Benchmarking Models for Low-Resource Nepali Event Extraction with Trigger Phrase Identification and Event Classification

Sujal Maharjan[1]*, Astha Shrestha[1]*, Lakshmojee Koduru[2], Sweta Poudel[3], Shuvam Shiwakoti[4], Rabin Thapa[1], Kritesh Rauniyar[5], Surendrabikram Thapa[4]
[1]IIMS College, Kathmandu, Nepal, [2] Google
[3]Kathmandu Engineering College, Tribhuvan University, Kathmandu, Nepal
[4]Virginia Tech, USA, [5]Macquarie University, Australia
sujalmaharjan007@gmail.com, aasthashrestha688@gmail.com

Abstract

Research on Event Extraction (EE) in South Asian languages is crucial for understanding information dissemination and enabling automated news analysis in morphologically rich, low-resource settings. To address the scarcity of high-quality, publicly available datasets, we present Nepali Event Extraction (NepEE), a manually annotated corpus comprising 10,226 Devanagari sentences. The dataset includes annotations for trigger identification and event type classification, achieving high inter-annotator agreement with Fleiss' $\kappa = 0.812$ for trigger identification and $\kappa = 0.855$ for event classification. Our dataset was developed through a rigorous iterative three-phase protocol involving five expert native speakers to ensure linguistic precision. We conduct benchmarking across a broad spectrum of approaches, including classical feature-based models, five fine-tuned Transformer encoders, and contemporary instruction-tuned Large Language Models (LLMs) using zero-shot and fixed few-shot prompting. Our analysis shows that Indic-specialized Transformers achieve superior classification performance, while traditional methods and few-shot prompting struggle with the challenges of exact span extraction in morphologically complex contexts. Furthermore, we quantify performance differences between sentence-level and span-level tasks, establishing strong baselines for future research. The findings and released NepEE dataset provide a valuable resource for advancing event understanding in low-resource languages (LRLs). The dataset, code, and experimental resources are publicly available at GitHub/SUJAL390/EEUCA-ACL-2026.

1 Introduction

Event Extraction (EE) is an important task in Information Extraction (IE), moving beyond entity recognition toward identifying structured event information from unstructured text (Hürriyetoğlu et al., 2025; Xiang and Wang, 2019; Li et al., 2022). The task is commonly divided into two interrelated subtasks: Trigger Identification, which involves anchoring an occurrence to its most salient lexical unit, and Event Type Classification, which maps that anchor to a specific node in a predefined semantic taxonomy. While the field has witnessed a paradigm shift toward high-performance neural architectures facilitated by mature benchmarks such as the Automatic Content Extraction (ACE) 2005 and the Event and Relation Extraction (ERE) corpora (Doddington et al., 2004; Walker et al., 2006), these advances have largely bypassed low-resource languages (LRLs) like Nepali. This digital divide creates a significant bottleneck for the deployment of context-aware systems in the South Asian region, where timely IE is a prerequisite for applications ranging from automated news synthesis to real-time disaster response monitoring (Grishman, 2019).

Nepali, an Indo-Aryan language spoken by approximately 30 million individuals, exhibits a complex set of linguistic peculiarities that challenge traditional extraction frameworks based on Western European languages. Nepali, a morphologically rich SOV language, employs intricate agglutinative structures and compound verb clusters, e.g., नियुक्त भए (*was appointed*), to denote actions. Unlike English, where a trigger is often a distinct lexical unit, Nepali triggers are frequently nominalized or split, where a verbal noun such as सम्झौता (*agreement*) carries the primary semantic load of the event. Furthermore, the extensive use of honorifics and auxiliary inflections necessitates granular character-level span detection to

*The authors contributed equally to this work and are designated as joint first authors. The author order follows alphabetical order by last name.

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 58–71

avoid the inclusion of extraneous morphological markers. Recent research in the region, including the Nepali Language Understanding Evaluation (NLUE) benchmark (Nyachhyon et al., 2025), has successfully pioneered foundational tasks like Named Entity Recognition (NER) and Part-of-Speech (POS) tagging; however, event-level semantic parsing remains underexplored.

To bridge this infrastructure gap, we present a novel, high-quality, human-annotated dataset specifically curated for Nepali trigger identification and event type classification. Recognizing the inherent subjectivity and linguistic nuance involved in semantic labeling, we implemented a rigorous annotation protocol involving five native Nepali speakers with advanced expertise in linguistics and media analysis. To ensure the scientific validity and reproducibility of the corpus, we conducted an exhaustive inter-annotator agreement analysis, yielding a Fleiss' κ of 0.812 for trigger identification and 0.855 for event type classification. According to the diagnostic benchmarks established by Landis and Koch (1977), these coefficients indicate near-perfect agreement, validating our annotation guidelines as a robust and scalable framework for capturing the nuances of Devanagari event semantics. Our schema covers eight diverse event categories, including *Disaster and Accidents*, *Political*, and *Economic Event*, providing a representative cross-section of the contemporary Nepali media landscape.

Beyond the introduction of the corpus, this paper establishes a comprehensive computational baseline by evaluating the proposed dataset across four distinct modeling paradigms to delineate the current performance ceiling. We contrast classical machine learning frameworks, such as feature-engineered Support Vector Machines (SVM) and Random Forests, with state-of-the-art (SOTA) Transformer-based architectures and Large Language Models (LLMs). This evaluation includes massive multilingual encoders such as mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), alongside regionally optimized models like IndicBERTv2 (Doddapaneni et al., 2023) and the specialized nepaliBERT (Ghimire, 2022). Furthermore, we benchmark the generative capabilities of leading LLMs including Qwen-2.5 (Qwen et al., 2025), Gemma-3 (Team et al., 2025), Phi-4 (Abdin et al., 2024), and Llama-3.1 (Grattafiori et al., 2024) under both zero-shot and fixed few-shot prompting configurations. Our results not only provide a rigorous performance benchmark but also offer a diagnostic analysis of the hard cases in Nepali EE, such as indirect phrasing and nominalized triggers, that continue to challenge modern Natural Language Understanding (NLU). Through this work, we provide the foundational data and empirical framework necessary to support future research on South Asian NLU, fostering a more inclusive and linguistically diverse global AI ecosystem.

2 Related Work

To contextualize the challenges addressed by NepEE, we review prior research spanning event extraction benchmarks, multilingual transformer architectures, and Nepali natural language understanding.

2.1 Foundational Benchmarks and the Evolution of Event Extraction

The trajectory of EE as a distinct sub-discipline of IE has been fundamentally shaped by the availability of high-quality, human-annotated corpora. Early foundational efforts were anchored by the ACE 2005 program (Doddington et al., 2004) and the subsequent ERE datasets (Walker et al., 2006), which established the canonical two-stage paradigm: Trigger Identification and Argument Role Labeling. While these benchmarks facilitated the transition from pattern-matching heuristics to statistical and neural models, their linguistic foundations are deeply rooted in Western European syntax.

Recent scholarship has highlighted the inadequacy of these schemas when applied to languages with distinct typological features. As noted by Grishman (2019), the reliance on distinct, monolexemic triggers in English does not easily port to languages where event anchors are distributed across complex morphological clusters. To address the limitations of the small-scale ACE corpora, the community introduced massive general-domain datasets such as MAVEN (Wang et al., 2020) and RAMS (Ebner et al., 2020), which expanded the taxonomic depth of events to thousands of categories. However, these datasets continue to exhibit a significant high-resource bias. Our work bridges this gap by introducing a standardized semantic benchmark for Nepali, focusing on the structural complexities of trigger identification and event type classification.

2.2 The Cross-Lingual Gap in Transformer Architectures

The advent of pre-trained Transformer architectures has redefined the state-of-the-art for sequence labeling and semantic parsing. Multilingual encoders, such as mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), have demonstrated remarkable cross-lingual transfer capabilities by mapping diverse languages into a shared embedding space. While foundational research in domain adaptation suggests that learning shared representations is an optimal strategy for transfer learning (Glorot et al., 2011; Bengio, 2012), recent empirical studies on Low-Resource Languages (LRLs) reveal a more nuanced reality. Specifically, massive multilingual models often exhibit a "curse of multilinguality," where the representation quality for any single language may be diluted as the number of supported languages increases under a fixed parameter budget. However, this limitation is not an absolute law; recent work suggests that increasing model capacity, utilizing Mixture-of-Experts (MoE), or leveraging high-quality instruction-tuning data (such as Bactrian-X (Li et al., 2023)) can effectively mitigate these bottlenecks. Despite the rise of Large Language Models (LLMs), fine-tuned encoders remain indispensable for specific extraction tasks, as zero-shot prompting of generalist models can still underperform compared to dedicated, task-specific baselines (Chen et al., 2024).

For morphologically rich scripts, research indicates that monolingual pre-training or specialized vocabulary adaptation significantly outperforms zero-shot cross-lingual transfer (Pfeiffer et al., 2020). In the South Asian context, the development of Nepali-BERT Ghimire (2022) represented a significant milestone, providing a model trained on native Devanagari corpora that captures the unique distributional semantics of Nepali more effectively than generalized multilingual counterparts. Our benchmarking framework extends this line of inquiry by evaluating whether these monolingual advantages translate to high-level semantic tasks like sentence-level trigger identification, particularly in the presence of complex agglutinative inflections and honorific markers that are often absent in multilingual training sets.

2.3 Event Extraction in Indic and South Asian Languages

Research into EE for the Indo-Aryan language family has gained momentum, yet remains fragmented. Studies in Hindi have leveraged deep neural architectures for event-centric extraction tasks (Sahoo et al., 2020). However, Nepali presents a unique typological profile characterized by its specific handling of nominalized triggers and compound verb clusters. Unlike Hindi, where certain auxiliary structures are more standardized, Nepali exhibits a higher degree of verbal agglutination where the semantic core of an event may be embedded within a multi-token cluster (e.g., नियुक्त भए) (*was appointed*).

Furthermore, existing IE efforts in the region often conflate Nepali with other Devanagari-based languages, failing to account for its distinct verb-final (SOV) constraints and its extensive use of verbal nouns as action anchors. Recent work on the IndicNLPSuite (Kakwani et al., 2020) provided foundational resources for many South Asian languages, yet high-level tasks like event-centric NLU for Nepali were not included in the original benchmarks. By introducing a human-annotated EE corpus with near-perfect inter-annotator agreement ($\kappa > 0.8$), we provide one of the first large-scale efforts to formalize event-level semantics for Nepali, distinguishing it from general Indic-language extraction models.

2.4 The Consolidation of Nepali Natural Language Understanding

Historically, Nepali NLP was confined to foundational tasks such as rule-based (POS) tagging, Named Entity Recognition (NER) like Everest-NER (Niraula and Chapagain, 2022), and basic sentiment Analysis (Bal, 2004). The landscape underwent a rapid modernization with the introduction of the Nep-gLUE and NLUE benchmark (Timilsina et al., 2022; Nyachhyon et al., 2025). These benchmarks provided standardized datasets for NER, question answering, and document classification, thereby creating a performance baseline for the language.

Despite these advances, a significant gap remains in the domain of high-level semantic extraction (Rauniyar et al., 2023; Thapa et al., 2023). Current Nepali NLU benchmarks primarily focus on entity-level or document-level tasks, leaving the intermediate layer of sentence-level event se-

mantics unaddressed. As Nyachhyon et al. (2025) argue, the development of sophisticated NLU for Nepali requires moving beyond foundational syntax toward structured knowledge extraction. We present a novel gold-standard dataset for Nepali EE, covering both trigger identification and event type classification. Our work establishes a benchmark for this task by evaluating a range of models, including classical machine learning algorithms, multilingual transformers, domain-specific Nepali transformers, and LLMs (zero-shot and fixed few-shot prompting). This framework provides a rigorous diagnostic for how different architectures handle the unique semantic and structural challenges of the Devanagari script.

3 Dataset

In this section, we describe our data collection process and the iterative annotation schema developed for the Nepali EE task.

3.1 Data Collection

The dataset was constructed from a publicly released corpus of 65,000 Nepali sentences (Paudyal, 2017). From this base corpus, we curated sentences through a controlled selection procedure designed to ensure event salience and annotation suitability.

Sentences containing explicit eventive expressions, including verbal predicates and nominalized forms, were prioritized. Each sentence was manually reviewed by five trained native Nepali speakers to verify semantic completeness and contextual interpretability prior to annotation. During selection, we continuously monitored class frequencies and applied corrective sampling to maintain balanced representation across the eight event categories. The final dataset consists of 10,226 sentences with a stable class distribution, as shown in Table 2.

3.2 Annotation Process

To ensure high-quality annotations, we engaged five experienced native Nepali speakers possessing a deep understanding of local linguistic structures and media discourse. Annotators were provided with comprehensive guidelines, complete with illustrative examples, for the two primary tasks: trigger identification and event type classification. To maximize inter-annotator consistency and resolve linguistic ambiguities, we implemented a structured, iterative three-phase annotation schema (illustrated in Figure 1). This protocol consisted of an initial dry run, an instruction revision phase, and a final conflict resolution phase.

- **Initial Dry Run**: We initiated the annotation process with a dry run of 40 sample sentences. This phase was crucial in gauging the effectiveness of the guidelines. Initially, annotators faced confusion in identifying the minimal span for compound verb clusters. For instance, in the phrase अनुदान दिएको छ (*has provided a grant*), some annotators selected the entire phrase, while others selected only the core nominalized trigger अनुदान (*grant*). These edge cases were logged for subsequent guideline refinement.

- **Instruction Revision Phase**: Building upon insights from the dry run, the annotation process entered a second phase where 100 additional sentences were annotated. During this phase, annotators were provided with refined instructions, which were adjusted based on the feedback from the initial dry run. This step aimed to enhance the clarity and precision of annotations, particularly in identifying split predicates and character-level boundaries for triggers.

- **Conflict Resolution**: In the final stage, annotators engaged in a collaborative discussion to address discrepancies that arose while annotating 100 sentences after the revision of instructions. This consensus-building process allowed for a thorough review of annotations and a shared understanding of the final guidelines. The resolution of occasional ambiguities was achieved through regular meetings and consultations with experts in annotation. The resolution of ambiguities ensured consistency and accuracy of annotations, enhancing the overall quality of the Nepali Event Extraction (NepEE) dataset.

3.3 Annotation Guidelines

To ensure annotation consistency and linguistic consistency, we devised detailed annotation guidelines to assist the annotators. Given a sentence, it was annotated for two primary interdependent tasks: trigger identification and event type classification.

Table 1: Comparative Summary of Benchmarking and Event Extraction Tasks. EN = English, ZH = Chinese, AR = Arabic, NE = Nepali, Trigger ID = Trigger Identification, Event Class. = Event Type Classification,

Work	Task	Datasets	Data Size	Language
Doddington et al. (2004)	Entity, Relation, Event Ext.	ACE	∼1.1M words	EN, ZH, AR
Wang et al. (2020)	Event Detection (ED)	MAVEN	118,732 instances	EN
Ebner et al. (2020)	Multi-Sentence Arg. Linking	RAMS	9,124 events	EN
Kakwani et al. (2020)	Multilingual NLU Bench.	IndicGLUE	2,473,708 instances	11 Indic
Timilsina et al. (2022)	Nepali NLU Benchmarking	Nep-gLUE	286,941 annotations	NE
Nyachhyon et al. (2025)	NLU Benchmarking (12 tasks)	NLUE	∼341K instances	NE
Ours	**Trigger ID & Event Class.**	**NepEE**	**10,226 sentences**	**NE**

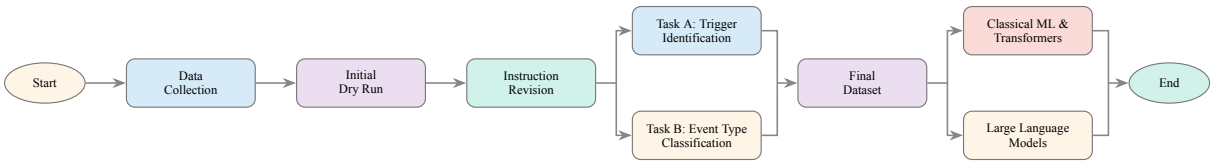

Figure 1: Overview of the end-to-end pipeline for the NepEE dataset.

Table 2: Distribution of class labels for the Event Type Classification task.

Task Label	#Samples	%
Political Event	1,419	13.8
Other Event	1,397	13.7
Economic Event	1,371	13.4
Sports Event	1,289	12.6
Entertainment Event	1,238	12.1
Health Event	1,227	12.0
Disaster and Accidents	1,174	11.5
Education Event	1,111	10.9
Total	**10,226**	**100.0**

A. Trigger Identification: This task identifies the minimal lexical span that signals an event. In a morphologically rich language like Nepali, triggers are frequently complex. Annotators were guided by the following criteria:

- Verbal Triggers: Identifying the minimal span capturing the action, including auxiliary verbs or compound forms e.g., नियुक्त भए (*was appointed*) when they constitute a single semantic unit.
- Nominalized Triggers: Accepting verbal nouns such as सम्झौता (*agreement*) or घोषणा (*announcement*) when they function as the primary action of the sentence.

B. Event Type Classification: Upon identifying a trigger, annotators assigned one of eight predefined event types based on the semantic context (Table 3).

Table 3: Event type classification schema used in annotation. "Disaster & Acc." denotes Disaster and Accidents.

Event Type	Description
Political	Governance-related occurrences such as appointments, resignations, or elections.
Economic	Business activities, trade agreements, and fiscal policy developments.
Sports	Matches, tournaments, victories, or athletic achievements.
Entertainment	Film premieres, festivals, award ceremonies, and artistic performances.
Health	Disease outbreaks, medical updates, or public health campaigns.
Disaster & Acc.	Natural disasters, fires, or transportation accidents.
Education	Academic results, institutional announcements, or education reforms.
Other	Clear events not covered by the above categories.

C. Special Cases: Annotators were provided with instructions for linguistic outliers. For Multiple Events in a single sentence, the primary trigger was chosen. For Idioms or indirect phrasing, annotators focused on clearly realized actions to avoid over-annotation. This methodical approach ensured the reliability and comprehensiveness of the NepEE corpus.

4 Data Analysis

This section provides a detailed analysis of the dataset.

Table 4: Fleiss' Kappa across annotation phases.

Phase	Annotators	κ_{Trig}	κ_{Type}
Pilot Phase	$\alpha_1, \alpha_2, \alpha_3$	0.621	0.764
	$\alpha_2, \alpha_3, \alpha_4$	0.645	0.781
	$\alpha_3, \alpha_4, \alpha_5$	0.638	0.775
Final Phase	$\alpha_1, \alpha_2, \alpha_3$	0.805	0.849
	$\alpha_2, \alpha_3, \alpha_4$	0.818	0.855
	$\alpha_3, \alpha_4, \alpha_5$	0.814	0.861
Aggregated	**Final avg.**	**0.812**	**0.855**

4.1 Inter-annotator Agreement

The reliability of a human-annotated semantic resource is fundamentally predicated on the degree of consensus achieved among independent judges (Fleiss, 1971; Falotico and Quatto, 2015). To quantify this metric for the NepEE corpus, we employed Fleiss' Kappa (κ) to measure agreement across our five native speakers for both the span-level and category-level tasks. Our final analysis yielded an overall agreement of $\kappa = 0.812$ for trigger identification and $\kappa = 0.855$ for event type classification.

A comparative analysis of the agreement coefficients across the longitudinal phases of the project (Table 4) underscores the efficacy of our iterative instruction refinement. The initial pilot phase revealed a moderate agreement for trigger spans, primarily due to the morphological ambiguity inherent in Devanagari verb-particle clusters. However, the subsequent instruction revision phase, which formalized the boundaries for split-predicates and nominalized action anchors, resulted in a substantial performance uplift. According to the interpretative benchmarks of Landis and Koch (1977), the terminal scores indicate near-perfect agreement, ensuring that the resulting dataset serves as a high-quality benchmark for assessing the semantic granularity of modern Nepali NLU systems.

4.2 Linguistic Analysis: Keywords and Topic Salience

To characterize the thematic and lexical properties of the NepEE corpus, we implemented a robust computational pipeline designed to address the morphological complexity and orthographic variations of the Nepali language. Our methodology integrates Unicode NFC normalization to ensure consistent character composition, Devanagari-specific regex filtering, and a custom-curated stopword removal process targeting high-frequency functional noise. Leveraging a class-based TF-IDF (c-TF-IDF) framework, a supervised adaptation of the procedure popularized by the BERTopic framework (Grootendorst, 2022), we quantified the salience of tokens across the eight event categories to identify the distinctive lexical features associated with each class.

The c-TF-IDF results (see Appendix for lexical distribution) demonstrate high semantic density and categorical distinctiveness. For instance, the *Economic Event* class is characterized by domain-specific anchors such as सेयर (*share*) and नेप्से (*NEPSE*), while *Disaster and Accidents* exhibits a strong correlation with vehicular and environmental lexemes like बस (*bus*) and पहिरो (*landslide*). The emergence of these highly relevant keywords validates the gold-standard manual annotations, confirming that the dataset successfully captures the underlying thematic distributions of the Nepali news corpus. It is also worth noting how domain-specific context influences trigger semantics. For example, in the *Education Event* category, विजयी (*victorious*) frequently acts as a trigger for student union elections or academic competitions, while सञ्चालन (*operation*) typically anchors events related to the opening or running of educational institutions.

Furthermore, we conducted a trigger word ambiguity analysis using a cross-tabulation matrix to visualize trigger-class overlap (Figure A2). This analysis quantifies trigger polysemy, the phenomenon where a single lexeme sparks different event types depending on the sentential context. A primary example of semantic overlap is observed in the trigger मृत्यु (*death*), which appears frequently in both *Disaster and Accidents* ($n = 85$) and *Health Event* ($n = 27$). Similarly, while the trigger सार्वजनिक (*public*) shows a strong bias toward *Entertainment Event* ($n = 77$), its distribution across multiple domains reflects its status as a high-utility functional anchor in Nepali media discourse. This dual methodology of keyword salience and trigger ambiguity analysis establishes a rigorous linguistic baseline, highlighting the challenges of contextual disambiguation inherent in automated Nepali EE.

5 Experimental Results and Analysis

To ensure a rigorous and reproducible evaluation, our experimental design treats supervised and generative paradigms differently. For all supervised models, including discriminative classical classi-

Table 5: Model performance for Trigger Identification. **Bold** indicates the best model and F1 score within a category; † indicates the overall highest performer.

Model	Precision	Recall	Macro F1
Classical Models			
CRF	0.6149	0.1916	0.2719
LogisticRegression	0.5569	0.2609	0.3400
RandomForest	0.6074	0.2104	0.2985
PassiveAggressive	0.4208	0.3394	**0.3506**
Transformer-based			
IndicBERTv2	0.7302	0.6908	0.7093
XLM-RoBERTa	0.7445	0.6748	0.7058
NepaliBERT	0.6927	0.6083	0.6419
MuRIL	0.3044	0.3333	0.3182
mBERT†	0.7331	0.6955	**0.7132**
LLM (Zero-shot prompting)			
Qwen2.5	0.3095	0.2495	**0.2651**
Gemma-3	0.2258	0.2128	0.2074
Phi-4	0.1494	0.4299	0.1929
Llama-3.1	0.1018	0.1959	0.1149
LLM (Few-shot prompting)			
Qwen2.5	0.3035	0.3156	**0.2889**
Gemma-3	0.2679	0.3157	0.2722
Phi-4	0.2307	0.4999	0.2878
Llama-3.1	0.1886	0.2131	0.1820

Table 6: Model performance for Event Type classification. **Bold** indicates the best model and F1 score within a category; † indicates the overall highest performer.

Model	Precision	Recall	Macro F1
Classical Models			
SVM	0.7998	0.7768	**0.7856**
RandomForest	0.7420	0.7267	0.7320
LogisticRegression	0.7864	0.7779	0.7813
MultinomialNB	0.7762	0.7664	0.7700
Transformer-based			
IndicBERTv2†	0.8520	0.8576	**0.8536**
XLM-RoBERTa	0.8276	0.8363	0.8298
NepaliBERT	0.8235	0.8262	0.8245
MuRIL	0.8002	0.8078	0.7923
mBERT	0.7915	0.7950	0.7925
LLM (Zero-shot prompting)			
Qwen2.5	0.7015	0.6891	0.6852
Gemma-3	0.7769	0.6701	**0.6977**
Phi-4	0.7353	0.6691	0.6898
Llama-3.1	0.7004	0.5111	0.5259
LLM (Few-shot prompting)			
Qwen2.5	0.7171	0.7124	0.7096
Gemma-3	0.7655	0.7474	**0.7538**
Phi-4	0.7318	0.6707	0.6848
Llama-3.1	0.7084	0.6574	0.6386

fiers and fine-tuned transformer encoders, we utilize a standardized 80/10/10 split (seed = 42) to train, validate, and test the models strictly on unseen data. Conversely, because the instruction-tuned Large Language Models (LLMs) underwent no parameter updates or fine-tuning, there is no risk of parameter leakage. Therefore, to obtain the most statistically robust and comprehensive measure of their zero-shot and few-shot inference capabilities, the LLMs were evaluated across the entire dataset of 10,226 sentences.

5.1 Experimental Methodology

Linguistic Pre-processing Engine To address the morphological richness and orthographic variations of Nepali, we developed a specialized linguistic processor. The engine performs NFKC Unicode normalization and handles zero-width joiners. A core component of our pipeline is an Orthographic Consonantal Root Extractor, which generates a structural representation by stripping all vowel signs (known as *matras* in Nepali) from the token. This method mitigates inflectional variance and was used as a fuzzy-matching fallback to align triggers during data labelling for classical models. Furthermore, we integrate a TnT POS Tagger (Brants, 2000) trained on the Nepali portion of the Indian languages corpus to provide shallow morpho-syntactic signals derived from POS tags and suffix heuristics for our discriminative baselines.

Discriminative Baselines For trigger identification, we framed the task as a token-wise classification problem using the BIO (Beginning, Inside, Outside) tagging formulation (Ramshaw and Marcus, 1995). We evaluated Conditional Random Fields (CRF) (Lafferty et al., 2001) alongside Passive-Aggressive (Crammer et al., 2006), Random Forest, and Logistic Regression classifiers. These models were vectorized via a contextual feature window (w_{i-1}, w_i, w_{i+1}) incorporating extracted consonantal roots and POS tags. For event type classification, we established lexical baselines using SVM (Cortes and Vapnik, 1995), Random Forest, Logistic Regression, and Multinomial Naive Bayes. All classical models were optimized through the Optuna framework (Akiba et al., 2019) across 60 trials per model.

Supervised Transformer Fine-tuning We fine-tuned five SOTA encoder architectures: IndicBERTv2 (IndicBERTv2-MLM-Sam-TLM),

MuRIL (muril-base-cased), XLM-RoBERTa (xlm-roberta-base), NepaliBERT, and mBERT (bert-base-multilingual-cased). Trigger identification was implemented via token classification layers using subword-to-character offset_mapping. Unlike the classical models, transformers for trigger identification used strict literal matching for trigger alignment to ensure character-perfect span offsets. All models were trained for 6 epochs with a learning rate of 2×10^{-5} and a batch size of 32 on Kaggle's T4 GPUs.

Generative LLM Evaluation We assessed the zero-shot and few-shot prompting ($k = 2$) capabilities of four leading models (see Appendix for prompt templates): Qwen-2.5 (Qwen2.5-7B-Instruct), Gemma-3 (gemma-3-4b-it), Phi-4 (Phi-4-mini-instruct), and Llama-3.1 (Llama-3.1-8B-Instruct). Because these models operate strictly in inference mode, they were evaluated across the full corpus to maximize statistical confidence. Performance for trigger identification was measured via token-level overlap F1 to accommodate the generative nature of the models. To ensure evaluation rigor, a heuristic-based label mapper normalized generative outputs to our predefined event categories. All LLM inference was performed on Modal.com using NVIDIA L4 GPUs, 16GB RAM and the vLLM engine (Kwon et al., 2023) for optimized throughput.

5.2 Performance Analysis and Insights

5.2.1 Task A: Trigger Identification

Span-level extraction (Table 5) proved substantially more complex than classification. Supervised mBERT achieved the highest F1-macro (0.7132). Within the classical paradigm, the token-wise Passive-Aggressive classifier ($F1 = 0.3506$) outperformed the CRF ($F1 = 0.2719$), indicating that high-dimensional local contextual signals are highly discriminative for Nepali triggers.

5.2.2 Task B: Event Type Classification

As summarized in Table 6, the supervised IndicBERTv2 model achieves the overall highest macro F1-score (0.8536). A significant finding is the robustness of optimized classical models; the SVM baseline ($F1 = 0.7856$) outperformed several zero-shot LLM configurations, suggesting that specialized morphological features effectively capture event-thematic distributions in the Nepali news domain. Among LLMs, Gemma-3 demonstrated the strongest few-shot performance ($F1 = 0.7538$).

5.3 Error Analysis on Trigger Identification

Trigger Identification for LLMs was evaluated using a strict token-level overlap metric. Under this framework, generative models heavily underperformed, ranging from 0.1149 (Llama-3.1-8B) to 0.2651 (Qwen2.5-7B) in zero-shot settings, peaking at 0.2889 (Qwen2.5-7B) with few-shot prompting. Because strict boundary-based metrics severely penalize boundary inflation alongside correct spans, they obscure the models' actual semantic comprehension. We established a classification taxonomy (Tables A2 and A3) to dissect these false negatives.

The primary driver of precision failure is contextual over-extraction. Instead of isolating single event triggers, models default to summarization, inflating token denominators. Given the event text "गत असार २४ को भोटेकोशीको बाढीले रसुवागढी बन्द भएपछि चीनसँगको आयात व्यापारको अन्तिम विकल्प तातोपानी नाका मात्र हो।" (*After the Bhotekoshi flood on Asar 24 closed Rasuwagadhi checkpoint, the only remaining option for import trade with China is the Tatopani border point.*) (truth: बाढीले) (*due to the flood*), Llama-3.1 extracts an entire argument summary: "भोटेकोशी बाढी रसुवागढी आयात व्यापार तातोपानी" (*Bhotekoshi River, flood, Rasuwagadhi checkpoint, import, trade, Tatopani border point*). This generative habit drives Llama's severe in-sentence span mismatch rate (46.3% zero-shot prompting; 65.8% few-shot prompting).

Furthermore, morphologically rich Nepali text induces systematic boundary misalignment. Models frequently capture inflectional suffixes and auxiliary verbs alongside the root trigger. For the nominal trigger वृद्धि (*increase*), Gemma-3 expands to वृद्धि भएको (*has increased*), while Phi-4-mini generates वृद्धि भएको देखिएको छ (*has been observed to have increased*). These mismatches account for 4.3% (Llama) to 12.7% (Phi-4-mini) of few-shot errors.

Ultimately, these low F1-scores are partly influenced by the strict boundary-based evaluation protocol. While exact-match metrics penalize span expansion, qualitative analysis indicates that generative models often identify semantically relevant event regions but fail to isolate minimal trigger spans. This suggests that span boundary precision,

rather than semantic localization, remains the primary challenge for generative models in Nepali trigger identification.

6 Conclusion

In this paper, we presented NepEE, a manually annotated Nepali EE dataset comprising 10,226 Devanagari sentences with span-level trigger annotations and eight event categories. The dataset establishes a comprehensive benchmark for trigger identification and event type classification in Nepali. Through systematic evaluation across classical machine learning models, transformer-based encoders, and instruction-tuned LLMs, we provide strong baselines and highlight key linguistic challenges including morphological variation, nominalized triggers, and predicate ambiguity. High inter-annotator agreement further supports the reliability of the annotations. The proposed dataset lays a critical foundation for structured IE in Nepali and aims to stimulate broader research on inclusive and multilingual NLP systems.

7 Limitations

Despite its contributions, this work has several limitations. The dataset is derived primarily from a single data source, which may not reflect informal or conversational Nepali and may limit cross-domain generalization. The annotation schema focuses solely on trigger spans and event-type labels and does not include argument roles such as participants, time, and location, which restricts full event-structure modeling. While NepEE currently focuses on these foundational steps, it serves as the first milestone in a broader roadmap. Future iterations of the corpus will extend the schema to include full argument role labeling (e.g., agents, locations, temporal markers) and event coreference, bringing Nepali NLU closer to comprehensive, ACE-style event extraction pipelines.

Furthermore, event boundaries in Nepali can be linguistically complex due to compounding, light verb constructions, and context-dependent trigger interpretation, which makes precise span identification inherently challenging. While careful guidelines were followed, certain edge cases require semantic judgment that may not always be uniformly resolved.

Third, our evaluation of LLMs for trigger identification relies on strict exact-match token overlap metric, which inherently penalizes generative models that produce semantically correct but morphologically misaligned spans (e.g., including auxiliary verbs). Future work should incorporate partial-match F1 metrics and explore advanced prompting strategies (such as Chain-of-Thought reasoning, schema-constrained generation, or Parameter-Efficient Fine-Tuning) to better harness generative capabilities for exact boundary extraction.

In addition, the benchmarks presented in this study are intended to establish competitive baselines rather than define upper performance limits. Continued progress may be achieved through larger-scale data collection, domain diversification, and exploration of more specialized architectures tailored to morphologically rich languages.

8 Ethical Considerations

The foundational sentences for this work originate from Sanjaal Corps' open-source Nepali collection distributed under the Apache-2.0 license by Sanjaal Corps. As the source material is openly licensed for research use, explicit individual consent was not required. Protecting the integrity of the Sanjaal Corps source material was a priority during our processing phase. The dataset does not introduce additional personally identifiable information beyond what is present in the original corpus.

For the annotation process, trained native Nepali speakers were engaged and annotators received a fair wage that matched current local pay scales for linguistic work. Annotators were provided with detailed guidelines and the workflow began with pilot rounds aimed at tightening inter-annotator consistency across complex categories. Given that certain sentences involve topics such as conflict, crime, or public health events, annotators were informed in advance about the nature of the content. Participation was voluntary, and annotators retained the right to withdraw from the task at any stage. Supervisory support was available to address concerns during the annotation process.

As with any curated corpus, the dataset may reflect biases present in the underlying source material. While high inter-annotator agreement supports the internal consistency of the annotations, it does not eliminate potential societal or distributional bias. Researchers are therefore encouraged to exercise caution when deploying models trained on this dataset in sensitive applications.

Finally, we encourage environmentally respon-

sible research practices. Efficient model training, transparent reporting of computational resources, and the use of carbon footprint estimation tools are recommended to reduce the environmental impact of large-scale experimentation.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Bal Krishna Bal. 2004. Structure of nepali grammar. *PAN Localization, Madan Puraskar Pustakalaya, Kathmandu, Nepal*, pages 332–396.

Yoshua Bengio. 2012. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36. JMLR Workshop and Conference Proceedings.

Thorsten Brants. 2000. Tnt–a statistical part-of-speech tagger. In *Sixth Applied Natural Language Processing Conference*, pages 224–231.

Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024. Is a large language model a good annotator for event extraction? In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17772–17780.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7(Mar):551–585.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, Ralph M Weischedel, and 1 others. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8057–8077.

Rosa Falotico and Piero Quatto. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49(2):463–470.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Rajan Ghimire. 2022. NepaliBERT. https://huggingface.co/Rajan/NepaliBERT. Accessed: 2023-02-25.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ralph Grishman. 2019. Twenty-five years of information extraction. *Natural Language Engineering*, 25(6):677–692.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Ali Hürriyetoğlu, Surendrabikram Thapa, Hristo Tanev, and Surabhi Adhikari. 2025. Findings and insights from the 8th workshop on challenges and applications of automated extraction of socio-political events from text. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts*, pages 1–5.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite:

Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4948–4961.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Qian Li, Jianxin Li, Jiawei Sheng, Shiyao Cui, Jia Wu, Yiming Hei, Hao Peng, Shu Guo, Lihong Wang, Amin Beheshti, and 1 others. 2022. A survey on deep learning event extraction: Approaches and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 35(5):6301–6321.

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

Jinu Nyachhyon, Mridul Sharma, Prajwal Thapa, and Bal Krishna Bal. 2025. Consolidating and developing benchmarking datasets for the nepali natural language understanding tasks. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1906–1925.

Kushal Paudyal. 2017. NepaliDataSets: Publicly released Nepali datasets of Sanjaal Corps. https://github.com/sanjaalcorps/NepaliDataSets. Accessed: 2025-04-17.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7654–7673.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third workshop on very large corpora*.

Kritesh Rauniyar, Sweta Poudel, Shuvam Shiwakoti, Surendrabikram Thapa, Junaid Rashid, Jungeun Kim, Muhammad Imran, and Usman Naseem. 2023. Multi-aspect annotation and analysis of nepali tweets on anti-establishment election discourse. *IEEE Access*, 11:143092–143115.

Sovan Kumar Sahoo, Saumajit Saha, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A platform for event extraction in hindi. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2241–2250.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Surendrabikram Thapa, Kritesh Rauniyar, Shuvam Shiwakoti, Sweta Poudel, Usman Naseem, and Mehwish Nasim. 2023. Nehate: Large-scale annotated data shedding light on hate speech in nepali local election discourse. *Frontiers in Artificial Intelligence and Applications*, 372:2346–2353.

Sulav Timilsina, Milan Gautam, and Binod Bhattarai. 2022. Nepberta: Nepali language model trained in a large corpus. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 273–284.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *(No Title)*.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. Maven: A massive general domain event detection dataset. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1652–1671.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

A Appendix

A.1 Prompts

Table A1 presents the lexical distribution of major event categories in the NepEE dataset, including the most frequent trigger words and salient

category-specific keywords ranked using c-TF-IDF. Figure A1 illustrates the zero-shot and few-shot prompting strategies employed for trigger identification and event classification tasks.

A.2 Trigger word ambiguity analysis

Trigger words in event extraction often exhibit varying degrees of semantic ambiguity, where the same lexical item may be associated with multiple event categories depending on contextual usage. To better characterize this challenge in the NepEE dataset, we analyze the distribution of frequently occurring trigger words across different event types. Figure A2 presents a heatmap illustrating the association between selected Nepali trigger words and event categories. The visualization reveals that while some triggers are strongly aligned with a single category, others appear across multiple event types, indicating substantial contextual overlap. For example, certain trigger words commonly associated with political or entertainment events also occur in educational or economic contexts. Such ambiguity highlights the need for context-aware modeling approaches that go beyond isolated trigger identification and incorporate broader semantic and syntactic cues.

A.3 Error analysis

Tables A2 and A3 present the distribution of generative failure modes for zero-shot and few-shot trigger identification, respectively. Across most models, the dominant source of error is *In-Sentence Mismatch*, where the predicted trigger originates from the input sentence but does not correspond to the annotated gold trigger. This suggests that models are often capable of identifying event-relevant lexical spans, yet struggle to precisely localize the correct trigger expression.

Hallucination errors are particularly prominent for Llama-3.1-8B and Phi-4-mini in the zero-shot setting, indicating a tendency to generate unsupported or fabricated trigger words. In contrast, Qwen2.5-7B demonstrates comparatively lower hallucination rates and achieves the highest exact-match performance among the evaluated models. Few-shot prompting generally reduces hallucination and under-extraction errors, especially for Llama-3.1-8B, but also increases morphological mismatch rates in several cases, suggesting that demonstrations encourage semantically related but morphologically inconsistent outputs.

Additionally, Phi-4-mini exhibits substantially higher over-extraction behavior across both settings, frequently generating longer phrases or multiple tokens instead of concise trigger spans. Overall, the findings highlight that trigger identification in Nepali remains challenging not only because of semantic ambiguity, but also due to morphological variation and the generative tendencies of LLMs. This behavior may also reflect inconsistencies between semantically plausible trigger expressions and the single annotated trigger span present in the dataset. In several cases, models generate alternative lexical forms that are contextually appropriate but differ from the gold annotation due to synonym usage or inflectional variation. These findings suggest that future Nepali event extraction systems may benefit from more flexible evaluation schemes and annotation strategies.

Table A1: Lexical distribution and keyword salience across event categories. Trigger frequency (Freq) is shown in parentheses; keywords are ranked by c-TF-IDF.

Event	Top Triggers (Freq)	Significant Keywords (Ranked by salience)
Disaster & Acc.	मृत्यु (*death*, 85), दुर्घटना (*accident*, 71)	मृत्यु (*death*), दुर्घटना (*accident*), बस (*bus*), बाढी (*flood*), पहिरो (*landslide*), नदी (*river*)
Political	छलफल (*discussion*, 36), निर्णय (*decision*, 29)	निर्वाचन (*election*), पार्टी (*party*), संविधान (*constitution*), निर्वाचित (*elected*), मत (*vote*), सरकार (*government*)
Economic	लगानी (*investment*, 22), खर्च (*expense*, 20)	सेयर (*share*), लगानी (*investment*), बैंक (*bank*), कारोबार (*transaction*), मूल्य (*price*), भुक्तानी (*payment*), नेप्से (*NEPSE*)
Sports	पराजित (*defeated*, 48), गोल (*goal*, 27)	रन (*run*), खेल (*game*), विकेट (*wicket*), गोल (*goal*), क्रिकेट (*cricket*), लिग (*league*)
Entertainment	सार्वजनिक (*released*, 77), प्रदर्शन (*screening*, 21)	चलचित्र (*movie*), गीत (*song*), फिल्म (*film*), नाटक (*drama*), अवार्ड (*award*), संगीत (*music*)
Health	उपचार (*treatment*, 28), मृत्यु (*death*, 27)	रोग (*disease*), क्यान्सर (*cancer*), औषधि (*medicine*), अस्पताल (*hospital*), संक्रमण (*infection*), खोप (*vaccine*)
Education	विजयी (*victorious*, 14), सञ्चालन (*operation*, 11)	परीक्षा (*exam*), भर्ना (*admission*), विद्यालय (*school*), विश्वविद्यालय (*university*), शैक्षिक (*academic*)

Prompt: As a domain expert and native Nepali annotator, extract the event trigger word(s) from the text. A trigger is the specific word or phrase indicating that an event has occurred. Return only the trigger word(s).

Prompt: As a domain expert and native Nepali annotator, classify the sentence into one of these categories:
Political Event, Other Event, Economic Event, Sports Event, Entertainment Event, Health Event, Disaster and Accidents, Education Event.
Return only one category name.

Prompt: As a domain expert and native Nepali annotator, extract the event trigger word(s) from the text. A trigger is the specific word or phrase indicating that an event has occurred. Return only the Nepali trigger word(s).

P1	P2
Sentence: काठमाडौंमा ट्रक ठोकिए। *Trucks collided in Kathmandu.* **Answer:** ठोकिए (*collided*)	**Sentence:** नेपाल र भारतबीच व्यापार सम्झौता भयो। *A trade agreement was made between Nepal and India.* **Answer:** सम्झौता भयो (*An agreement was made.*)

Prompt: As a domain expert and native Nepali annotator, classify the sentence into one of these categories:
Political Event, Other Event, Economic Event, Sports Event, Entertainment Event, Health Event, Disaster and Accidents, Education Event.
Return only one category name.

P1	P2
Sentence: बाढीले गाउँ बगायो। *The village was swept away by the flood.* **Answer:** Disaster and Accidents	**Sentence:** नेपाल र भारतबीच व्यापार सम्झौता भयो। *A trade agreement was made between Nepal and India.* **Answer:** Economic Event

Figure A1: Zero-shot and few-shot prompts used for evaluation.

Table A2: Generative failure mode distribution for zero-shot trigger identification. Values represent the percentage of total dataset predictions ($N = 10,226$) falling into each taxonomy category.

Model	Exact Match	Over-extract	Under-extract	Morphological Mismatch	In-Sentence Mismatch	Hallucination	Abstention
Qwen2.5-7B	16.3%	0.5%	15.2%	2.0%	64.5%	1.3%	0.1%
Gemma-3-4B	12.2%	1.5%	9.1%	4.8%	61.5%	10.5%	0.5%
Llama-3.1-8B	2.8%	10.6%	3.6%	2.4%	46.3%	34.3%	0.0%
Phi-4-mini	3.8%	31.7%	1.7%	5.6%	21.1%	30.1%	6.1%

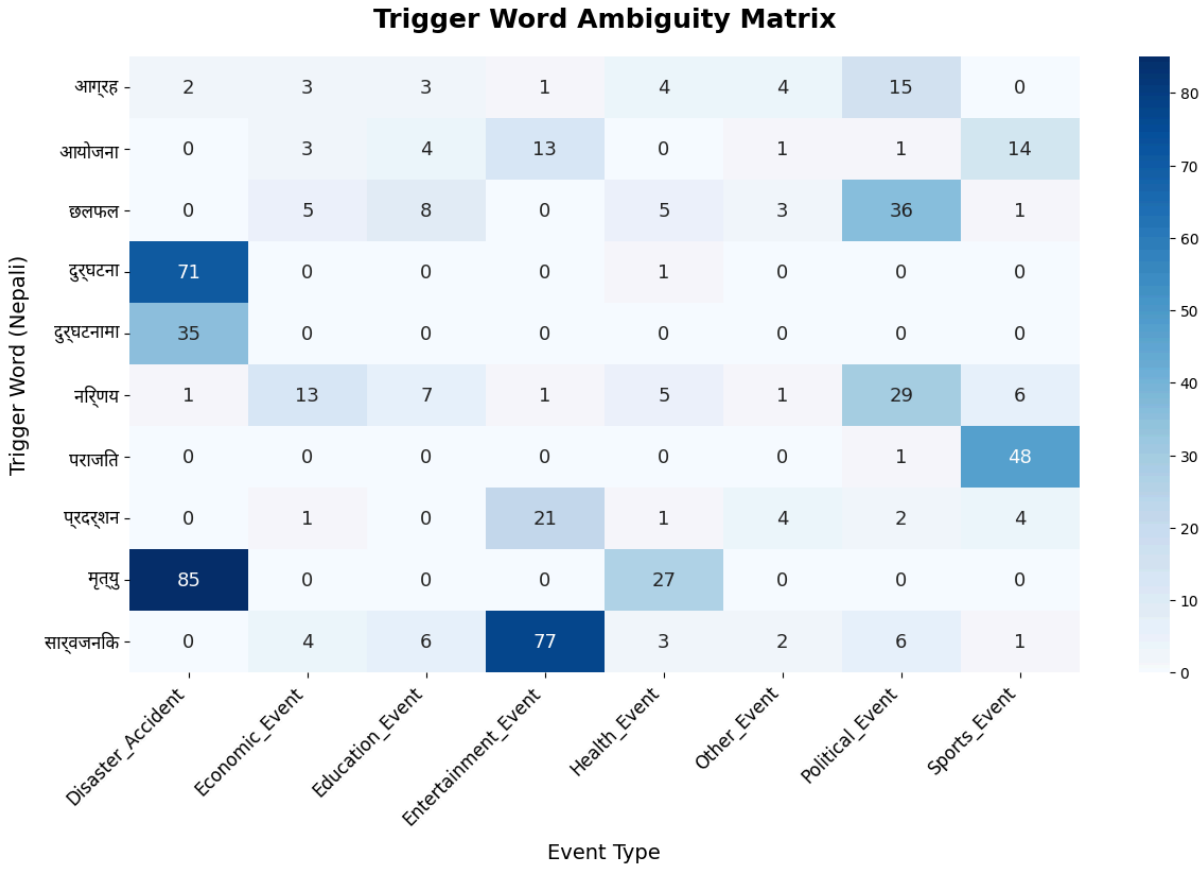

Figure A2: Heatmap of trigger-word ambiguity across event types in the NepEE dataset. Due to font rendering limitations during figure generation, some Nepali Unicode characters may appear slightly distorted in the visualization.

Table A3: Generative failure mode distribution for few-shot trigger identification. Values represent the percentage of total dataset predictions ($N = 10,226$) falling into each taxonomy category.

Model	Exact Match	Over-extract	Under-extract	Morphological Mismatch	In-Sentence Mismatch	Hallucination	Abstention
Qwen2.5-7B	14.5%	3.1%	10.1%	9.0%	61.1%	2.0%	0.1%
Gemma-3-4B	12.5%	4.4%	6.1%	11.8%	51.9%	13.3%	0.1%
Llama-3.1-8B	9.6%	4.4%	6.3%	4.3%	65.8%	9.6%	0.0%
Phi-4-mini	6.3%	30.4%	1.3%	12.7%	26.9%	19.5%	3.0%