Improving Public Health Safety in Low-Resource Languages Using a Human-Verified Health Misinformation Corpus and Large Language Models

Sujal Maharjan[1,*], Astha Shrestha[1,*], Laxmi Thapa[2], Sweta Poudel[3], Shuvam Shiwakoti[4], Rabin Thapa[1], Kritesh Rauniyar[5], Surendrabikram Thapa[4]

[1]IIMS College, Kathmandu, Nepal, [2]O.P. Jindal Global University, India
[3]Kathmandu Engineering College, Tribhuvan University, Kathmandu, Nepal
[4]Virginia Tech, USA, [5]Macquarie University, Australia
{sujalmaharjan007, aasthashrestha688}@gmail.com

Abstract

The proliferation of health misinformation in Low-Resource Languages (LRLs) poses a severe threat to public health, yet automated detection remains critically under-studied due to the scarcity of high-quality benchmarks. We address this gap by introducing Nep-Health-Misinfo, a novel human-verified corpus for health misinformation identification in Nepali. The dataset was developed by adapting four foundational benchmarks (Monkeypox-V1, Monkeypox-V2, COVID-19, and CoAID) through a systematic Machine Translation Post-Editing (MTPE) protocol involving native experts. Our evaluation of Neural Machine Translation (NMT) systems reveals a significant translation asymmetry: while state-of-the-art (SOTA) systems achieve a BLEU score of 43.21 on factual health data, performance degrades sharply on deceptive narratives, with BLEU and TER scores dropping to 19.11 and 62.42, respectively. To establish robust baselines, we benchmark seven recent open-weight Large Language Models (LLMs), including Qwen2.5-7B-Instruct, Gemma-3-4B-IT, and Ministral-8B-Instruct, across zero-shot and few-shot settings. For the few-shot evaluation, we compare stochastic sampling against a K-means centroid-based approach for semantically representative exemplar selection. Experimental results indicate that Qwen2.5-7B-Instruct achieves a peak Macro F1-score of 0.8488, improving over its zero-shot performance (0.7188) on the same dataset. Our findings demonstrate that while few-shot prompting effectively mitigates distribution shifts in low-resource medical contexts, performance remains highly sensitive to the semantic density of exemplars. This work provides the first human-verified Nepali health misinformation corpus. All code and resources are available at https://github.com/SUJAL390/Nep-Health-Misinfo-CHIPSAL-LREC.

Keywords: Nepali NLP, health misinformation, large language models, machine translation post-editing

1. Introduction

The digital revolution has fundamentally restructured the ecosystem of health information dissemination, shifting the locus of authority from centralized medical specialists to decentralized, peer-to-peer online platforms (Swire-Thompson and Lazer, 2020; Eysenbach et al., 2008). While this democratization of data enables the rapid sharing of life-saving medical knowledge, it simultaneously provides a frictionless medium for the unchecked proliferation of health misinformation. The World Health Organization (WHO) has characterized this phenomenon as a global infodemic, a state where an overabundance of information, inclusive of both factual and deceptive narratives, complicates the public's ability to identify trustworthy guidance (Zarocostas, 2020; Cinelli et al., 2020). In high-stakes health domains, the cost of this informational decay is tangible: false claims regarding therapeutic efficacy, vaccine safety, and disease etiology have been dirdectly linked to treatment delays, increased vaccine hesitancy, and heightened mortality rates across diverse populations.

Despite the global nature of this crisis, a significant digital language divide persists in the development and deployment of automated misinformation detection systems (Joshi et al., 2020). Recent advances in Large Language Models (LLMs) have illustrated an exceptional proficiency in recognizing factual inconsistencies and adversarial narratives within high-resource languages such as English (Thapa et al., 2024; Zhou et al., 2023; Arora et al., 2025). However, these capabilities often fail to generalize to low-resource languages (LRLs), where a lack of high-quality, human-verified benchmarks hampers the fine-tuning and evaluation of safety protocols. Nepali, a language spoken by over 30 million people in the South Asian region, remains critically underserved in this regard. The scarcity of localized health corpora leaves Nepali speakers uniquely vulnerable to deceptive health narratives that exploit specific regional linguistic nuances and cultural contexts that global models, optimized primarily for Western medical corpora, frequently overlook (Alber et al., 2025; Acharya, 2023).

* 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 Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 209–218
16 May 2026.

Table 1: Comparative Summary of Benchmarking and Misinformation Tasks. EN = English, NE = Nepali. vLMs = Vision-Language Models; QA = Question Answering.

Work	Task	Datasets	Model	Language
Zhang et al. (2025)	Health Misinfo	1	vLMs	EN
Maheshwari et al. (2025)	General QA	1	LLMs	11 Indic
Thapa et al. (2024)	Health Misinfo	4	LLMs	EN
Singh et al. (2024)	Multilingual Gen	1	LLMs	29 Indic
Ghimire and Shrestha (2025)	Fake News	1	XLM-R	NE/EN
Ours	**Health Misinfo**	**4**	**LLMs**	**NE**

A primary technical impediment to bridging this resource gap is the phenomenon of Translation Asymmetry. While modern Neural Machine Translation (NMT) systems like IndicTrans2 (Gala et al., 2023) have significantly improved general-purpose translation for Indic languages, their robustness degrades sharply when encountering the deceptive, grammatically non-standard, or adversarial linguistic patterns characteristic of misinformation. This decay in translation fidelity introduces semantic noise, where the underlying factual inconsistencies of a health claim are obscured by linguistic artifacts, ultimately degrading the downstream performance of detection frameworks. Consequently, simply translating English datasets into LRLs via automated means is insufficient for meticulous benchmarking; it necessitates a human-centric approach to ensure medical and cultural alignment. To address these challenges, this study introduces the Nep-Health-Misinfo corpus, a high-quality dataset for health misinformation identification in Nepali. We curated this corpus by adapting four foundational English benchmarks including Monkeypox-V1/V2 (Crone, 2022), COVID-19 (Patwa et al., 2021), and CoAID (Cui and Lee, 2020). These specific datasets were selected following the framework established by Thapa et al. (2024) and were processed utilizing a highly intensive Machine Translation Post-Editing (MTPE) protocol. By employing two native Nepali annotators with medical context awareness, we ensure that the dataset preserves the deceptive content of the original claims while reflecting natural Nepali syntax. This process allows us to quantify for the first time the translation gap between factual and deceptive health text in a low-resource environment.

Building upon the benchmarking framework established by Thapa et al. (2024), we evaluate seven recent open-weight LLMs: Qwen2.5-7B-Instruct (Qwen2.5-7B), Gemma-3-4B-IT (Gemma-3-4B), Ministral-8B-Instruct (Ministral-8B), Falcon3-7B-Instruct (Falcon3-7B), Phi-4-Mini-Instruct (Phi-4-Mini), Mistral-7B-Instruct-v0.3 (Mistral-7B), and Falcon-H1-3B-Instruct (Falcon-H1-3B). Following their experimental setup, our evaluation spans zero-shot and few-shot settings, comparing random sampling against a K-means centroid-based approach for selecting semantically representative exemplars. Our findings indicate that few-shot settings outperform zero-shot baselines. The contributions of this work are threefold: (i) we provide the Nep-Health-Misinfo corpus, the first human-verified health misinformation dataset for Nepali; (ii) we present a detailed technical analysis of how translation asymmetry affects the quality of low-resource health resources; and (iii) we establish a performance baseline for recent open-weight LLMs in a South Asian LRL context.

2. Related Work

The evolution of the digital landscape has facilitated an unprecedented shift in how medical knowledge is consumed and contested. Historically, research into health misinformation focused on the transition from centralized medical authority to decentralized, algorithmically driven information ecosystems (Swire-Thompson and Lazer, 2020). As Generative Artificial Intelligence (GenAI) matures, the challenge has shifted from identifying simple deceptive patterns to addressing sophisticated AI-driven narratives that mimic human-level credibility and expert reasoning (Zhou et al., 2023; Vraga and Bode, 2020).

2.1. Computational Paradigms in Detection

Automated misinformation detection has historically been grounded in heuristic-systematic psychological frameworks (Metzger, 2007). Early computational efforts relied on feature engineering and classical Machine Learning (ML) methods, such as Support Vector Machines (SVM) and Decision Trees, to classify textual patterns.

The introduction of the Transformer architecture marked a definitive paradigm shift from these manual feature-based approaches. Pre-trained Language Models (PLMs) like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) utilized

self-attention mechanisms to capture latent semantic relationships, significantly outperforming earlier methods (Devlin et al., 2019; Liu et al., 2019). Consequently, research during the COVID-19 pandemic largely transitioned toward these deep semantic models, though classical ML methods remained relevant for establishing baselines on static claims (Bojjireddy et al., 2021).

In the South Asian context, the research landscape remains largely anchored in this discriminative stage. Giri et al. (2024) utilized Convolutional Neural Networks (CNNs) to explore deceptive linguistic patterns in Nepali digital media, while Ghimire and Shrestha (2025) introduced bilingual classification frameworks utilizing XLM-RoBERTa for cross-lingual news classification. Despite their efficacy, these models are primarily classification-oriented and lack the zero-shot reasoning capabilities required to disentangle the nuanced factual inconsistencies found in modern medical misinformation.

2.2. LLM Safety and Hallucination Gap

The advent of LLMs has introduced Chain-of-Thought (CoT) reasoning, allowing models to perform multi-step fact-checking (Wei et al., 2022). Recent evaluations demonstrate that frontier models can surpass human benchmarks on medical exams; however, this performance is often deceptive. Huang et al. (2025) and Alber et al. (2025) highlight the hallucination gap, where models generate plausible but clinically dangerous advice due to statistical biases in their training data.

Global safety benchmarks have attempted to quantify these risks. Arora et al. (2025) introduced HealthBench, which employs physician-defined rubrics to evaluate clinical safety, and Thapa et al. (2024) established a critical framework for evaluating LLM susceptibility to medical myths. Furthermore, Zhang et al. (2025) developed MM-Health, a multimodal dataset illustrating how AI-generated health content can deceive existing state-of-the-art (SOTA) classifiers. A structured comparison of these benchmarking efforts is presented in Table 1. As evident from this comparison, existing benchmarks remain predominantly English-centric, leaving the safety profile of LLMs in LRLs largely unmapped.

2.3. Linguistic Resource Scarcity

A significant linguistic gap persists, as the majority of safety-critical research excludes languages like Nepali. Recent initiatives like IndicParam (Maheshwari et al., 2025) and IndicGenBench (Singh et al., 2024) have begun benchmarking South Asian languages for general tasks, but they lack domain-specific medical depth. A major bottleneck in LRLs research is the unreliability of NMT; as noted by Kamath et al. (2025), raw translation often distorts the medical intent and cultural nuance of health claims.

Our work addresses these challenges by introducing the Nep-Health-Misinfo corpus, a human-verified dataset developed via a highly intensive MTPE protocol. By extending the established experimental framework of Thapa et al. (2024) to this novel corpus, we provide the first comprehensive analysis of how recent open-weight models handle the semantic complexities of health misinformation in a South Asian LRLs environment. This study connects local discriminative methods with broader generative safety benchmarking, establishing a foundational baseline for public health safety in the region.

3. Dataset and Methodology

A critical contribution of this work is the creation of a high-quality Nepali dataset. To safeguard against the semantic drift common in low-resource translation, we employed a MTPE workflow to ensure semantic fidelity and cultural appropriateness.

3.1. Language Resource Construction

We introduce Nep-Health-Misinfo, a human-verified corpus designed to serve as a high-quality benchmark for health misinformation detection in the Devanagari script. The dataset comprises 4,924 unique instances aggregated from four established English datasets: CoAID, COVID-19, Monkeypox-V1, and Monkeypox-V2. Prior to translation, we applied a thorough data-cleaning protocol to the English source corpora. This involved systematically filtering out noisy, incomplete, or low-quality source rows and purging extensive cross-dataset duplicates. This preprocessing step ensured that only high-quality, medically relevant claims were advanced to the translation pipeline. Detailed statistics, including the breakdown of *real* and *fake* claims for each sub-corpus, are presented in Table 2. This resource was specifically curated to evaluate how modern LLMs handle the semantic nuances of sensitive medical claims when transitioned from high-resource English to the low-resource South Asian context of the Nepali language.

The construction of Nep-Health-Misinfo followed a multi-stage pipeline designed to maximize linguistic and medical accuracy. In the initial phase, we adhered to specific preservation protocols where @mentions, #hashtags, abbreviations (e.g., WHO, CDC), and specific identifiers like COVID-

Table 2: Distribution of real and fake health claims across the four sub-corpora of the Nep-Health-Misinfo dataset.

Dataset	Label	# Samples
CoAID	Real	441
	Fake	25
COVID-19	Real	1,050
	Fake	930
Monkeypox-V1	Real	1,329
	Fake	903
Monkeypox-V2	Real	158
	Fake	88
Total		**4,924**

19 were retained in their original English format during translation. We then established an automated baseline by processing the source text through three SOTA NMT engines: Google NMT, NLLB-200 (600M), and mBART-50. These models provided a diverse set of initial hypotheses, allowing us to identify consistent failure points in Devanagari script generation, specifically regarding the transliteration of Western biomedical terms versus native Nepali nomenclature.

To elevate the dataset to a gold-standard level, we conducted a rigorous MTPE phase on the entire corpus. Two native Nepali speakers were recruited for a careful sentence-by-sentence manual review. The editors were tasked with three primary objectives: ensuring strict medical fidelity, correcting grammatical errors and mistranslations introduced by NMT, and adapting Western-centric misinformation narratives into natural, culturally relevant Nepali phrasing without altering the ground truth. The original labels from the source datasets were preserved throughout translation and post-editing, and annotators did not modify class assignments. Disagreements between annotators were resolved through discussion, and unresolved cases were adjudicated by a senior reviewer.

Following the MTPE process, a final verification pass was conducted to ensure absolute uniqueness and linguistic consistency. Additionally, we anonymized specific user mentions to @MENTION to ensure models focus on the informational content rather than exhibiting bias toward specific entities; however, #hashtags and abbreviations were preserved in their original form.

Finally, we quantified the necessity of this human-verified data by evaluating the raw NMT outputs against our final human-edited version using BLEU, chrF, and Translation Edit Rate (TER) metrics. As summarized in Table 3, while Google NMT provided the strongest baseline, it achieved a TER of 40.95 on the CoAID dataset, indicating a high level of edit operations required during post-editing, reflecting substantial translation effort. Furthermore, our label-specific analysis revealed a complexity gap: misinformation instances *fake* yielded a significantly higher TER (62.42) compared to factual news (40.95). This suggests that health misinformation relies on linguistic irregularities that defy automated translation, thereby necessitating rigorous MTPE used in this study.

Table 3: Quantitative evaluation of translation models across datasets, stratified by veracity. Bold values highlight the top model per metric for each label, while red text indicates the global best for that metric across the specific dataset.

Dataset	Label	Metric	Google	NLLB	mBART
CoAID	Real	BLEU	**43.21**	28.44	17.39
		chrF	**71.76**	61.84	51.53
		TER ↓	**40.95**	57.33	67.04
	Fake	BLEU	**19.11**	12.56	2.14
		chrF	**65.92**	58.82	36.58
		TER ↓	**62.42**	78.34	85.99
COVID-19	Real	BLEU	**32.16**	23.57	15.62
		chrF	**66.91**	57.50	51.27
		TER ↓	**48.72**	62.44	72.10
	Fake	BLEU	**34.05**	22.99	15.75
		chrF	**68.85**	59.79	52.01
		TER ↓	**49.98**	68.82	74.70
Monkeypox-V1	Real	BLEU	**32.68**	21.94	16.87
		chrF	**64.35**	52.28	47.95
		TER ↓	**53.76**	73.70	76.25
	Fake	BLEU	**29.62**	21.66	13.89
		chrF	**60.08**	49.32	44.70
		TER ↓	**56.56**	70.93	79.81
Monkeypox-V2	Real	BLEU	**34.90**	19.02	15.04
		chrF	**65.39**	53.05	48.37
		TER ↓	**53.53**	71.82	80.85
	Fake	BLEU	**30.55**	22.26	14.51
		chrF	**56.73**	46.30	42.88
		TER ↓	**57.48**	70.35	78.74

3.2. Models and Experimental Settings

We benchmarked seven open-weight models selected for their parameter efficiency and multilingual capabilities: Qwen2.5-7B, Gemma-3-4B, Ministral-8B, Falcon3-7B, Phi-4-Mini, Mistral-7B, and Falcon-H1-3B. To rigorously evaluate in-context learning (ICL) capabilities in a low-resource setting, we tested each model under five distinct configurations: (i) zero-shot classification,

Example of Zero-shot prompt

Task: Health Misinformation Detection
Instruction: Determine if the following tweet in Nepali contains health misinformation.
Output 'Yes' if it is misinformation (fake) or 'No' if it is factual (real).
Output ONLY the word 'Yes' or 'No'.

Figure 1: Example of a Zero-shot prompt for Health Misinformation Detection.

Example of Few-shot prompt

Task: Health Misinformation Detection, **Instruction:** Determine if the following tweet in Nepali contains health misinformation. Output 'Yes' if it is misinformation (fake) or 'No' if it is factual (real). Output ONLY the word 'Yes' or 'No'.

P1	**P2**	**P3**	**P4**
Tweet: बिरालो र कुकुरले कोरोनाभाइरस फैलाउँछन् । **Translation:** Cats and dogs spread coronavirus.	**Tweet:** वैज्ञानिकहरू भन्छन् कि अमेरिका मंकीपक्स विरुद्धको लडाइँ हार्दै गइरहेको हुन सक्छ। **Translation:** Scientists say that America might be losing the fight against monkeypox.	**Tweet:** गेट्स फाउन्डेसनसँग यस कोरोनाभाइरसको प्याटेन्ट छ। **Translation:** The Gates Foundation has the patent for coronavirus.	**Tweet:** खैर यो मान्छेलाई पक्कै पनि समलिङ्गी भएको कारण मंकीपक्स भएको छ। **Translation:** Well, this guy definitely got monkeypox because he is gay.
Misinformation: Yes	**Misinformation:** No	**Misinformation:** Yes	**Misinformation:** Yes
Output: Yes	**Output:** No	**Output:** Yes	**Output:** Yes

Figure 2: Example of a Few-shot prompt for Health Misinformation Detection.

(ii) 5-shot random, (iii) 10-shot random, (iv) 5-shot sampled, and (v) 10-shot sampled. In the random setting, exemplars were selected uniformly at random from the training pool using a fixed seed.

Following Thapa et al. (2024), we used a clustering-based exemplar selection strategy for the sampled few-shot setting. For each shot size $k \in \{5, 10\}$, we constructed a label-balanced exemplar set by selecting approximately $k/2$ examples from each class *real* and *fake* whenever possible. Within each class subset, we represented Nepali texts using TF-IDF features with up to 500 dimensions and both unigram and bigram n-grams. We then applied K-means clustering with the number of clusters set to the required number of exemplars for that class. For each cluster, we selected the instance closest to the centroid as the exemplar. If clustering was not feasible due to insufficient data or feature sparsity, we fell back to random sampling within the class. The zero-shot prompt template used across experiments is illustrated in Figure 1, while the structure of the few-shot prompt with labeled exemplars is shown in Figure 2. All experiments were conducted using the vLLM library (Kwon et al., 2023) on L4 GPUs within the Google Colab ecosystem. To ensure reproducibility, random selection of few-shot exemplars was controlled with a fixed seed (seed = 1). For clustering-based sampled selection, K-means initialization was fixed internally (seed = 42), producing deterministic exemplar sets. This careful control of randomization ensures that all experiments can be exactly reproduced, allowing consistent comparison of model performance across different prompting settings and datasets. It also minimizes variability introduced by random exemplar selection while maintaining methodological rigor.

4. Results and Analysis

Our comprehensive benchmarking revealed distinct performance hierarchies among the evaluated architectures. The complete Macro F1 performance comparison across datasets and prompting configurations is presented in Table 4. In Table 4, random few-shot results were obtained using a fixed seed (seed = 1), while K-means-based sampled settings used a fixed initialization seed (seed = 42). Qwen2.5-7B was the most consistently strong overall performer across datasets. It achieved the highest peak Macro F1-scores on multiple sub-corpora, including 0.8488

Table 4: Performance comparison across datasets under different prompting settings. Bold text highlights the optimal model for each setting, while red text indicates the overall best model for each dataset.

Dataset	Setting	Qwen2.5-7B	Gemma-3-4B	Ministral-8B	Falcon3-7B	Phi-4-Mini	Mistral-7B	Falcon-H1-3B
Monkeypox-V1	Zero-Shot	**0.7188**	0.3255	0.3361	0.4922	0.4966	0.4170	0.5946
	5-Shot Sampled	**0.8029**	0.7032	0.4622	0.7659	0.7163	0.6187	0.7564
	10-Shot Sampled	**0.8121**	0.6945	0.6827	0.7475	0.7679	0.7012	0.7240
	5-Shot Random	0.6662	0.8181	0.8301	0.7647	0.7448	0.6764	0.8033
	10-Shot Random	**0.8235**	0.7505	0.7546	0.7552	0.7836	0.7566	0.7956
Monkeypox-V2	Zero-Shot	**0.7061**	0.2903	0.3179	0.4971	0.4582	0.3480	0.5908
	5-Shot Sampled	0.6849	0.6938	0.4359	0.6819	**0.6928**	0.5115	0.6534
	10-Shot Sampled	0.7283	0.6549	0.5731	**0.8039**	0.7130	0.5947	0.6690
	5-Shot Random	0.7918	0.7449	0.5913	**0.8028**	0.7945	0.6014	0.6530
	10-Shot Random	0.8488	0.7748	0.8449	0.8060	0.8003	0.6862	0.6970
CoAID	Zero-Shot	0.3413	0.0678	0.2254	0.1480	0.2325	0.1452	**0.6427**
	5-Shot Sampled	0.5771	0.4890	0.1723	0.6634	0.3643	0.6183	**0.7252**
	10-Shot Sampled	**0.7453**	0.5731	0.1995	0.5671	0.4132	0.5917	0.6303
	5-Shot Random	**0.7594**	0.3679	0.0606	0.5912	0.2108	0.4914	0.6693
	10-Shot Random	0.7798	0.6366	0.6220	0.6460	0.4812	0.6729	0.6290
COVID-19	Zero-Shot	**0.7153**	0.4231	0.4417	0.6153	0.6556	0.6414	0.5107
	5-Shot Sampled	0.7386	**0.8161**	0.5268	0.6025	0.7755	0.7782	0.6249
	10-Shot Sampled	0.7458	0.8368	0.8358	0.5488	0.7764	0.7822	0.5207
	5-Shot Random	0.6806	**0.7979**	0.7058	0.6443	0.7472	0.7451	0.6287
	10-Shot Random	0.7315	**0.8317**	0.8170	0.6865	0.7054	0.7324	0.7324

on Monkeypox-V2 and 0.7798 on CoAID in the 10-shot random setting. However, no single model dominated all configurations; models such as Gemma-3-4B and Ministral-8B achieved competitive or superior performance in specific datasets and prompting settings, indicating that model effectiveness varies depending on task characteristics and experimental setup. While Qwen2.5-7B demonstrates strong cross-dataset robustness, attributing this performance solely to pretraining factors (e.g., exposure to Asian linguistic data or Devanagari scripts) remains speculative and requires further controlled analysis.

Conversely, zero-shot performance was generally lower than few-shot settings across most datasets and models; for instance, Gemma-3-4B achieved a Macro F1 of only 0.068 on CoAID. This zero-shot gap underscores the necessity of in-context grounding for LRL, as models frequently fail to map Nepali medical terminology to the correct label without structured demonstrations. Interestingly, while K-means sampling was hypothesized to outperform random selection, our results show that 10-shot random prompting often yielded higher peak scores. This suggests that in the presence of the semantic noise inherent in misinformation, simple random sampling may provide sufficient lexical and contextual diversity, performing comparably to or even better than centroid-based exemplar selection strategies.

4.1. The Random Sampling Advantage

A critical and novel finding of this study is the divergence of LRL performance from established trends in English-centric benchmarks. While prior literature typically advocates for clustering-based sampling to ensure diverse semantic coverage, our results for Nepali indicate that 10-shot random selection frequently enables models to achieve higher performance peaks.

To validate this trend, we performed an explicit multi-seed evaluation for our leading models, Qwen2.5-7B and Falcon3-7B, across three distinct random seeds (1, 123, and 2026). On the Monkeypox-V2 dataset, Qwen2.5-7B achieved a peak Macro F1-score of 0.8488 (seed = 1), while Falcon3-7B reached 0.8060 (seed = 1). Both scores significantly exceeded their deterministic clustering-based baselines. This trend persisted on the larger COVID-19 dataset, where Qwen2.5-7B reached a peak of 0.7612 (seed = 2026), surpassing the 0.7458 achieved through K-means clustering.

While multi-seed evaluation shows that performance is sensitive to exemplar selection, results vary noticeably across random seeds. Despite this, random sampling often outperformed clustering-based sampling in our reported runs. We hypothesize that in LRL settings, clustering-based selection may limit the diversity of contextual signals provided to the model, whereas random sampling captures a broader range of linguistic patterns. This suggests that, in our setting, lexical and semantic diversity in exemplars may be more important than centroid-based selection.

4.2. Inter-Model Agreement

Assessing the consistency of predictions across diverse architectures is crucial for ensuring system reliability. We utilized Fleiss' Kappa (κ) to evalu-

Table 5: Fleiss Kappa Scores Across Different Training Settings. Red text indicates the best setting for each dataset.

Dataset	Setting	Fleiss' Kappa
COVID-19	10-Shot Random	**0.5115**
	10-Shot Sampled	0.4025
	5-Shot Random	0.3549
	5-Shot Sampled	0.2910
	Zero-Shot	0.0874
CoAID	10-Shot Random	**0.2988**
	10-Shot Sampled	0.1627
	5-Shot Random	0.0010
	5-Shot Sampled	0.1479
	Zero-Shot	0.0148
Monkeypox-V1	10-Shot Random	**0.4568**
	10-Shot Sampled	0.3667
	5-Shot Random	0.4204
	5-Shot Sampled	0.2798
	Zero-Shot	0.0036
Monkeypox-V2	10-Shot Random	**0.4764**
	10-Shot Sampled	0.2700
	5-Shot Random	0.2923
	5-Shot Sampled	0.1740
	Zero-Shot	0.0073

ate agreement among the seven evaluated models. The resulting inter-model agreement scores across datasets and prompting configurations are summarized in Table 5. Our analysis reveals that inter-model agreement increases with the number of in-context demonstrations provided.

In zero-shot settings, inter-model agreement was extremely low ($\kappa \approx 0.00$–0.08), indicating highly inconsistent predictions across models. However, in 10-shot settings, agreement rose to moderate levels ($\kappa > 0.51$ on COVID-19). This convergence suggests that providing sufficient exemplars encourages divergent architectures to adopt a more consistent decision framework for identifying Nepali misinformation, effectively aligning the models through the provided context.

4.3. Qualitative Analysis: Systemic Failure Modes

To deepen our understanding of model limitations, we conducted a granular error analysis, isolating instances where all seven models failed across both 10-shot settings ($n = 14$). This analysis revealed three systemic failure modes that characterize the current limitations of generative AI in the Nepali health domain.

First, models exhibited a pronounced Authority Bias. They consistently misclassified fabricated claims as Real when the text contained the names of high-prestige international or national institutions. For example, a fabricated claim stating "UNICEF releases coronavirus prevention guidelines" (ID 647) was universally accepted as factual. As shown in Table 6, this heuristic shortcut appears to override critical analysis; similar failures occurred with claims attributing false information to national medical bodies such as AIIMS (ID 1192).

Second, we observed a vulnerability to Pseudo-Scientific Mimicry. Misinformation utilizing sophisticated medical vocabulary was frequently labeled as factual. This was evident in instances where models failed to flag deceptive content containing terms like "genome" and "host immunity" (ID 868). This suggests that models rely on a complexity heuristic, where domain-specific vocabulary is treated as a proxy for truthfulness, failing to distinguish between genuine medical discourse and stylistic mimicry.

Finally, we observed a significant trend of Safety-Alignment Bias. Factual news reporting on LGBTQ+ health issues, specifically regarding Monkeypox, was frequently flagged as *fake*. As seen in Table 6 (ID 273 and 272), legitimate updates regarding vaccine availability and symptom reporting within the community were misclassified. This may indicate that alignment or safety tuning can sometimes make models overly cautious in cases involving minority identities and infectious disease.

5. Conclusion

This study introduces Nep-Health-Misinfo, the first human-verified Nepali health misinformation dataset. Through a comprehensive evaluation of seven recent open-weight models, we established that Qwen2.5-7B serves as a robust baseline for deployment in South Asian LRL settings. Our experimental results yield a significant methodological finding: stochastic (random) exemplar selection consistently reaches higher performance peaks than deterministic clustering-based strategies. This suggests that for Nepali health misinformation, capturing broad lexical diversity is more critical than maximizing semantic centroid representation. This finding simplifies the technical path for LRL deployment by reducing the computational overhead required for complex data preprocessing without compromising downstream accuracy.

However, our qualitative analysis revealed critical systemic vulnerabilities, specifically Author-

Table 6: Systemic Failure Analysis: Representative examples where all models failed ($n = 14$). The samples illustrate specific failure modes, including: Authority Bias (misleading reliance on prestige entities like UNICEF/AIIMS), Pseudo-Science (confusion caused by medical jargon), and Safety-Alignment Bias (misclassification of factual LGBTQ+ health news).

Dataset	Failure Mode	ID	Original Text	Translation	Truth	Pred
COVID-19	**Authority Bias**	647	UNICEF releases coronavirus prevention guidelines.	UNICEF ले कोरोनाभाइरस रोकथाम दिशानिर्देशहरू जारी गर्यो।	Fake	Real
		1192	AIIMS has released a list of respiratory symptoms of Covid-19 and similar diseases.	AIIMS ले Covid-19 र यस्तै रोगहरूको श्वासप्रश्वासका लक्षणहरूको सूची जारी गरेको छ।	Fake	Real
COVID-19	**Pseudo-Science**	868	there is variation in mortality and infection rate based upon the genome of the virus. Host immunity is also playing a role.	भाइरसको जीनोममा आधारित मृत्युदर र संक्रमण दरमा भिन्नता छ।होस्टको प्रतिरक्षाले पनि भूमिका खेलिरहेको छ।	Fake	Real
		216	The private health system began offering antibody tests to detect COVID-19. At the same time that the Ministry of Health said it would buy antigen tests.	निजी स्वास्थ्य प्रणालीले COVID-19 पत्ता लगाउन एन्टिबडी परीक्षणहरू प्रस्ताव गर्न थाल्यो। उही समयमा स्वास्थ्य मन्त्रालयले एन्टिजेन परीक्षणहरू खरिद गर्ने बतायो।	Fake	Real
Monkeypox-V1	**Safety Bias**	273	Pride Center Offers Free Vaccinations as Monkeypox Surges in Broward County	ब्रोवार्ड काउन्टीमा मङ्कीपक्स बढ्दै जाँदा प्राइड सेन्टरले निःशुल्क खोपहरू प्रदान गर्दछ।	Real	Fake
		272	I shot video interviews of gay men talking about their #monkeypox symptoms for my @MENTION article. They talk about the first signs of the infection and about severe pain. via @MENTION	मैले मेरो @MENTION लेखका लागि समलिङ्गी पुरुषहरूले आफ्ना #monkeypox लक्षणहरूबारे कुरा गरिरहेको भिडियो अन्तर्वार्ताहरू खिचें। उनीहरू सङ्क्रमणको पहिलो सङ्केत र गम्भीर पीडाको बारेमा कुरा गर्छन्। @MENTION मार्फत	Real	Fake

ity Bias and Safety-Alignment Bias. These findings serve as a necessary caution: current frontier models remain susceptible to deceptive authoritative tones and may inadvertently censor factual discourse concerning marginalized communities. By providing both a high-quality dataset and a quantified analysis of the current performance gap, this work establishes a foundational framework for bridging the digital language divide in public health safety. We anticipate that the Nep-Health-Misinfo corpus will catalyze future research into cross-lingual alignment and the development of safer, more equitable multilingual AI systems.

6. Limitations and Ethical Considerations

6.1. Linguistic and Data Limitations

While Nep-Health-Misinfo represents the first human-verified health misinformation dataset for Nepali health safety, several limitations persist. First, the dataset is derived from four English-centric benchmarks; consequently, it may not fully capture misinformation narratives that are indigenous to the South Asian socio-political landscape, such as specific Ayurvedic or local folk-medicine fallacies. Second, while our MTPE protocol utilized native experts to mitigate translation asymmetry, the source material's Western origin remains a latent bias in the dataset's thematic distribution. Future iterations should incorporate natively authored Nepali misinformation from local social media platforms to capture regional rhetorical styles.

6.2. Computational and Model Scope

Our benchmarking was restricted to seven open-weight models under 10 billion parameters to ensure feasibility for low-resource deployment. While Qwen2.5-7B emerged as a consistent model, the performance of significantly larger frontier models remains unmapped in this specific context. Furthermore, our study focused on text-only misinformation; however, real-world health deception in Nepal often propagates through multimodal formats, including deepfake audio and manipulated imagery, which fall outside the current scope.

6.3. Ethical Considerations

The curation and release of misinformation datasets necessitate strict ethical oversight. To prevent the accidental propagation of harmful claims, the Nep-Health-Misinfo corpus is intended strictly for research purposes. All instances in the dataset have been clearly labeled to prevent confusion with factual medical advice. Furthermore, during the construction phase, we performed rigorous anonymization of all user-specific metadata, replacing handles with `@MENTION`, to protect the privacy of the original posters in accordance with global data protection standards.

Finally, our discovery of safety-alignment bias regarding LGBTQ+ health reporting highlights a significant ethical risk: the potential for AI systems to inadvertently marginalize vulnerable populations under the guise of safety. We advocate for the responsible use of this dataset to audit and de-bias LLMs, ensuring that public health safety initiatives do not result in the systemic censorship of factual, life-saving information for minority communities.

References

U Acharya. 2023. Promoting digital literacy with scarce resources. *Managing the Misinformation Effect—The State of Fact-Checking in Asia. International Federation of Journalists*.

Daniel Alexander Alber, Zihao Yang, Anton Alyakin, Eunice Yang, Sumedha Rai, Aly A Valliani, Jeff Zhang, Gabriel R Rosenbaum, Ashley K Amend-Thomas, David B Kurland, et al. 2025. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, 31(2):618–626.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Sirisha Bojjireddy, Soon Ae Chun, and James Geller. 2021. Machine learning approach to detect fake news, misinformation in covid-19 pandemic. In *Proceedings of the 22nd Annual International Conference on Digital Government Research*, pages 575–578.

Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The covid-19 social media infodemic. *Scientific reports*, 10(1):16598.

Stephen Crone. 2022. Monkeypox misinformation: Twitter dataset.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

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.

Gunther Eysenbach et al. 2008. Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. *Journal of medical Internet research*, 10(3):e1030.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages.

Plan Ghimire and Pranjal Shrestha. 2025. Bilingual fake-news detection in low-resource media: A transformer-based framework for nepali–english content. *Journal of Innovations in Engineering Education*, 8(1):133–138.

Saroj Giri, Shiva Ram Dam, Rajesh Kamar, and Suraj Basant Tulachan. 2024. Fake news detection using convolutional neural networks. *Technical Journal*, 4(1):64–68.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Anusha Kamath, Kanishk Singla, Rakesh Paul, Raviraj Bhuminand Joshi, Utkarsh Vaidya, Sanjay Singh Chauhan, and Niranjan Wartikar. 2025. Benchmarking hindi llms: A new suite of datasets and a comparative analysis. In *Proceedings of the 1st Workshop on Benchmarks, Harmonization, Annotation, and Standardization for Human-Centric AI in Indian Languages (BHASHA 2025)*, pages 52–68.

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*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ayush Maheshwari, Kaushal Sharma, Vivek Patel, and Aditya Maheshwari. 2025. Indicparam: Benchmark to evaluate llms on low-resource indic languages. *arXiv preprint arXiv:2512.00333*.

Miriam J Metzger. 2007. Making sense of credibility on the web: Models for evaluating online

information and recommendations for future research. *Journal of the American society for information science and technology*, 58(13):2078–2091.

Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an infodemic: Covid-19 fake news dataset. In *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*, pages 21–29. Springer.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. Indicgenbench: A multilingual benchmark to evaluate generation capabilities of llms on indic languages. *arXiv preprint arXiv:2404.16816*.

Briony Swire-Thompson and David Lazer. 2020. Public health and online misinformation: challenges and recommendations. *Annual review of public health*, 41:433–451.

Surendrabikram Thapa, Kritesh Rauniyar, Hariram Veeramani, Aditya Shah, Imran Razzak, and Usman Naseem. 2024. Did you tell a deadly lie? evaluating large language models for health misinformation identification. In *International Conference on Web Information Systems Engineering*, pages 391–405. Springer.

Emily K Vraga and Leticia Bode. 2020. Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation. *Political Communication*, 37(1):136–144.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

John Zarocostas. 2020. How to fight an infodemic. *The lancet*, 395(10225):676.

Zhihao Zhang, Yiran Zhang, Xiyue Zhou, Liting Huang, Imran Razzak, Preslav Nakov, and Usman Naseem. 2025. From generation to detection: A multimodal multi-task dataset for benchmarking health misinformation. *arXiv preprint arXiv:2505.18685*.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–20.