Evaluating Nepali NER and POS Tagging on the Achhami Dialect

Samikshya Dhamala, Subresh Thakulla, Rishav Beejukchen, Bikash Kadayat, Supriya Khadka
Sunway College Kathmandu, Birmingham City University, Kathmandu, Nepal
{samikshya_25123833, subresh_23189651, rishav_24128439}@sunway.edu.np
{bikash_kadayat_a25, supriya}@sunway.edu.np

Abstract

Nepali Natural Language Processing (NLP) models are typically trained and tested on Standard Nepali, creating bias against regional dialects. This study examines Named Entity Recognition (NER) and Part-of-Speech (POS) tagging model performance on Achhami, a Far-Western dialect of Nepal. We constructed a parallel corpus of 300 sentence pairs covering news, cultural information, and conversations, with Achhami translations produced by native speakers to ensure linguistic authenticity. Our evaluation compared fine-tuned Transformer models (NepBertA, mBERT, and XLM-RoBERTa) against three large language models (GPT-4o Mini, Claude 3.5 Haiku, and Llama 3.1 70B) via zero-shot prompting. Across both tasks, all models showed consistent performance degradation on the Achhami dialect. For NER, F1 scores reduced by 2.12 to 3.97%. Claude 3.5 Haiku achieved the best overall NER results (89.10% F1 on Standard Nepali, 86.98% on Achhami), while unexpectedly, monolingual NepBertA outperformed multilingual mBERT, challenging assumptions about multilingual advantages. POS tagging results mirrored this disparity. While XLM-RoBERTa accuracy dropped from 78% on Standard Nepali to 72% on Achhami, the evaluated LLMs exhibited similar vulnerabilities, suffering accuracy degradations ranging from 2.9% to 7.0%. These findings quantify the "Kathmandu-centric" bias in Nepali NLP and demonstrate the need for dialectally diverse training data to ensure equitable language technology.

Keywords: Nepali NLP, Achhami Dialect, Named Entity Recognition, POS Tagging, Dialectal Robustness

1. Introduction

Nepali is an Indo-Aryan language belonging to the Indo-European family, believed to have developed from Sanskrit (Subba, 2009). Serving as the official language of Nepal, Nepali is also spoken in parts of India, Bhutan, and among the global Nepali diaspora. Nepal's complex topography, ranging from the Himalayan peaks to the Terai plains, has fostered substantial linguistic diversity. Despite its standardized written form, Nepali demonstrates significant dialectal variation shaped by geographical and socio-cultural factors (Clements and Khatiwada, 2015).

Nepali dialects are broadly categorized into three geographical zones: Eastern, Central, and Western varieties. Each exhibits distinct phonological, morphological, and lexical features (Rai, 2023). Eastern dialects, spoken in regions like Jhapa, Ilam, and Dhankuta, feature specific vowel qualities and retain certain archaic grammatical forms. These varieties show influence from neighboring Tibeto-Burman languages and possess distinctive intonation patterns (Pradhan, 2016). Central dialects, particularly the Kathmandu Valley variety, form the basis of Standard Nepali. This prestige dialect is used in education, media, and official communication, though notable micro-variations exist even within the Kathmandu Valley across Newar-influenced communities (Rai, 2023). Western dialects display the most substantial divergence from Standard Nepali (Prasad, 2026). A prominent example is the Achhami dialect, spoken primarily in the Achham district of western Nepal and in neighboring regions like Kailali due to migration. Achhami is classified alongside other western varieties such as Doteli, Bajhangi, Bajurali, Soradi, and Darchula (Prasad, 2026).

While largely similar to Standard Nepali, Achhami exhibits highly distinctive phonological and morphological features. For instance, the pronoun तिनी (/ti.ni/), meaning "he/she" in Standard Nepali, is adapted in Achhami as तिनु (/ti.nu/) to denote plural reference. Another distinction is the Standard Nepali possessive pronoun तिम्रो (/tim.ro/, meaning "your"), which corresponds to ताम्रो (/ta:m.ro/) in Achhami. In written contexts, this creates significant ambiguity. Standard readers might interpret ताम्रो (/ta:m.ro/) as "coppery" or "having copper-like elements," a meaning derived from the word तामा (/ta:.ma/, "copper"). Furthermore, Achhami frequently utilizes nasalization in pronominal forms, a feature often omitted in standard written representations (Aryal, 2026).

These dialectal characteristics present substantial challenges in computational contexts. Because natural language processing (NLP) systems are overwhelmingly trained on standardized language data, dialectal variation can degrade model performance in core sequence-labeling tasks such as Named Entity Recognition (NER) and Part-of-Speech (POS) tagging. NER depends on lexical and contextual cues that can shift under dialectal substitutions and ambiguous

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 210–221
11–12 May 2026.

orthography (Halpern, 2018), while POS tagging is additionally sensitive to morphosyntactic variation that affects inflectional patterns, pronominal paradigms, and token-level distributions (Moeller et al., 2021). To address this gap and systematically evaluate dialectal robustness in Nepali NLP, our work presents the following contributions:

- We introduce 300 Standard Nepali to Achhami sentence pairs, featuring native-speaker translations and complete NER and POS annotations across three domains.
- We provide quantitative evidence of F1 decrease across five NER models when evaluated on the Achhami dialect, and show parallel POS tagging degradation, indicating that dialectal mismatch affects both semantic labeling (NER) and syntactic labeling (POS) for Far-Western dialectal communities.
- We demonstrate that large language models outperform standard BERT architectures in dialectal robustness, and that monolingual models can exceed multilingual models in handling these specific regional variations.

2. Related Work

2.1. Dialectal Variation and Linguistic Features

Dialectal variation can manifest in orthography, phonology, morphology, syntax, and lexicon. Studies on English dialects demonstrate how regional constructions and syntactic patterns differ across varieties, such as Southern U.S. English, New Zealand English, and other regional forms (Joshi et al., 2025). In some cases, dialects of the same language may even approach mutual unintelligibility, illustrating the depth of variation that can occur within a single linguistic system. Examples include the relationship between Cantonese and Mandarin Chinese (Matthews and Yip, 2013) or between Bavarian and Standard German (Wiesinger, 2013).

For Nepali, descriptive studies report marked phonological and morphological differences between Western varieties (including Achhami and Doteli) and the Kathmandu-centered standard, while mutual intelligibility is generally maintained (Prasad, 2026; Khatiwada, 2009). These patterns suggest that dialectal variation in Nepali extends beyond lexical substitution and can affect grammatical cues that are central to automatic text processing.

2.2. Dialects and NLP Challenges

A consistent finding in dialectal NLP is that models trained on standardized varieties often transfer poorly to dialectal data. In dialectal Arabic NER studies, authors report notable performance drops when training data comes primarily from standardized forms (Khalifa et al., 2021). This kind of dialectal challenges manifest across diverse language families. In Chinese, NER systems face substantial difficulties with Cantonese compared to Mandarin, particularly for transliterated foreign names and location entities that use different character conventions (Nivre et al., 2020). NER models also show degradation on Swiss German due to different compound formation patterns and lexical items (Hollenstein and Aepli, 2014). Indian languages present particularly complex dialectal landscapes. Hindi NER systems show 8-15% degradation on regional varieties (Awadhi, Bhojpuri, Magahi) due to morphological differences in case marking and verb agreement (Sharma et al., 2025). These South Asian findings are especially relevant for Nepali, given typological similarities and shared Indo-Aryan heritage.

Computational research indicates that incorporating dialectal variation into model training can improve performance. Datta (2023) demonstrated that two-stage training and multi-task learning approaches improved English NER accuracy by approximately 3% and 1.2%, respectively. These results suggest that explicit modeling of dialectal diversity during training enhances generalization, though the magnitude of improvement depends on data availability, dialectal distance, and architectural choices. For low-resource languages like Nepali, where dialectal annotated data is virtually non-existent, understanding baseline patterns becomes essential before implementing data-intensive mitigation strategies.

2.3. NER and Dialect Sensitivity

NER identifies and classifies mentions of entities such as persons, locations, and organizations (Tedeschi et al., 2021). Dialectal variation can affect NER through orthographic inconsistencies and shifts in contextual cues that signal entity boundaries and types (Lin et al., 2024). Prior studies also indicate that entity types are not equally affected. Person entities can be particularly sensitive due to interactions with titles, honorifics, and surrounding morphosyntax, whereas location names and organizations often show greater orthographic stability across varieties (Benajiba and Rosso, 2008; Darwish, 2013; Khalifa et al., 2020; Hamed et al., 2025).

For Nepali, existing NER systems report roughly 75% to 85% F1 on Standard Nepali benchmarks

such as EverestNER (Niraula and Chapagain, 2022), typically using BERT-based architectures and neural sequence labeling approaches (Maharjan et al., 2019). Importantly, current datasets and evaluations largely focus on formal Standard Nepali from news and official sources, leaving dialectal robustness under-examined. Given documented Achhami differences in pronominal systems, pluralization, nasalization, and orthographic ambiguity, it is plausible that NER performance on Standard Nepali overestimates effectiveness for Far-Western dialect communities.

2.4. POS Tagging Evaluation

POS tagging assigns syntactic categories (for example, noun, verb, adjective) at the token level and is widely used as an upstream component in downstream NLP pipelines. In morphologically rich and variational settings, POS tagging can be especially sensitive to changes in inflectional patterns and function-word inventories, which are known to vary across Nepali dialects (Prasad, 2026; Khatiwada, 2009). For Nepali, prior POS work has explored both classical sequence models (for example, HMM-based approaches) and neural architectures, including deep learning models and more recent transformer-based methods (Yajnik, 2014; Paul et al., 2015; Prabha et al., 2018; Abdalazeim and Meziane, 2021; Pradhan and Yajnik, 2024).

Despite this progress, POS tagging research, like NER, has primarily been evaluated on Standard Nepali data. As a result, it remains unclear whether reported accuracies reflect performance under dialectal variation. By evaluating POS tagging alongside NER on parallel Standard Nepali and Achhami text, our study positions POS as a complementary diagnostic: NER probes robustness in semantically grounded labeling, while POS probes robustness in morphosyntactic labeling under dialect shift.

3. Methodology

In this section, we first describe the construction of a 300-sentence parallel corpus and the translation and quality-assurance process. We then outline the NER and POS annotation scheme and dataset format, followed by the model evaluation setup used to compare performance across Standard Nepali and Achhami. Our overall system flowchart is given in Figure 1.

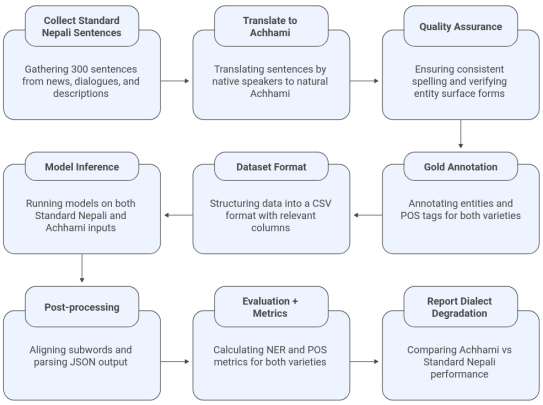

Figure 1: Overall System Flowchart

3.1. Dataset Creation

To support this evaluation, we constructed a parallel dataset of 300 sentence pairs[1]. Each instance contains a Standard Nepali sentence typical of formal media and education, paired with a natural Achhami translation reflecting daily spoken usage.

3.1.1. Source Data Collection

To ensure domain diversity and ecological validity, the 300 Standard Nepali base sentences were collected from three contexts. First, 100 sentences were extracted from recent news articles on politics, sports, and local events from Nepali news portals such as Kantipur[2] and OnlineKhabar[3]. Second, 100 sentences were constructed as conversational dialogues, including greetings, daily planning, and casual discussions. Third, 100 descriptive sentences covered locations, cultural traditions, and festivals, resembling encyclopedic or travel writing.

All Achhami translations were produced by two native speakers of the Achhami dialect, who are also fluent in Standard Nepali. The translation process emphasized natural, contextually appropriate dialectal usage rather than literal word-for-word mapping. Unlike English, where dialectal variation often includes clear lexical markers or "telltale" forms (e.g., capitalization for named-entities), Nepali dialectal variation particularly in Achhami is more subtle and frequently involves shifts in morphology, phonology, and auxiliary usage rather than easily identifiable lexical substitutions.

To ensure consistency and quality, both translators independently worked on portions of the dataset and then conducted cross-review and joint

[1] Dataset is publicly available on https://github.com/samikshyadhamala/NER-and-POS.git

[2] https://ekantipur.com/

[3] https://www.onlinekhabar.com/

discussions to resolve discrepancies. This process ensured that the final Achhami sentences reflect agreed upon, natural dialectal forms without artificial standard language influence. For NER, we ensured that the majority of sentences contain at least one named entity (Person, Location, or Organization). For POS tagging, we additionally aimed for diverse distributions of syntactic categories, including nouns, verbs, adjectives, adpositions, and auxiliary elements, so that tagging performance could be assessed across varied constructions in both Standard Nepali and Achhami. All gold annotations for both NER and POS tagging were manually created and verified by the authors.

3.1.2. Dialect Translation Process

Two of the authors, who are also native Achhami speakers, translated the dataset. Rather than literal, word-for-word translation, the focus was on capturing authentic, conversational Achhami phrasing. Key linguistic adaptations included utilizing localized verbal morphology and auxiliary verbs, such as replacing जान्छु (/d͡zan.t͡sʰu/, *jaanchu*), meaning "I go" / "I will go" in Standard Nepali) with जान्या हु (/dzan.ja hu/, *jaanya hu*).

Translations also incorporated dialectal suffixes such as -कौ (/kau/, *-kau*), -ऐक (/aik/, *-aik*), and -रैछ (/rʌi.t͡sʰʌ/, *-raichha*), and regionally appropriate lexical choices, like भोल्या (/bʰo.lja/, *bholya*) for "tomorrow." The workload was split evenly, with each translator producing 150 Achhami sentences.

3.1.3. Quality Assurance and Annotation

After translation, the translators cross-reviewed each other's sentences to ensure that the Achhami text did not read as minimally modified Standard Nepali. Because Achhami lacks a standardized orthography, the reviewers also agreed on consistent spellings for frequently occurring dialectal vocabulary. During quality assurance, the annotators verified that named entities remained identifiable under dialectal morphology, including suffixation. For example, the city धनगढी (/dʰʌn.gʌ.ɖʰi/, *Dhangadi*) was checked for correct recognition when realized as धनगढी-कौ (/dʰʌn.gʌ.ɖʰi kau/, *Dhangadi-kau*). Discrepancies were resolved via consensus.

The finalized dataset was compiled as a CSV with fields `ID`, `Category`, `Standard_Nepali`, `Dialect_Text`, `Entities`. The `Entities` field contains manually verified ground-truth NER annotations formatted as Entity (Type), for example धनगढी (*Dhangadi*) (LOC) or राम बहादुर (*Ram Bahadur*) (PER), with notes specifying the surface morphological form in each sentence version. The dataset also includes a `POS_Tags` field with token-level POS annotations for both Standard Nepali and Achhami, using the UD tagset. Each token is assigned a UD category such as NOUN, PROPN, VERB, AUX, ADP, ADJ, or PUNCT, enabling fine-grained morphosyntactic evaluation across the two varieties.

3.2. Model Selection

To evaluate dialectal robustness for Nepali NER and POS tagging, we selected models from two paradigms: fine-tuned transformer encoders and general-purpose large language models (LLMs) evaluated via zero-shot prompting (Table 1). This design enables comparison between monolingual and multilingual encoders, as well as between task-specific fine-tuning and instruction-following LLM behavior under dialect shift.

For NER, we evaluated two fine-tuned transformer models: (i) NepBERTa (fine-tuned for Nepali NER) (Gautam et al., 2022; SynapseHQ, 2026), a monolingual model trained on Nepali text, and (ii) an mBERT-based Nepali NER model[4], representing multilingual pretraining with Nepali task fine-tuning. We additionally evaluated three LLMs using zero-shot prompting via the OpenRouter[5] API: GPT-4o Mini, Claude 3.5 Haiku, and Llama 3.1 70B Instruct.

For POS tagging, we used XLM-RoBERTa fine-tuned for UD POS tagging[6] as the transformer baseline, and evaluated the same three LLMs (GPT-4o Mini, Claude 3.5 Haiku, and Llama 3.1 70B Instruct) with zero-shot prompting to enable direct cross-paradigm comparison on both Standard Nepali and Achhami.

Model	Task	Paradigm
NepBERTa (fine-tuned)	NER	Transformer
mBERT Nepali NER	NER	Transformer
XLM-RoBERTa UD-POS	POS	Transformer
GPT-4o Mini	NER + POS	LLM
Claude 3.5 Haiku	NER + POS	LLM
Llama 3.1 70B	NER + POS	LLM

Table 1: Models used for NER and POS evaluation.

3.3. Experimental Setup

We evaluated all models under two conditions to quantify dialectal degradation for both NER and POS tagging. First, models were evaluated on the 300 Standard Nepali sentences to establish a baseline against gold annotations. Second, the same procedure was applied to the Achhami translations. For NER, we used the same gold entity

[4] mbert-Nepali-NER on Hugging Face
[5] https://openrouter.ai/
[6] xlm-roberta-base-ft-udpos28-en on HuggingFace

labels, with minor updates only when an entity's surface form changed due to dialectal morphology or orthographic conventions. For POS tagging, we used gold UD token-level annotations for both varieties, since the underlying morphosyntactic categories are shared even when surface forms differ.

3.3.1. Traditional Transformer Models (BERT-based)

For the fine-tuned transformer encoders used in NER (NepBERTa and mBERT-Nepali-NER) and POS tagging (XLM-RoBERTa UD-POS), we performed standard token-classification inference. Because these models use subword tokenization, predictions were aligned back to the original word tokens to ensure that each token received a single NER or POS label. For NER, model outputs were mapped to the dataset label inventory using the BIO tagging scheme. For POS tagging, outputs were mapped to the UD tagset to enable consistent evaluation across Standard Nepali and Achhami.

3.3.2. LLMs and Prompting Strategy

For the LLM-based models (GPT-4o Mini, Claude 3.5 Haiku, and Llama 3.1 70B), we used prompting via the OpenRouter API for both tasks and evaluated the same prompts on Standard Nepali and Achhami inputs.

For NER, we used a standardized few-shot prompt that defined the task and required a strict JSON output, extracting entities verbatim and assigning them to the categories `PER`, `LOC`, `ORG`, `DATE`, and `MISC`. For POS tagging, we used a zero-shot prompt that required a strict JSON array of UD POS tags, with the number of output tags exactly matching the number of input tokens and restricted to the 17 UD coarse-grained categories. For Achhami inputs, the prompt additionally stated that the sentence is in the Achhami dialect of far-western Nepal and that UD tagging rules should be applied accordingly. We report the full prompts in Appendix A.1 (NER) and Appendix A.2 (POS).

3.3.3. Evaluation Methodology

For NER, we report entity-level Precision, Recall, and F1 using strict matching, requiring correct entity boundaries and entity type. For POS tagging, we report token-level Accuracy and label-based F1 variants. For both tasks, we additionally compute token-level Accuracy, Macro Macro F1, and Weighted F1 to facilitate cross-model comparison. We use Macro F1 as the primary summary metric because it weights frequent and rare labels equally, and we quantify dialectal degradation by comparing Achhami scores against the Standard Nepali baseline for each model and task.

4. Results

4.1. Evaluating Dialectal Robustness

We evaluated five NER models (two fine-tuned transformer encoders and three LLMs) and four POS tagging models (one fine-tuned transformer encoder and three LLMs) on both Standard Nepali and Achhami. Table 2 reports performance degradation when transitioning from Standard Nepali to Achhami, which we use as the primary measure of dialectal robustness across both tasks.

For NER, Claude 3.5 Haiku showed the strongest robustness, with only a 2.12% F1 drop, while also achieving the best absolute performance on both Standard Nepali (89.10) and Achhami (86.98). NepBERTa, despite having the lowest Standard Nepali baseline among the NER models, exhibited comparatively strong robustness (2.64 drop), outperforming GPT-4o Mini (3.56 drop) and mBERT-Nepali (3.97 drop) in terms of degradation. Notably, mBERT-Nepali had the largest NER degradation, indicating that multilingual pre-training alone does not guarantee robustness to intra-language dialectal variation.

For POS tagging, degradation was larger overall. Claude 3.5 Haiku again degraded least (2.91 drop), whereas GPT-4o Mini and Llama 3.1 70B showed larger drops of 7.02 and 5.29, respectively, suggesting greater sensitivity to Achhami morphosyntactic differences. The XLM-RoBERTa UD-POS model experienced the largest degradation across all model-task combinations (14.49 drop), despite being a POS-specific model fine-tuned on multilingual UD treebanks. Overall, the results show consistent performance loss on Achhami for both tasks and indicate that broad-coverage LLMs can be comparatively more robust than specialized transformer taggers in this low-resource dialect setting.

4.2. Degradation Patterns

Examining metric-specific degradation reveals distinct patterns for NER and POS (Figures 2 and 3). For NER, recall shows the smallest degradation across models (1.76% to 3.37%), suggesting that models continue to identify many entity mentions despite dialectal variation. Precision degrades more strongly (2.46% to 4.94%), indicating increased false positives and boundary or type mismatches on Achhami text. Accuracy shows larger drops for the LLMs (3.39% to 5.33%), reflecting cumulative token-level errors across entity and non-entity labels. Claude 3.5 Haiku exhibits the smallest recall drop (1.76%), while GPT-4o Mini shows the largest precision drop (4.94%).

For POS tagging, degradation patterns differ. Recall exhibits the largest drops (7.40% to

Model	Task	F1 (Std. Nepali)	F1 (Achhhami)	Drop (%)
Claude 3.5 Haiku	NER	89.10	86.98	2.12
GPT-4o Mini	NER	86.21	82.65	3.56
Llama 3.1 70B	NER	83.48	80.98	2.51
mBERT-Nepali	NER	66.67	62.70	**3.97**
NepBERTa	NER	64.24	62.60	2.64
Claude 3.5 Haiku	POS	75.13	72.22	2.91
GPT-4o Mini	POS	77.14	70.12	7.02
Llama 3.1 70B	POS	57.04	51.76	5.29
XLM-RoBERTa UD	POS	64.18	49.69	**14.49**

Table 2: Performance degradation (F1 Score) across models for NER and POS tagging. Bold values in the Drop (%) column highlight the worst (highest) performance degradation.

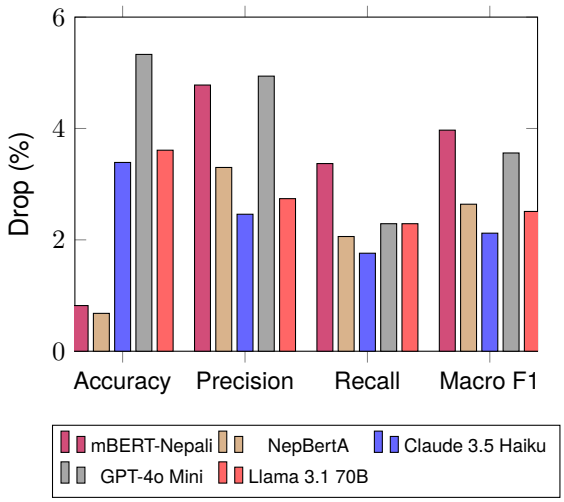

Figure 2: NER Percentage degradation across evaluation metrics for all the models.

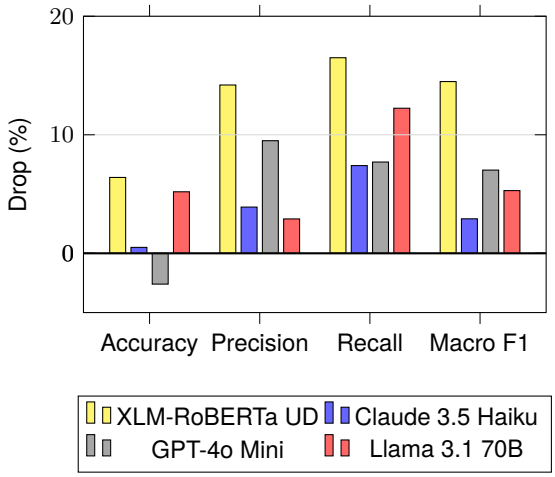

Figure 3: POS Tagging Percentage degradation across evaluation metrics from Standard Nepali to Achhami dialect for POS tagging models. Negative values indicate improvement on Achhami over Standard Nepali.

16.50%), indicating that correct tagging becomes more difficult under dialectal morphology and lexical variation. Precision drops range from 2.90% to 14.20%, and accuracy drops are comparatively smaller for the LLMs (0.50% to 5.19%) but remain substantial for XLM-RoBERTa UD (6.40%). Notably, GPT-4o Mini shows a negative accuracy drop (-2.60%), meaning its token-level accuracy slightly improves on Achhami, even though its Macro F1 decreases by 7.02%. This mismatch indicates that accuracy can be influenced by shifts in tag distribution, while Macro F1 is more sensitive to errors that affect less frequent tags and harder distinctions.

4.3. Error Pattern Analysis

To identify likely sources of degradation, we analyzed confusion matrices and compared systematic error patterns across model types and dialect conditions. For NER, the transformer-based BERT models showed substantial confusion between `ORG` and `LOC`, with a 15% `ORG`→`LOC` misclassification rate on Standard Nepali, alongside high false negative rates of around 40% across entity types. On Achhami, these issues intensified: `ORG`→`LOC` confusion increased to 23%, false negatives for `PER` increased by 12%, and a `PER`→`MISC` confusion emerged that was not observed on Standard Nepali. These shifts are consistent with the hypothesis that dialectal morphology and orthographic variation alter local contextual cues and token shapes that the models rely on for boundary detection and entity typing.

For POS tagging, the XLM-RoBERTa UD model showed the most pronounced and structured confusions. On Standard Nepali, its dominant errors were between `AUX` and `VERB`, reflecting ambiguity in Nepali auxiliary constructions. On Achhami, these confusions increased, and additional errors appeared for function-word categories, with `PART` often misclassified as `ADP` or `SCONJ`. This pattern aligns with Achhami's divergence in auxiliary and particle usage. The absence of Nepali or closely related varieties in the model's UD fine-tuning data is a plausible contributor to these generalization failures under dialect shift.

In contrast, LLM error patterns were comparatively stable across dialects for both tasks. For NER, the main recurring LLM confusion on Standard Nepali involved PER→ORG errors (7% to 9%) in contexts with ambiguous institutional titles. On Achhami, confusion patterns were largely similar, with only modest increases in false positives (3% to 4%), suggesting occasional over-prediction of entities rather than systematic type collapse. For POS tagging, LLM errors were concentrated in low-frequency tags such as PART, DET, and SCONJ, where performance remained weak in both varieties, likely due to limited support in the evaluation set. Across both dialects, AUX remained a consistent difficulty; for example, Claude 3.5 Haiku achieved AUX recall of 0.61 on Standard Nepali and 0.63 on Achhami, suggesting that auxiliary tagging errors stem more from structural ambiguity than from dialect-specific forms. Overall, these confusion-matrix analyses indicate that dialect shift amplifies boundary and type errors for transformer taggers, while LLM degradations are smaller and appear driven primarily by ambiguous contexts in NER and sparse POS categories.

5. Discussion

Our benchmarking of five NER models and four POS tagging models on Achhami shows consistent performance degradation relative to Standard Nepali across architectures and both tasks. For NER, F1 decreases by 2.12 to 3.97 percentage points. For POS tagging, Macro F1 drops are larger, ranging from 2.91 to 14.49 percentage points. This gap suggests that dialect shift affects both semantic labeling (NER) and morphosyntactic labeling (POS), with the impact being more pronounced for token-level syntactic categorization. Across both tasks, LLMs exhibit stronger dialectal robustness than fine-tuned transformer taggers. Claude 3.5 Haiku shows the smallest degradation for both NER (2.12) and POS tagging (2.91), indicating that broad multilingual pre-training at scale may yield more stable representations under dialectal variation than task-specific fine-tuning on standard-language data alone.

5.1. Findings and Interpretation

5.1.1. Architectural Factors for Robustness

Claude 3.5 Haiku's strong performance (89.10 F1 on Standard Nepali and 86.98 on Achhami, corresponding to a 2.12 drop) is consistent with several architectural and data-scale factors. As a large-scale model, it may have been exposed during pre-training to diverse registers and non-standard orthography that partially resemble dialectal variation. In addition, broader contextual modeling can help infer entity boundaries and types from sentential semantics even when surface forms differ from Standard Nepali. Multilingual exposure may further support abstraction over form, allowing entities and syntactic roles to be recognized via higher-level contextual regularities rather than only dialect-specific orthographic patterns.

The smaller recall degradation (1.76) compared to the precision drop (2.46) suggests that Claude largely preserves its ability to detect entity mentions in Achhami, while making more boundary or type errors in dialectal contexts. This aligns with the view that LLM robustness is primarily limited by ambiguity and boundary decisions rather than a failure to locate entity-bearing spans. A similar trend is visible in POS tagging, where Claude's comparatively small Macro F1 drop (2.91) suggests better stability under morphosyntactic variation than the transformer POS tagger.

5.1.2. The Monolingual Advantage

The finding that monolingual NepBERTa exhibits smaller NER degradation (2.64) than multilingual mBERT (3.97) complicates the common assumption that multilingual models are inherently more robust to linguistic variation. Three non-exclusive factors may explain this pattern.

First, pre-training data composition may matter more than multilinguality per se. NepBERTa's Nepali-only pre-training corpus may have included more informal or regionally diverse Nepali than the Nepali subset encountered by mBERT, yielding representations that better tolerate non-standard forms. Second, multilingual representation learning can introduce cross-lingual constraints that help transfer across languages but may reduce flexibility for within-language variation if dialectal forms are not well represented. Third, fine-tuning on Standard Nepali NER data may interact differently with monolingual versus multilingual initialization. In particular, mBERT may more strongly converge to standard orthographic and contextual cues during fine-tuning, whereas a monolingual model may retain broader within-language variability. Overall, these results suggest that, for Nepali dialect robustness, curated monolingual pre-training on diverse sources may be at least as important as multilingual pre-training, and that dialectal evaluation should be reported alongside Standard Nepali benchmarks for both NER and POS tagging.

5.2. Design Implications

5.2.1. Rethinking Evaluation Practices

Our results show that evaluation limited to Standard Nepali benchmarks can overestimate real-

world performance for dialectal populations. For example, a model achieving 89% F1 on Standard Nepali NER drops to 87% on Achhami, and POS tagging shows even larger sensitivity, with Macro F1 drops ranging up to 14.49 percentage points. Although some degradations appear small in absolute terms, they can compound in practical deployments, increasing missed entities, spurious extractions, and incorrect syntactic analyses. We therefore advocate dialectally stratified evaluation as a standard practice in Nepali NLP. Future model releases should report performance across multiple Nepali varieties, including at minimum Standard Nepali and Far-Western varieties, and ideally additional regional groupings when data is available, to provide transparent robustness profiles for both NER and POS tagging.

5.2.2. Training Data Collection Priorities

The competitive dialectal robustness of the monolingual NepBERTa relative to mBERT suggests that pre-training data composition and diversity may be as important as multilingual architecture. Future Nepali model development should prioritize three areas. First, researchers should build more dialectally diverse pre-training corpora by collecting text from regional outlets, social media from dialectal regions, and transcribed speech, rather than relying primarily on Kathmandu-centered formal sources. Second, the community should expand annotation efforts to include dialectal NER and POS labels, since these tasks are affected differently by dialect shift and benefit from explicit supervision. Finally, data collection should be community-driven, with partnerships in Far-Western and other regions to ensure authentic representation and to avoid forcing dialectal content into Standard Nepali orthographic norms.

5.2.3. Model Selection Guidelines for Practitioners

For deployments in dialectally diverse settings, our results suggest trade-offs between accuracy and robustness across both NER and POS. In higher-resource scenarios, practitioners can consider large-scale LLMs such as Claude 3.5 Haiku, which showed the smallest degradation in both tasks in our evaluation. In more resource-constrained scenarios, monolingual models such as NepBERTa may be preferable to multilingual alternatives when dialectal robustness is critical, particularly for NER. For high-stakes applications, including healthcare or legal services, human-in-the-loop validation remains important, since dialectal variation can affect both entity extraction and downstream syntactic analyses that depend on POS tags.

5.2.4. Addressing Kathmandu-Centric Bias

Our quantitative evidence of consistent performance degradation on Achhami for both NER and POS tagging supports concerns about Kathmandu-centric evaluation and development practices in Nepali language technology. This bias is not only technical; it can contribute to unequal utility of NLP systems across regions when Standard Nepali performance is treated as representative. To reduce this inequity, research funding and public-sector AI initiatives should explicitly incentivize dialectal inclusivity and evaluation beyond Standard Nepali. In addition, venues and service providers should encourage reporting of dialectal performance profiles for core tasks, including both NER and POS tagging, and promote clear documentation of known limitations and targeted robustness goals.

6. Conclusion

We present the first systematic benchmark of dialectal robustness in Nepali for both Named Entity Recognition and Part-of-Speech tagging by evaluating nine model configurations on Standard Nepali and Achhami. All models degrade on Achhami. For NER, F1 drops by 2.12 to 3.97 percentage points across five models. For POS tagging, degradation is larger, with Macro F1 drops ranging from 2.91 to 14.49 percentage points, indicating that Standard Nepali bias can be more severe for syntactic labeling than for entity recognition. Claude 3.5 Haiku is the most robust model in both tasks (2.12 drop for NER and 2.91 for POS), and LLMs are generally more robust than the evaluated fine-tuned transformer taggers in this setting. We also find that monolingual NepBERTa is more robust than multilingual mBERT for NER (2.64 vs. 3.97 drop), suggesting that multilingual pre-training does not inherently yield dialectal robustness. For POS tagging, XLM-RoBERTa UD shows the largest drop (14.49), reinforcing that existing POS resources and training data do not adequately cover Nepali dialectal variation. Overall, our results quantify dialect-driven performance gaps in Nepali NLP across both NER and POS tagging and motivate dialect-inclusive data collection, dialect-stratified evaluation, and the development of morphosyntactic resources that better represent Nepal's linguistic diversity.

7. Limitations

This study has several limitations. First, we focus on Achhami, so the results may not generalize to other Nepali varieties with different contact influences and morphosyntactic properties. Second,

our dataset is small (300 parallel sentence pairs), which limits statistical power and increases uncertainty in model comparisons, particularly when differences are a few percentage points. This size also restricts fine-grained analysis of rare NER classes (for example, `DATE` and `MISC`) and yields sparse support for several UD POS categories (for example, `PART`, `DET`, `SCONJ`, and `INTJ`), which can make Macro F1 sensitive to label sparsity. Third, because the Achhami data is translation-based, it may underrepresent naturally occurring dialect phenomena such as code-switching and spontaneous syntactic constructions, which could affect POS tagging in particular. Fourth, our POS transformer baseline (XLM-RoBERTa UD) was not fine-tuned on Nepali UD data, so its performance reflects a mix of cross-lingual transfer and dialect shift rather than dialect robustness alone. Finally, LLM results depend on prompting choices and structured output reliability, and alternative prompt designs may yield different robustness profiles.

8. Ethical Considerations

We constructed our Achhami-Nepali NER and POS dataset from publicly accessible news articles. All texts were already in the public domain, and no private or sensitive user data were collected.During annotation, we removed any explicit personal identifiers beyond named entities required for the NER task. POS annotation was performed solely on linguistic structure and contains no personally identifiable information. The dataset is intended solely for research and educational purposes.

9. Data/Code Availability Statement

All the data and code supporting the findings of this study is publicly available on GitHub.

10. References

Alaa Abdalazeim and Farid Meziane. 2021. A review of the generation of requirements specification in natural language using objects uml models and domain ontology. *Procedia Computer Science*, 189:328–334.

Parma Nanda Aryal. 2026. Pronominal in english, nepali and achhami.

Yassine Benajiba and Paolo Rosso. 2008. Arabic named entity recognition using conditional random fields. In *Proc. of Workshop on HLT & NLP within the Arabic World, LREC*, volume 8, pages 143–153.

N Clements and Rajesh Khatiwada. 2015. Cooccurrence constraints on aspirates in nepali. *Features in Phonology and Phonetics: Posthumous Writings by Nick Clements and coauthors*.

Kareem Darwish. 2013. Named entity recognition using cross-lingual resources: Arabic as an example. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1558–1567.

Samiha Datta. 2023. Investigating english-language dialect-adjusted models.

Milan Gautam, Sulav Timilsina, and Binod Bhattarai. 2022. Nepberta: Nepali language model trained in a large corpus. 2:273–284.

Jack Halpern. 2018. Very large-scale lexical resources to enhance chinese and japanese machine translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Injy Hamed, Caroline Sabty, Slim Abdennadher, Ngoc Thang Vu, Thamar Solorio, and Nizar Habash. 2025. A survey of code-switched arabic nlp: Progress, challenges, and future directions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4561–4585.

Nora Hollenstein and Noëmi Aepli. 2014. Compilation of a swiss german dialect corpus and its application to pos tagging. In *Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects*, pages 85–94.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Computing Surveys*.

Muhammad Khalifa, Hesham Hassan, and Aly Fahmy. 2021. Zero-resource multi-dialectal arabic natural language understanding.

Salam Khalifa, Nasser Zalmout, and Nizar Habash. 2020. Morphological analysis and disambiguation for gulf arabic: The interplay between resources and methods. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3895–3904.

Rajesh Khatiwada. 2009. Nepali. *Journal of the International Phonetic Association*, 39:373–380.

Pin-Jie Lin, Merel Scholman, Muhammed Saeed, and Vera Demberg. 2024. Modeling orthographic variation improves nlp performance for

nigerian pidgin. *ACL Anthology*, pages 11510–11522.

Gopal Maharjan, Bal Krishna Bal, and Santosh Regmi. 2019. Named entity recognition (ner) for nepali. In *Conference on Creativity in Intelligent Technologies and Data Science*, pages 71–80. Springer.

Stephen Matthews and Virginia Yip. 2013. *Cantonese: A comprehensive grammar*. Routledge.

Sarah Moeller, Ling Liu, and Mans Hulden. 2021. To pos tag or not to pos tag: The impact of pos tags on morphological learning in low-resource settings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 966–978.

Nobal Niraula and Jeevan Chapagain. 2022. Named entity recognition for nepali: data sets and algorithms. In *The International FLAIRS Conference Proceedings*, volume 35.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajic, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4034–4043.

Abhijit Paul, Bipul Syam Purkayastha, and Sunita Sarkar. 2015. Hidden markov model based part of speech tagging for nepali language. In *2015 international symposium on advanced computing and communication (ISACC)*, pages 149–156. IEEE.

Greeshma Prabha, PV Jyothsna, KK Shahina, B Premjith, and KP Soman. 2018. A deep learning approach for part-of-speech tagging in nepali language. In *2018 international conference on advances in computing, communications and informatics (ICACCI)*, pages 1132–1136. IEEE.

Ashish Pradhan and Archit Yajnik. 2024. Parts-of-speech tagging of nepali texts with bidirectional lstm, conditional random fields and hmm. *Multimedia Tools and Applications*, 83(4):9893–9909.

Uma Pradhan. 2016. *Ethnicity, equality, and education: a study of multilingual education in Nepal*. Ph.D. thesis, University of Oxford.

Upadhayaya Gopal Prasad. 2026. The subject-verb agreement in the english language and achhami dialect.

Sushma Rai. 2023. *Language and the Question of Identity: A Study of the Nepali Language Movement in India (1956-1992)*. Ph.D. thesis.

Shalini Sharma, Piyush P Singh, et al. 2025. Named entity recognition for hindi current landscape and emerging trends. *Journal of Information Technology, Cybersecurity, and Artificial Intelligence*, 2(2):133–144.

Tanka Bahadur Subba. 2009. *Indian Nepalis: issues and perspectives*. concept publishing company.

SynapseHQ. 2026. Securly ai chat.

Simone Tedeschi, Simone Conia, Francesco Cecconi, and Roberto Navigli. 2021. Named entity recognition for entity linking: What works and what's next.

Peter Wiesinger. 2013. The central and southern bavarian dialects in bavaria and austria. In *The dialects of modern German*, pages 438–519. Routledge.

Archit Yajnik. 2014. Ann based pos tagging for nepali text. *International Journal of Natural Language Computing (IJNLC)*, 3(3). Open access via publisher site.

Appendix A: LLM Prompt Template

A.1 NER model

The complete prompt template provided to the Large Language Models consisted of the following role definition, operational rules, and three-shot examples.

Role and Task Definition

You are a Named Entity Recognition expert for the Nepali language. Extract ALL named entities from this text and classify them: PER (people), LOC (places), ORG (organizations or institutions), DATE (dates or times), and MISC (other named entities).

Operational Rules

Only include entities appearing verbatim in the text; do not guess or hallucinate. If no entities are present, return `{"entities": []}`. Output must be valid JSON on a single line without markdown formatting or code fences.

Few-Shot Examples

We provided three demonstration examples to establish expected behavior.

Prompt Format: Example 1

Input: प्रधानमन्त्री केपी शर्मा ओलीले सिंहदरबारमा नयाँ आर्थिक नीति प्रस्तुत गरे।
Output: `{"entities": [{"text": "केपी शर्मा ओली", "type": "PER"}, {"text": "सिंहदरबार", "type": "LOC"}]}`

Prompt Format: Example 2

Input: राम र सीता काठमाडौँस्थित त्रिभुवन अन्तर्राष्ट्रिय विमानस्थल गए।
Output: `{"entities": [{"text": "राम", "type": "PER"}, {"text": "सीता", "type": "PER"}, {"text": "काठमाडौँ", "type": "LOC"}, {"text": "त्रिभुवन अन्तर्राष्ट्रिय विमानस्थल", "type": "LOC"}]}`

Prompt Format: Example 3

Input: रमाइलो मेलामा जनकपुर नगरपालिका र अञ्चल अस्पतालले स्टल राखे।
Output: `{"entities": [{"text": "जनकपुर नगरपालिका", "type": "ORG"}, {"text": "अञ्चल अस्पताल", "type": "ORG"}]}`

A.2 POS models

We used a zero-shot prompt for POS tagging, because the task requires deterministic token-level labeling with a fixed tag inventory. The prompt consists of (i) a role and task definition, (ii) an optional dialect note for Achhami inputs, and (iii) operational rules enforcing a strict JSON output format.

Role and Task Definition You are a Nepali linguistics expert for UD POS tagging. Given a sequence of tokens, assign exactly one UD POS tag to each token.

Dialect-Specific Note (Achhami only) NOTE: Tokens are in the Achhami dialect spoken in the Achham and Bajhang districts of far-western Nepal. Apply UD POS rules accordingly for this regional variety.

Operational Rules

1. Return only a JSON array containing exactly N POS tags, where N equals the number of input tokens.
2. Use only the following 17 UD tags: `NOUN`, `PROPN`, `VERB`, `AUX`, `ADJ`, `ADV`, `ADP`, `PRON`, `DET`, `NUM`, `CCONJ`, `SCONJ`, `PART`, `PUNCT`, `INTJ`, `SYM`, `X`.
3. Always tag punctuation marks such as `,` , `.`, `?`, and `!` as `PUNCT`.
4. Do not skip any token. The number of output tags must equal the number of input tokens.
5. Do not include any explanation, markdown formatting, or code fences.
6. Output format: `["TAG1","TAG2",…,"TAGN"]`.

Prompt Template

POS Prompt Template: Standard Nepali

Role: You are a Nepali linguistics expert for UD POS tagging.
Input: TOKENS (N): [token$_1$, token$_2$, …, token$_N$]
Rules:

1. Return ONLY a JSON array of N POS tags
2. Tags: NOUN PROPN VERB AUX ADJ ADV ADP PRON DET NUM CCONJ SCONJ PART PUNCT INTJ SYM X
3. No explanation, no markdown
4. Always tag and punctuation as PUNCT
5. Count your tags before responding — must equal N
6. Format: `["TAG1","TAG2",…]`

OUTPUT:

POS Prompt Template: Achhami Dialect

Role: You are a Nepali linguistics expert for UD POS tagging.
NOTE: Tokens are in ACHHAMI dialect (Achham district, western Nepal). Apply UD POS rules for this dialect.
Input: TOKENS (N): [token$_1$, token$_2$, …, token$_N$]
Rules:

1. Return ONLY a JSON array of N POS tags
2. Tags: NOUN PROPN VERB AUX ADJ ADV ADP PRON DET NUM CCONJ SCONJ PART PUNCT INTJ SYM X
3. No explanation, no markdown
4. Always tag and punctuation as PUNCT
5. Count your tags before responding — must equal N
6. Format: `["TAG1","TAG2",…]`

OUTPUT:

Example Prompt and Response

POS Prompt: Example (Standard Nepali)

Input: TOKENS (7): [प्रधानमन्त्री, केपी, शर्मा, ओली, सिंहदरबार, गए, ।]
Output: `["NOUN", "PROPN", "PROPN", "PROPN", "PROPN", "VERB", "PUNCT"]`

POS Prompt: Example (Achhami Dialect)

Input: TOKENS (7): [प्रधानमन्त्री, केपी, शर्मा, ओली, सिंहदरबार, गईन, ।]
NOTE: Tokens are in ACHHAMI dialect (Achham district, western Nepal).
Output: `["NOUN", "PROPN", "PROPN", "PROPN", "PROPN", "VERB", "PUNCT"]`