a bias towards the high-resource languages.

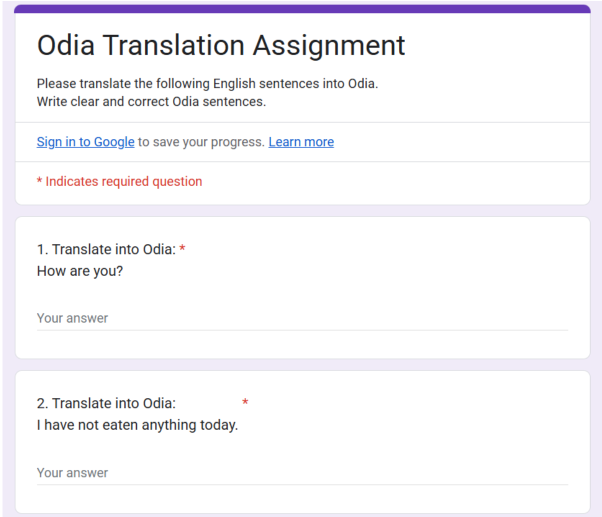
1 Introduction

The Sambalpuri dialect is spoken by approximately 10 million people across India. Even after having so many speakers, the dialect is still a low-resource language in digital formats. This makes it difficult for models to translate text to Sambalpuri without losing the social and cultural context. Here we will try to quantify the Lexical, Morphological and Syntactic Drift where models incorrectly default to Standard Odia (the high-resource parent language) or Hindi/Bengali in some cases during the translation tasks. In this work, we would use 100 sentences translated by an expert and quantify the lexical, morphological and syntactic changes made by LLMs during translation. Further, we would calculate a Fidelity Score between 0 and 4 and see where the model performs well and where it needs improvement. We provide a diagnostic analysis of model limitations in capturing dialectal nuances—such as unique verb conjugations, case marking and pronominal shifts—that are currently lost in mainstream multilingual representations.

2 Methodology

2.1 Data Collection

We curated a parallel corpus of 100 sentences across five categories: Lexical, Morphological, Syntactic, Cultural, and Honorific. The English sentences are collected from the day-to-day lives of native Sambalpuri speakers. Short sentences have been selected, along with a special focus on local traditions, idioms, and uncommon words. Further, the sentences have been translated to standard Odia by an expert researcher and to Sambalpuri by a Native speaker. Forms have been used to take inputs directly and easily collect the responses. Further, the differences between standard Odia and Sambalpuri has been categorised into Morphological, Lexical, Syntactic, Honorific and Cultural for further evaluation. These data is stored along with a unique id in a CSV file.


Odia Translation Assignment

Please translate the following English sentences into Odia.
Write clear and correct Odia sentences.

Sign in to Google to save your progress. Learn more

* Indicates required question

1. Translate into Odia: *
How are you?

Your answer

2. Translate into Odia: *
I have not eaten anything today.

Your answer

Figure 1: Data Collection Using Forms

2.2 Evaluation Metric: Fidelity Score

Each model output was scored on a scale of 0–4:

- **4 (Perfect):** Accurate dialectal markers and meaning.
- **2 (Partial):** Gives correct meaning but defaults to Standard Odia markers.
- **0 (Failure):** Semantic hallucination or script contamination.

3 Performance of SLMs

I first tested Phi 3.5, an SLM model developed by Microsoft. On prompting it to translate a simple text like ”How are you?” to Odia, it defaulted to its high-resource neighbour, Bengali, for script and grammar. Further, when asked about the Odia language, the model misclassified Odia as a ”Bengali-Assamese” language, proving a major gap in pre-training data for the Odia language. Further, I tested the Phi-4-mini model, the latest SLM in the Phi series. On asking for the same translation, it also resorted to the use of the Bengali Script. On asking the meaning of the sentence ” ” (Wake up and wash your face), it replied with ”Yes indeed”, highlighting total semantic hallucination. The model showed high confidence, offering help with Odia despite having no competence. This is a critical RLHF issue. Thus, we can see Meta Hallucination as well as Script Contamination, forcing us to use more powerful LLMs for further experiments.

4 Quantitative Analysis

4.1 Global Performance Metrics

The Mean Fidelity Score was found to be 2.14. This indicates that the model performs at ”partially correct” levels. This tells us that the model is able to grasp the meaning of the Odia sentence but fails to apply linguistic markers specific to the Sambalpuri language. It shows that the model has surface-level fluency and is not adapted to nuanced local variations. A Relative Levenshtein Gain (RLG) of 0.0651 shows that the model largely produced standard Odia outputs and made only superficial changes, failing to bridge the lexical gap between the two languages. A Pearson Correlation of 0.0244 means there is no relation between length and Fidelity Score. A small skewness of -0.3690 tells that the model consistently gives average performance without reaching native level fidelity. The following table summarises the primary statistical findings of the evaluation:

Metric	Value	Research Implication
Mean Fidelity Score	2.14	Moderate fluency; lacks nuance.
Mean RLG	0.0651	Minimal gain over Standard Odia.
Pearson Correlation (r)	0.0244	Failures are linguistic, not length-based.
Score Skewness	-0.3690	Consistent moderate underperformance.

Table 1: Summary of Quantitative Evaluation

4.2 Categorical Hit Rates

A "Hit" is defined as a Fidelity Score of 4. This shows the inability of the model to adapt to Syntactic and Cultural contexts, often resorting to word-for-word translation without the "flavour"

- **Morphological:** 20% (AI handles basic suffix swaps).
- **Syntactic/Cultural:** 0% (Total failure in structural adaptation).

5 Important Visual Lessons

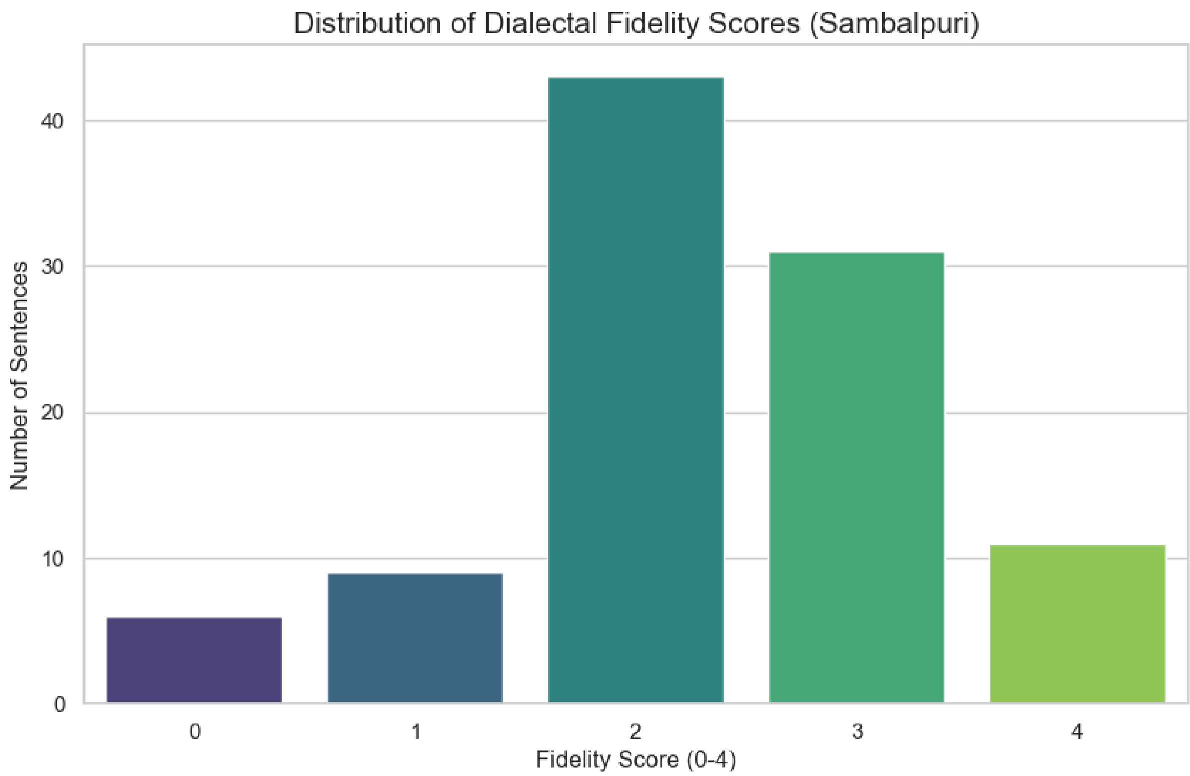


Figure 2: Distribution of Fidelity Scores

This graph shows that Fidelity Score has a maximum frequency in 2 and 3, showing mediocre performance. This reveals that while the model tries to translate, it seldom reaches the level of perfect native Sambalpuri. While the model is able to find the pronominal shift, and but fails to find the local term for day-to-day objects and verb conjugation.

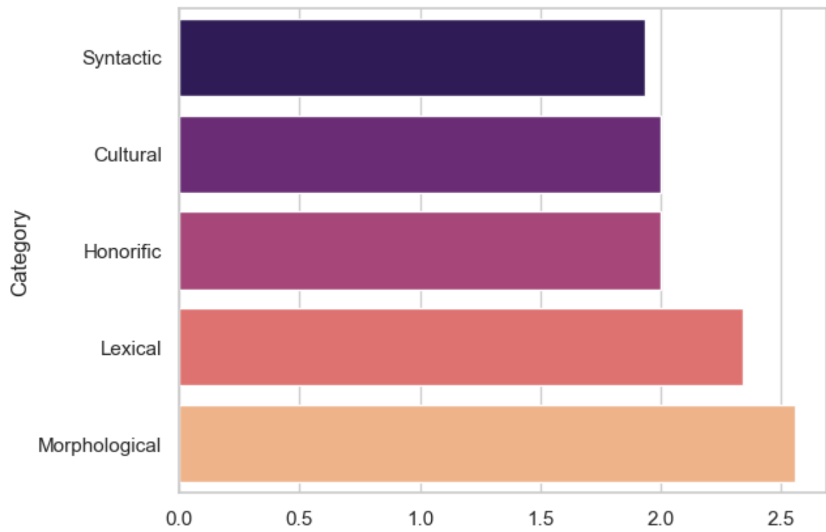


Figure 3: Fidelity Score by Category

The above graph tells us that the model performs worse in Syntactic, Cultural and Honorific categories compared to Lexical and Morphological ones. This shows that the model is not able to adapt to the syntactic and cultural nuances of the dialect. Also, a high morphological score reveals that the model is making fewer grammatical errors.

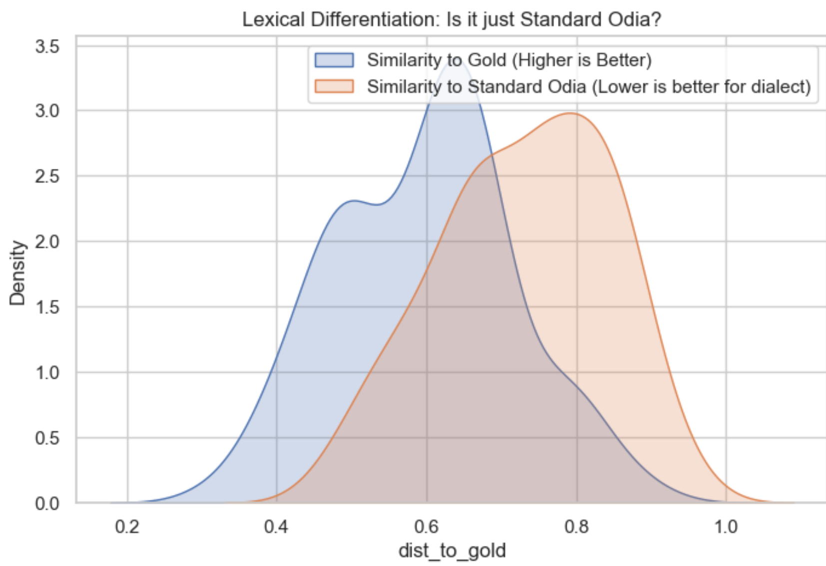


Figure 4: Lexical Differentiation

This is one of the most important findings. This analysis makes use of the Levenshtein ratio to check if the model just provides standard Odia output and labels it as Sambalpuri. The blue curve represents overlap between AI's output and the expert-annotated Sambalpuri ground truth. The orange curve represents how closely the AI's supposed Sambalpuri output resembles Standard Odia. The high density of the orange curve between the values 0.7-09 suggests that the model performs only surface-level changes and does not adapt to specific linguistic markers of the dialect. This is consistent with our earlier observations. The high overlap between both curves indicates the model plays safe and defaults to Odia Vocabulary to maintain fluency at the expense of dialectal fidelity.

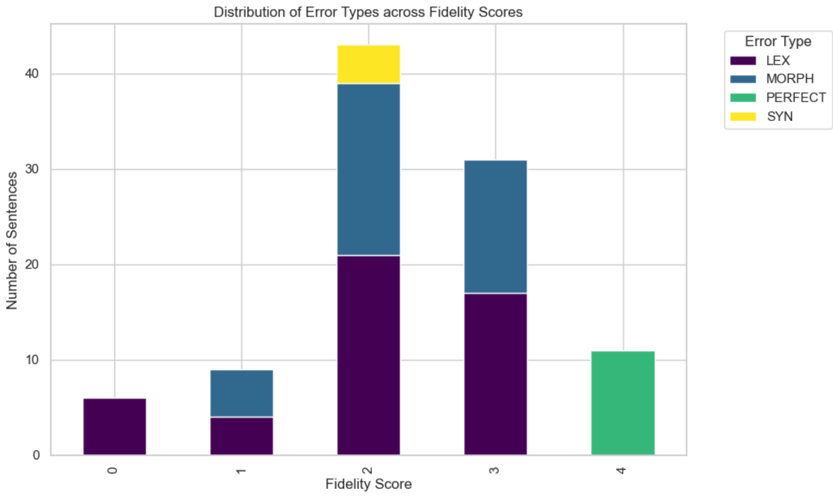


Figure 5: Distribution of Error Types across Fidelity Scores

The above graph shows that the majority of the errors are lexical, followed by the morphological ones. This proves that the model is unable to detect words native to the dialect and often uses standard Odia terms instead. Morphological error says that the model fails to frame sentences correctly. The 0 score is all Lexical error which indicates that the model fails to find the correct word while translating to Sambalpuri. We will see more examples in the qualitative analysis section.

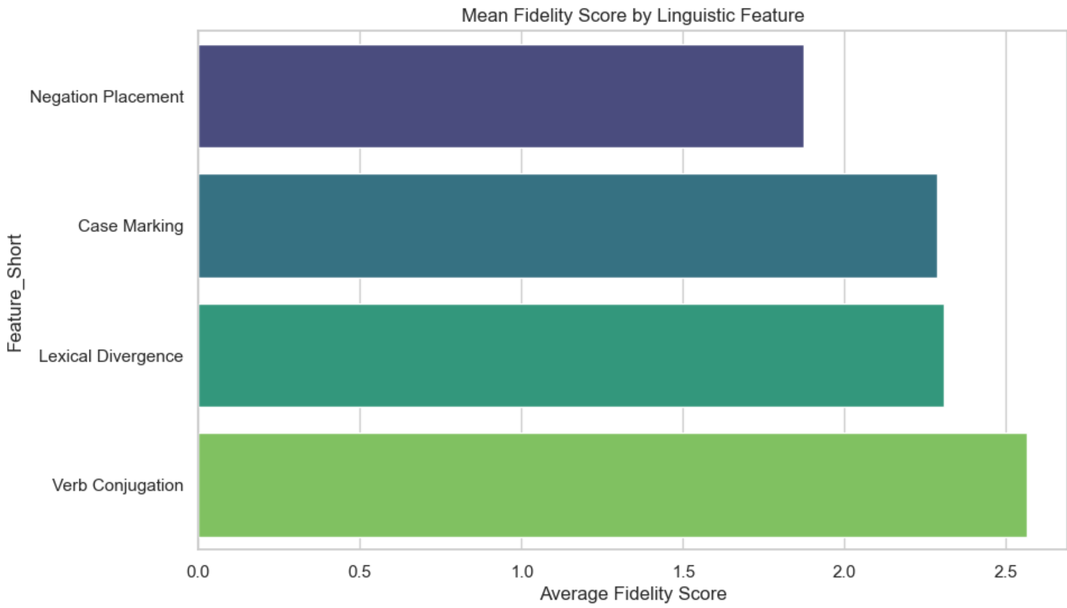


Figure 6: Mean Fidelity Score by Linguistic Feature

This graph shows that the model performs badly in Negation placement but excels in Verb Conjugation. ”No” is placed before the verb in Sambalpuri and after the verb in Odia. The model was unable to identify it and got it wrong in most cases. The models changed the verb suffixes correctly in the majority of cases, thus exhibiting a higher fidelity score. The model performs moderately case marking and Lexical Divergence.

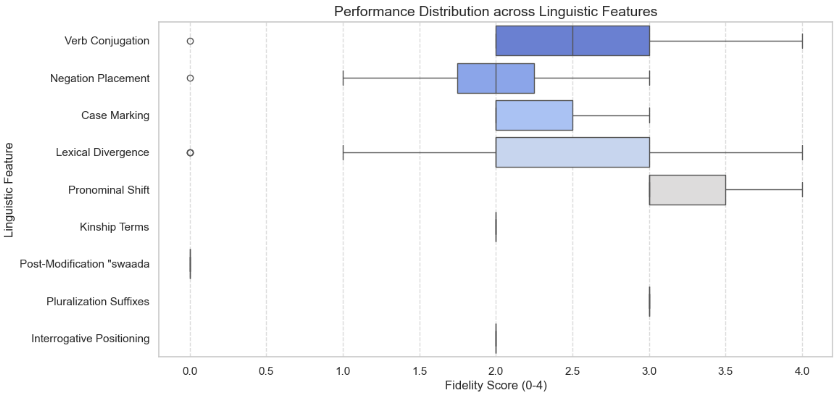


Figure 7: Performance Distribution across Linguistic Features

In this graph, we can see that the LLM excels in Pronominal Shift but offers quite a variation in Verb Conjugation and Lexical Devirgence. Thus it can perfectly convert pronouns like ”Mu” in Odia to ”Mui” in Sambalpuri.

6 Qualitative Analysis: The Hall of Shame

To contextualize the quantitative failures—specifically the 0% Syntactic hit rate—Table 2 documents instances where the AI-generated Sambalpuri failed completely (Fidelity Score 0). These examples highlight the ”Standardization Trap” where the model defaults to the high-resource parent language.

Standard Odia	AI-Generated (Model)	Gold (Sambalpuri)	Linguistic Diagnosis
ସୂର୍ଯ୍ୟୋଦୟ ପରେ ଆମେ କ୍ଷେତକୁ ଯାଉ ।	ଆମେ ସୁରୁଜ ଉଇଲା ପରେ ବିଲକେ ଯାଉଁ ।	ବେଲ୍ ଉଦ୍‌ଲେ ଆମେ କ୍ଷେତକେ ଯେସୁ ।	**Syntactic Leakage:** Failure to use dialectal verb form and incorrect lexical substitution.
ତେଲ କାହିଁକି ଏତେ ଦାମିକା?	ତେଲ ଏତେ ମହରଗ କାଏଁଗିରି?	ତେଲ୍ କେଁ ଯେ ଏତେ ମହଙ୍ଗା ।	**Morphological Failure:** Incorrect interrogative suffix; failed structure.
ଗାଈମାନଙ୍କ ପାଇଁ ଖାଦ୍ୟ ଆଣ ।	ଗୋରୁ ଲାଗି ଘାସ ଆଣା	ଗୋ ମାନ୍‌କର୍ ଲାଗି ଖାନା ଆନ୍ ।	**Semantic Shift:** Model hallucinated ”grass” (*ghasa*) instead of ”food” (*khana*).
ତୁମେ ଢୋଲ ଘୋଷଣା ଶୁଣିନାହଁକି?	ତୁଇ କାଏଁ ଢୋଲ ପିଟା ଶୁଣି ନାଇଁନ କାଏଁ?	ତୁମେ ଢୋଲ୍ ଘୋଷା ନେ ଶୁଣି କେ ?	**Honorific Failure:** Incorrect shift to informal and use of non-standard slang.
ସମୟ ଅପଚୟ କରିବା ବନ୍ଦ କର...	ବେଲା ନଷ୍ଟ ନାଇଁ କର...	ସମୟ ନେ ବର୍ବାଦ୍ କର୍...	**Standardization Bias:** Persistence of Standard Odia verb instead of dialectal .

Table 2: Hall of Shame: Qualitative Analysis of Fidelity Score 0 Instances.

7 Conclusion

While LLMs show promise in translation, they lack the structural ”depth” required for Sambalpuri. Future work should focus on syntactic-specific fine-tuning. A more rigorous approach is required to address people who speak diverse dialects. This research presents a comprehensive case study on how even the flagship LLM models can underperform on certain dialects.