Probing the Limits of Multilingual Language Understanding: Low-Resource Language Proverbs as LLM Benchmark for AI Wisdom

Surendrabikram Thapa[1,*], Kritesh Rauniyar[2,3,*], Hariram Veeramani[4], Surabhi Adhikari[5], Imran Razzak[6], Usman Naseem[7]
[1]Virginia Tech, USA, [2]Delhi Technological University, India,
[3]IIMS College, Nepal, [4]UCLA, USA, [5]Columbia University, USA,
[6]Mohamed bin Zayed University of Artificial Intelligence, UAE,
[7]Macquarie University, Australia

Abstract

Understanding and interpreting culturally specific language remains a significant challenge for multilingual natural language processing (NLP) systems, particularly for less-resourced languages. To address this problem, this paper introduces PRONE[1], a novel dataset of 2,830 Nepali proverbs, and evaluates the performance of various language models (LMs) in two tasks: (i) identifying the correct meaning of a proverb from multiple choices, and (ii) categorizing proverbs into predefined thematic categories. The models, including both open-source and proprietary, were tested in zero-shot and few-shot settings with prompts in English and Nepali. While models like GPT-4o demonstrated promising results and achieved the highest performance among LMs, they still fall short of human-level accuracy in understanding and categorizing culturally nuanced content, highlighting the need for more inclusive NLP.

1 Introduction

Language is a powerful medium for conveying culture, traditions, and shared human experiences. Training language models (LMs) to learn multiple languages and contexts can significantly enhance their ability to understand diverse human perspectives and communicate across cultural boundaries (Li et al., 2024; Hu et al., 2020; Thapa et al., 2025). While this can enable more inclusive and globally aware AI systems, it also presents substantial challenges in accurately capturing the unique nuances, idioms, and culturally specific references that vary widely between languages and societies (Liu et al., 2025; Agarwal et al., 2025; Aleem et al., 2024; Tao et al., 2024; Myung et al., 2024; Pawar et al., 2025).

For instance, what is considered common knowledge in one culture may not hold the same relevance in another. A phrase like `watching the ball drop' immediately invokes the image of New Year's Eve in Times Square for those familiar with American culture. At the same time, it may mean nothing to someone from a different cultural background. Similarly, in Japan, a *Hanami* or `flower-viewing party' carries deep cultural significance associated with cherry blossoms in spring. In contrast, it might simply be interpreted as a generic gathering in other parts of the world. Proverbs are another prime example of how deeply language is intertwined with culture. Unlike general phrases or idioms, proverbs frequently rely on metaphors, analogies, and references unique to their origin (Kordoni, 2018; Qiang et al., 2023; Verma and Vuppuluri, 2015; Abebe Fenta and Gebeyehu, 2023). They are not just linguistic expressions but also cultural artifacts that reflect the lived experiences and shared understanding of a community.

For example, the proverb 'शङ्कर सहायता गर्छन् त भयङ्करको के पिर' (If Shankar is helpful, then what is there to fear?), reflects a deeply rooted cultural belief in divine protection and faith. In Hinduism, Shankar (name for Lord Shiva) is revered as a powerful god, and the proverb suggests that if a divine force is on one's side, there is no need to worry about any dangers or challenges. For a language model unfamiliar with Hindu deities or the cultural context of Nepal, the significance of this proverb would likely be misunderstood or lost. For instance, the model might interpret `Shankar' as a common proper name for a person rather than recognizing it as a reference to Lord Shiva. Thus, it is crucial to develop language models that are not only proficient in multiple languages but also attuned to the cultural contexts and nuances that shape the meaning of expressions, idioms, and proverbs. While recent advancements in multilingual NLP for cultural understanding have focused on major languages like Hindi, Chinese, and Span-

[*] These authors contributed equally to this work and are listed as joint first authors.

[1] https://github.com/therealthapa/prone

Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025), pages 120–129

ish (Hu et al., 2020; Kakwani et al., 2020; Baucells et al., 2025), there remains a considerable gap when it comes to less-resourced languages such as Nepali (Thapa et al., 2024; Rauniyar et al., 2023). To address this gap, we introduce **PRONE**, a novel dataset of 2,830 Nepali proverbs and evaluate the performance of large language models (LLMs) in interpreting and categorizing them accurately. Our contributions are:

- We introduce **PRONE**, a manually curated novel dataset of 2,830 **PRO**verbs in **NE**pali, reflecting diverse cultural expressions and wisdom unique to Nepali.
- We manually classify the proverbs into five broad categories, capturing key themes and contextual nuances.
- We benchmark the performance of LLMs on two specific tasks: **Task A:** Evaluating the ability of LLMs to correctly identify the meaning of a proverb from a set of options consisting of one correct and three incorrect choices. **Task B:** Assessing the capacity of LLMs to accurately categorize the proverbs into predefined categories.

By focusing on Nepali proverbs, our study supports the United Nations Sustainable Development Goal (SDG) of *`Leave No One Behind'* by promoting linguistic inclusivity and cultural representation in AI.

2 Related Works

Prior research in figurative languages, such as metaphor detection, generation, and interpretation, has employed various approaches, including linguistic and visual embeddings, context-based analysis, and paraphrasing tasks, which are also relevant to understanding proverbs (Pramanick et al., 2018; Chakrabarty et al., 2021; Bizzoni and Lappin, 2018; Wachowiak and Gromann, 2023; Liu et al., 2022). Goren and Strapparava (2024) examine GPT-3.5's ability to detect word-level metaphors in proverbs using different prompting strategies. They expand the PROMETHEUS dataset (Özbal et al., 2016) with hypothetical contexts and test three prompting approaches. The results show that the model performs best with hypothetical context, followed by first providing the proverb's meaning. Similarly, there have been efforts to enhance language models' understanding of cultural and linguistic nuances, such as the work by Wibowo et al. (2024), who developed COPAL-ID, a dataset tailored for commonsense reasoning in Indonesian. They experiment with different LLMs, including open-source models such as XLM-R (Conneau et al., 2020), BLOOMZ (Muennighoff et al., 2023b), and PolyLM (Wei et al., 2023), as well as proprietary models such as ChatGPT and GPT-4, to evaluate their ability to handle the cultural and linguistic nuances embedded in the COPAL-ID dataset. Their findings indicate that while proprietary models like GPT-4 achieve relatively higher accuracy, they still fail human-level performance in understanding local nuances.

Expanding on the theme of evaluating the understanding of language models of culturally nuanced language, Liu et al. (2024) investigated the abilities of various language models, such as BLOOMZ (Muennighoff et al., 2023b), LLaMA-2 (Touvron et al., 2023), XGLM (Lin et al., 2022), XLM-R (Conneau et al., 2020), and mT0 (Muennighoff et al., 2023a), in reasoning with proverbs and sayings across different cultures. They evaluated these models using culturally diverse proverbs in six languages (English, German, Russian, Bengali, Mandarin Chinese, and Indonesian). Their findings showed that while these models could memorize proverbs to some extent, they often struggled to understand them in conversational contexts, particularly when dealing with figurative language and cross-cultural translations. However, these studies have primarily focused on high-resource languages, leaving less-resourced languages like Nepali largely unexamined. Our work is the first to address this gap, introducing a novel dataset of 2,830 Nepali proverbs and evaluating the ability of LLMs to interpret and categorize them effectively.

3 Dataset

We created a dataset of 2,830 Nepali proverbs collected from various sources, including online databases, literature, and local cultural repositories. The primary collection relied on three subject matter experts (SMEs), each with at least a master's degree in fields related to Nepali language, literature, or culture. The collected proverbs were checked among the SMEs to filter out any proverbs that were deemed irrelevant.

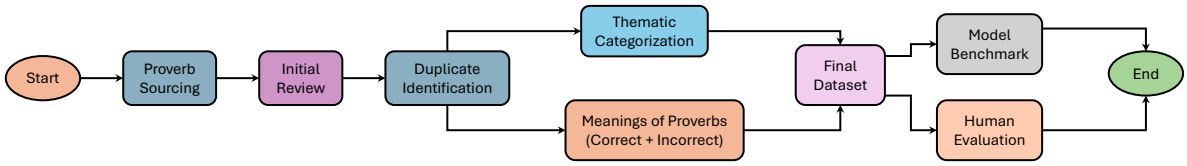

Figure 1: Overview of the End-to-End Pipeline for Annotating and Evaluating Nepali Proverbs

3.1 Deduplication

For deduplication, we used NepBERTa (Timilsina et al., 2022), a pre-trained language model, to generate embeddings, which capture the semantic meaning of each proverb. We then compute the cosine similarity between these embeddings to measure the similarity between pairs of proverbs. Using this approach, we identify semantically similar proverbs and treat them as near-duplicates. We manually visit the near-duplicates and remove if there are redundant proverbs.

3.2 Thematic Categorization

To categorize the proverbs, we manually annotated them into five categories: (i) Social Behavior and Relationships, (ii) Fate and Caution, (iii) Hard Work and Perseverance, (iv) Wisdom and Knowledge, and (v) Nature and Environment; using a rigorous annotation criterion (Appendix A). Each proverb was annotated by three annotators, and the final category was determined by majority agreement; in cases where all three annotators assigned different categories, the disagreement was resolved through a consensus Zoom meeting.

3.3 Final Dataset

For each proverb, as shown in Figure 1, we assigned one correct meaning and three plausible but incorrect meanings to test the interpretative capabilities of language models. The final dataset thus consists of proverbs categorized into five thematic groups (Table 1) and accompanied by multiple-choice options for their meanings.

Category	Proverbs
Social Behavior and Relationships	1274
Fate and Caution	1177
Hard Work and Perseverance	200
Wisdom and Knowledge	157
Nature and Environment	22
Total	**2830**

Table 1: Distribution of Nepali Proverbs.

4 Experimental Setup

4.1 Language Models

To evaluate the understanding and categorization of Nepali proverbs, we employed a range of language models, including open-source and proprietary ones. We conducted experiments in both zero-shot and few-shot settings for all models, prompting the models in both English and Nepali languages (prompts in Appendix B). The models evaluated included: **BERT-based LMs**: DistillBERT-Ne (Shrestha, 2023), RoBERTa-Ne (Chaudhary, 2023), NepBERTa (Timilsina et al., 2022), NepaliBERT (Ghimire, 2023), NepNewsBERT (Pudasaini, 2023). **Closed/ Proprietary Models**: GPT-3.5, GPT-4, GPT-4o (OpenAI, 2023), Gemini Pro 1.5, Gemini Flash 1.5, Mistral Medium (Mistral AI, 2024). **Open-sourced Models**: LLaMA-2 (7B) (Touvron et al., 2023), Mistral (7B) (Jiang et al., 2023), Gemma (7B) (Mesnard et al., 2024).

4.2 Evaluation Metrics

For Task A, we used accuracy to measure the proportion of correct selections by the LLMs from a set of options (one correct and three incorrect) as it directly reflects the models' ability to identify the correct meaning. Similarly, for Task B, we employed the F-score to evaluate the models' performance in categorizing proverbs into predefined categories, as it balances precision and recall, addressing the imbalanced distribution of categories.

5 Results and Discussion

Table 2 shows the performance of language models in Task A. Among all the models used, GPT-4o consistently performs the best across all the settings. The results show that across all model types, performance in the few-shot (FS) setting is consistently higher than in the zero-shot (ZS) setting, reflecting the benefit of additional context or

	Model	ZS-En	ZS-Ne	FS-En	FS-Ne
BERT Based LMs	DistillBERT-Ne	33.72	26.19	40.56	43.16
	RoBERTa-Ne	35.97	28.45	43.19	44.05
	NepaliBERT	36.41	29.71	45.87	45.87
	NepBERTa	38.67	32.84	45.34	47.64
	NepNewsBERT	40.41	35.76	50.23	50.18
Open Source	LLaMA-2	57.33	48.91	65.20	62.46
	Mistral	58.87	46.42	64.90	61.52
	Gemma	60.19	52.74	68.95	66.36
Closed/ Proprietary	GPT-3.5	14.88	15.62	22.14	25.09
	GPT-4	68.57	59.58	76.54	79.65
	GPT-4o	**80.92**	74.63	**86.19**	**87.99**
	Gemini Pro 1.5	79.93	**75.02**	83.72	83.75
	Gemini Flash 1.5	66.25	62.93	83.72	85.12
	Mistral Medium	17.14	53.57	24.39	50.99
Human Annotator		**95.17**			

Table 2: Accuracy of Different Language Models on Task A (Proverb Meaning Identification) Across Zero-Shot (ZS) and Fine-Tuned (FS) Settings in English (En) and Nepali (Ne).

	Model	ZS-En	ZS-Ne	FS-En	FS-Ne
BERT Based LMs	DistillBERT-Ne	31.87	21.05	32.05	34.09
	RoBERTa-Ne	33.86	23.74	32.98	35.22
	NepaliBERT	35.91	24.56	36.71	36.73
	NepBERTa	36.24	26.73	38.56	39.67
	NepNewsBERT	38.17	28.42	40.37	40.02
Open Source	LLaMA-2	46.88	37.76	45.37	41.59
	Mistral	44.61	35.74	44.23	40.38
	Gemma	49.87	39.04	45.37	42.42
Closed/ Proprietary	GPT-3.5	26.02	9.18	42.73	31.23
	GPT-4	50.95	43.90	49.53	47.29
	GPT-4o	**84.52**	**53.22**	**74.66**	**63.90**
	Gemini Pro 1.5	47.40	6.68	31.77	20.50
	Gemini Flash 1.5	58.18	12.69	55.67	52.93
	Mistral Medium	28.67	11.57	23.50	34.51
Human Annotator		**88.74**			

Table 3: Performance (F-score) of Various Language Models on Task B (Proverb Categorization) in Zero-Shot (ZS) and Few-Shot (FS) Settings for English (En) and Nepali (Ne).

examples. For example, the accuracy of BERT-based models improves from 26.19%-40.41% in the zero-shot setting to 43.16%-50.23% in the few-shot setting. Open-source models also show notable improvements with fine-tuning, where accuracy increases from 46.42%-60.19% in zero-shot to 61.52%-68.95% in few-shot. Similarly, closed/proprietary models such as GPT-4 and GPT-4o achieve much higher accuracy in few-shot settings, with GPT-4o reaching 87.99% compared to 74.63% in the zero-shot setting.

Table 3 presents the performance of various language models on Task B. The results indicate varying levels of performance across models and settings. BERT-based models show modest F-scores, ranging from 21.05% (DistillBERT-Ne, ZS-Ne) to 40.37% (NepNewsBERT, FS-En), with a slight improvement observed in the few-shot setting compared to zero-shot. Open-source models demonstrate moderate performance, with F-scores ranging from 35.74% (Mistral, ZS-Ne) to 49.87% (Gemma, ZS-En), indicating some capacity to handle the proverb categorization task. However, they do not reach the highest scores. Closed/proprietary models exhibit a wider range of F-scores, from as low as 6.68% (Gemini Pro 1.5, ZS-Ne) to as high as 84.52% (GPT-4o, ZS-En). Among these, GPT-4o consistently achieves the best performance, with the highest F-scores across all settings, particularly in the zero-shot English setting (84.52%) and the few-shot English setting (74.66%).

5.1 Human Evaluation

We also performed a human evaluation on both Tasks A and B to compare the performance of LLMs against human understanding. We employed a different set of three native Nepali speakers as annotators, each with at least a school-level education in Nepali. In Task A, human annotators achieved an accuracy of 95.17%, while in Task B, they obtained an F-score of 88.74%. These high-performance metrics indicate that these tasks are relatively straightforward for native speakers.

6 Conclusion

We evaluated various LLMs' abilities to understand and categorize Nepali proverbs using a novel dataset of 2,830 proverbs. While some models, such as GPT-4o, showed promising results, their performance still lags behind human annotators, who achieved the highest F-score of 95.17% and 88.74% in task A and task B, respectively. The gap highlights the need for further improvement in handling culturally specific content, particularly for less-resourced languages like Nepali. Future research should enhance models' understanding of diverse linguistic contexts to achieve more culturally inclusive NLP systems.

Limitations

While our study offers valuable insights into the performance of language models on the PRONE dataset, several limitations must be addressed.

First, the dataset, though substantial with 2,830 Nepali proverbs, may not encompass the full spectrum of cultural and contextual nuances inherent in Nepali language use. The limited scope of proverbs may restrict the models' ability to generalize across a broader range of culturally specific expressions. Second, despite promising results from models like GPT-4o, a noticeable gap remains compared to human annotators, highlighting challenges in achieving full cultural comprehension and accurate categorization. This indicates a need for enhanced training methods, possibly involving more diverse cultural data and improved model adaptation techniques. Additionally, our evaluation using English and Nepali prompts in zero-shot and few-shot settings may not fully capture the models' potential in varied real-world applications. Future work should explore alternative approaches, such as fine-tuning culturally rich datasets and developing hybrid models, to improve understanding and performance in less-resourced languages.

Ethics Statement

Data Collection and Privacy: The PRONE dataset of Nepali proverbs was created using publicly available sources, ensuring no personal or sensitive data was involved. We complied with all relevant data protection guidelines and model usage terms, focusing solely on non-commercial research. While the dataset aims to advance culturally inclusive NLP, we acknowledge the potential for biases in model outputs and caution against misuse that could reinforce cultural stereotypes. Comprehensive documentation is provided, but researchers should be aware of the dataset's limitations and apply it responsibly in diverse contexts.

Annotators Recruitment: The human annotators for this study were recruited at the local prevailing rate, ensuring fair compensation for their contributions. We adhered to ethical recruitment practices, and there were no ethical issues identified in this process. The annotators' native proficiency and cultural understanding were essential to the study, enhancing the quality and accuracy of the evaluations conducted.

References

Anduamlak Abebe Fenta and Seffi Gebeyehu. 2023. Automatic idiom identification model for amharic language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(8):1--9.

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1--21.

Mahwish Aleem, Imama Zahoor, and Mustafa Naseem. 2024. Towards culturally adaptive large language models in mental health: Using chatgpt as a case study. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, pages 240--247.

Irene Baucells, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, Jose Javier Saiz, Robiert Sepulveda Torres, Jeremy Barnes, Pablo Gamallo, Aitor Gonzalez-Agirre, German Rigau, and Marta Villegas. 2025. IberoBench: A benchmark for LLM evaluation in Iberian languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491--10519, Abu Dhabi, UAE. Association for Computational Linguistics.

Yuri Bizzoni and Shalom Lappin. 2018. Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the workshop on figurative language processing*, pages 45--55.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250--4261, Online. Association for Computational Linguistics.

Amit Chaudhary. 2023. roberta-base-ne: A roberta-based language model for nepali. Accessed: 2024-09-16.

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, Online. Association for Computational Linguistics.

Rajan Ghimire. 2023. Nepalibert: A bert-based language model for nepali. Accessed: 2024-09-16.

Gamze Goren and Carlo Strapparava. 2024. Context matters: Enhancing metaphor recognition in proverbs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3825--3830.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411--4421. PMLR.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, 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.

Valia Kordoni. 2018. Beyond multiword expressions: Processing idioms and metaphors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 15--16.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019--9052.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016--2039.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652--689.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437--4452.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Mistral AI. 2024. Mistral ai models: Getting started guide. Accessed: 2024-09-16.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023a. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991--16111, Toronto, Canada. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. 2023b. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991--16111.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104--78146.

OpenAI. 2023. Gpt (generative pre-trained transformer). `https://openai.com/research/gpt`. Accessed: 2024-09-16.

Gözde Özbal, Carlo Strapparava, and Serra Sinem Tekiroğlu. 2016. Prometheus: A corpus of proverbs annotated with metaphors. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3787--3793.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1--96.

Malay Pramanick, Ashim Gupta, and Pabitra Mitra. 2018. An lstm-crf based approach to token-level metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 67--75.

Shushant Pudasaini. 2023. Nepnewsbert: A bert-based language model for nepali news classification. Accessed: 2024-09-16.

Jipeng Qiang, Yang Li, Chaowei Zhang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2023. Chinese idiom paraphrasing. *Transactions of the Association for Computational Linguistics*, 11:740--754.

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.

Dipesh Shrestha. 2023. Nepali-distilbert: A distilbert-based language model for nepali. Accessed: 2024-09-16.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

Surendrabikram Thapa, Kritesh Rauniyar, Ehsan Barkhordar, Hariram Veeramani, and Usman Naseem. 2024. Which side are you on? investigating politico-economic bias in nepali language models. In *Annual Workshop of the Australasian Language Technology Association (22nd: 2024)*, pages 104--117. Association for Computational Linguistics (ACL).

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Surabhi Adhikari, Kengatharaiyer Sarveswaran, Bal Krishna Bal, Hariram Veeramani, and Usman Naseem. 2025. Natural language understanding of devanagari script languages: Language identification, hate speech and its target detection. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 71--82.

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*. Association for Computational Linguistics (ACL).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Rakesh Verma and Vasanthi Vuppuluri. 2015. A new approach for idiom identification using meanings and the web. In *Proceedings of the international conference recent advances in natural language processing*, pages 681--687.

Lennart Wachowiak and Dagmar Gromann. 2023. Does gpt-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018--1032.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

Haryo Wibowo, Erland Fuadi, Made Nityasya, Radityo Eko Prasojo, and Alham Aji. 2024. Copal-id: Indonesian language reasoning with local culture and nuances. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1404--1422.

A Annotation Details

To categorize the proverbs, we followed the following definitions and criteria:

1. **Social Behavior and Relationships**: Proverbs that focus on human interactions, social conduct, community norms, and interpersonal relationships. This category includes proverbs that highlight themes such as trust, deceit, friendship, love, and societal roles.

   Criteria: The proverb must relate to behaviors, expectations, or dynamics between individuals or groups within a social context.
   Examples: "जस्तो कुकुर, उस्तै पुच्छर।" (`Like the dog, like its tail') — emphasizes consistent behavior or traits, and "छिमेकीको घरमा आगो लाग्दा आफ्नो घर सुरक्षित छैन।" (`When a neighbor's house is on fire, your own house is not safe') — reflects interdependence in community relations.

2. **Fate and Caution**: Proverbs that deal with themes of destiny, luck, and the importance of caution or mindfulness in life. This category encompasses advice on being prudent, prepared, or aware of one's circumstances and external forces.

   Criteria: The proverb must convey messages related to fate, destiny, or the necessity of being cautious or aware in various situations.
   Examples: "चोक्टा पर्छु, ढुंगामा लाग्छु।" (`If I fall, I fall on a rock') — suggests the inevitability of misfortune, and "सर्पसँग दोस्ती गर्नु खतरा हुन्छ।" (`Friendship with a snake is dangerous') — emphasizes the need for caution in relationships.

Proverb	True Meaning	Incorrect Meaning	Note
श्राद्ध गर्न त सजिलो तर सिधा पुर्याउन गाह्रो	ठूलो कामभन्दा सानो काम गर्न गाह्रो हुन्छ । (It is easier to do big things when smaller things are in place.)	श्राद्ध गर्दा धेरै मानिसहरूलाई बोलाउनुपर्छ (You have to call a lot of people in funeral).	In Nepali culture, श्राद्ध (Shraddha) refers to an important ritual to honor deceased ancestors, where the ceremony may seem straightforward, but strict attention to every detail is vital and often more challenging.
कामकुरो एकातिर, कुम्लो बोकी ठिमीतिर।	एकातिरको गर्नुपर्ने काम छाडी अर्कातिर लाग्ने चाला (Neglecting task that needs to be done and instead tending toward something else.)	मानिसहरू सधैं आफ्नो बोझ अरूलाई दिन्छन् (People always tend to give their burden to others.)	Thimi is a town in Nepal, and in this context, it represents a place that is irrelevant to the original task or goal.
मियाँको दौड मस्जिदसम्म	व्यक्तिको प्रयास वा महत्वाकांक्षा सीमित दायरामा मात्र सीमित छ। (Person's efforts or ambitions are limited to a narrow range or familiar routine)	धार्मिक कर्तव्य पूरा गर्न दौडधुप गर्ने (To actively run around to fulfill the religious duties)	The use of मियाँ (Miyan), a respectful term for a Muslim man, and मस्जिद (mosque), denotes a routine or habitual practice.
'शङ्कर सहायता गर्छन् त भयङ्करको के पिर'	'यदि भगवान् साथ दिन्छन् भने डरको कुनै कारण छैन।' (If Lord is on your side, there is no reason to fear.)	यदि साथीले मद्दत गर्छ भने डराउनु पर्दैन। (If a friend helps, there is no need to be afraid.)	"शङ्कर" (Shankar) can be both a common proper name for a person and a reference to Lord Shiva in Hinduism; thus, without context, the proverb could be mistakenly interpreted as referring to an ordinary person's help rather than invoking divine protection.

Table 4: Examples of Nepali Proverbs with Their True and Incorrect Meanings, Along with Notes on Potential Misinterpretations

3. **Hard Work and Perseverance**: Proverbs that highlight the value of diligence, effort, endurance, and resilience in overcoming challenges or achieving goals. These proverbs often carry motivational or inspirational messages.

   Criteria: The proverb should focus on themes of hard work, persistence, or the rewards of sustained effort and commitment.
   Examples: "हलो जोते मुरी फल्छ, घाँस खाए गोरु मर्छ।" (`Plowing the field yields a harvest; eating the grass kills the ox') — underscores the benefits of hard work, and "धेरै मेहनत नगरी कुनै चीज प्राप्त हुँदैन।" (`Without much effort, nothing is obtained') — stresses the necessity of perseverance.

4. **Wisdom and Knowledge**: Proverbs that offer guidance or insights about life, learning, and understanding. These proverbs often reflect collective wisdom, experience, or philosophical reflections on human behavior or morality.

   Criteria: The proverb should convey a lesson or insight related to knowledge, learning, or the deeper understanding of life.
   Examples: "ज्ञान नै शक्ती हो।" (`Knowledge is power') — emphasizes the importance of wisdom, and "साँढेको आँसु दूध हुँदैन।" (`A bull's tears are not milk') — encourages recognizing reality and not being swayed by appearances.

5. **Nature and Environment**: Proverbs that use elements of nature (such as animals, plants, weather, or landscapes) to convey lessons or truths. These proverbs employ natural metaphors to illustrate human behavior, morality, or life lessons.

   Criteria: The proverb must use imagery from the natural world to communicate its message or lesson.
   Examples: "हावा खाएको जस्तै गर्छ, पानी परेको जस्तै भिजाउँछ।" (`It moves like the wind, wets like the rain') uses elements of nature to describe inevitability or impact, and "ओखती जति तीतो हुन्छ, रोग त्यति नै राम्रो हुन्छ।" (`The more bitter the medicine, the better the cure') draws on natural elements to illustrate a life lesson.

Using these definitions and criteria, we ensured that each proverb was categorized accurately, reflecting its central theme and underlying cultural context. This approach allowed us to create a well-defined dataset that can be effectively used to evaluate the performance of large language models in understanding culturally specific content.

B Prompt Templates

Example of Zero-shot Prompt in English for Meaning

Select the correct meaning for the given proverb among given options: `Proverb: _____ Options: A. _____ B. _____ C. _____ D. _____`. Only output the correct option as `A', `B', `C' or `D'. Explanations are not needed.

Example of Zero-shot Prompt in Nepali for Meaning

तल दिइएको उखानको सही उत्तर दिनुहोस् । कृपया `A', `B', `C' वा `D' मध्य सही विकल्पहरु मात्र उत्तर दिनुहोस् । व्याख्या नगर्नुहोस् । उखान: _____ विकल्पहरु: `A. _____ B. _____ C. _____ D. _____`.

Example of Few-shot Prompt in English for Meaning

Select the correct meaning for the given Nepali proverb among the given options. Only output the correct option as `A', `B', `C', or `D'. Explanations are not needed. Example 1: Proverb: अवसर चुकेपछि के काम,मौकामा नै काम गर्नुपर्छ। Options: A. अवसर चुकाउनु राम्रो हुन्छ, B. अवसर चुकाउनु भनेको सफलताको संकेत हो, C. अवसर चुकाएपछि अर्को अवसर आउँदैन, D. मौकामा नै काम गर्नुपर्छ Correct Option: D; Example 2: Proverb: 'भेडा भेडासँग, बाख्रा बाख्रासँग' Options: A. सबै प्राणी आफ्ना–आफ्ना जातिसँग मिल्छन्।, B. भेडा र बाख्रा कहिल्यै सँगै बस्न सक्दैनन् , C. भेडा र बाख्राको सम्बन्ध झगडालु हुन्छ, D. भेडा र बाख्राको बिचमा सधैं प्रतिस्पर्धा हुन्छ Correct Option: A; Example 3: Proverb: 'बाहिरका ठूला, भित्रका लुला' Options: A. देख्नलाई मात्र निकै भएजस्ता तर मनका फितला ।, B. बाहिर राम्रो देखिने तर भित्र कमजोर, C. बाहिर धनी तर भित्र गरिब, D. बाहिर बोल्न सक्ने तर भित्र डराउने, Correct Option: B; Example 4: Proverb: औषधी र उपदेश मीठो हुन्न । Options: A. जब तपाईं खराब अवस्थामा हुनुहुन्छ, प्रभावकारी सुझाव र समाधानहरू राम्रो वा सजिलो नलाग्न सक्छ।, B. औषधी र उपदेश सधैं तीतो हुन्छ, C. औषधी र उपदेश सधैं बेकार हुन्छ, D. औषधी र उपदेशले कहिल्यै काम गर्दैन Correct Option: A; Example 5: Proverb: पहिलो गाँसमै ढुङ्गा । Options: A. नराम्रो शुरुवात हुनु ।, B. पहिलो गाँसमै ढुङ्गा भनेको खाना पकाउन नजान्नु हो, C. पहिलो गाँसमै ढुङ्गा भनेको सधैं असफल हुनु हो, D. पहिलो गाँसमै ढुङ्गा भनेको ढुङ्गा खानु हो Correct Option: A; Now, select the correct meaning for the given Nepali proverb. Proverb: _____ Options: `A. _____ B. _____ C. _____ D. _____`.

Example of Few-shot Prompt in Nepali for Meaning

दिइएको नेपाली उखानको तल दिइएको विकल्पहरु मध्य सही अर्थ भएको विकल्प छान्नुहोस् । कृपया 'A', 'B', 'C' वा 'D' मध्य सही विकल्पहरु मात्र उत्तर दिनुहोस् । व्याख्या नगर्नुहोस् । उदाहरण १: उखान: अवसर चुकेपछि के काम,मौकामा नै काम गर्नुपर्छ । विकल्पहरु: A. अवसर चुकाउनु राम्रो हुन्छ, B. अवसर चुकाउनु भनेको सफलताको संकेत हो, C. अवसर चुकाएपछि अर्को अवसर आउँदैन, D. मौकामा नै काम गर्नुपर्छ सही विकल्प: D; उदाहरण २: उखान: 'भेडा भेडासँग, बाख्रा बाख्रासँग' विकल्पहरु: A. सबै प्राणी आफ्ना–आफ्ना जातिसँग मिल्छन्।, B. भेडा र बाख्रा कहिल्यै सँगै बस्न सक्दैनन् , C. भेडा र बाख्राको सम्बन्ध झगडालु हुन्छ, D. भेडा र बाख्राको बिचमा सधैं प्रतिस्पर्धा हुन्छ सही विकल्प: A; उदाहरण ३: उखान: 'बाहिरका ठूला, भित्रका लुला' विकल्पहरु: A. देख्नलाई मात्र निकै भएजस्ता तर मनका फितला ।, B. बाहिर राम्रो देखिने तर भित्र कमजोर, C. बाहिर धनी तर भित्र गरिब, D. बाहिर बोल्न सक्ने तर भित्र डराउने, सही विकल्प: B; उदाहरण ४: उखान: औषधी र उपदेश मीठो हुन्न । विकल्पहरु: A. जब तपाईं खराब अवस्थामा हुनुहुन्छ, प्रभावकारी सुझाव र समाधानहरू राम्रो वा सजिलो नलाग्न सक्छ।, B. औषधी र उपदेश सधैं तीतो हुन्छ, C. औषधी र उपदेश सधैं बेकार हुन्छ, D. औषधी र उपदेशले कहिल्यै काम गर्दैन सही विकल्प: A; उदाहरण५: उखान: पहिलो गाँसमै ढुङ्गा । विकल्पहरु: A. नराम्रो शुरुवात हुनु ।, B. पहिलो गाँसमै ढुङ्गा भनेको खाना पकाउन नजान्नु हो, C. पहिलो गाँसमै ढुङ्गा भनेको सधैं असफल हुनु हो, D. पहिलो गाँसमै ढुङ्गा भनेको ढुङ्गा खानु हो सही विकल्प: A; अब दिइएको उखानको सही विकल्प उत्तर दिनुहोस् । उखान: _____ विकल्पहरु: `A. _____ B. _____ C. _____ D. _____`.

Example of Zero-shot Prompt in English for Proverb Category

Classify the following Nepali proverb into one of the five categories: {`Wisdom and Knowledge', `Hard Work and Perseverance', `Social Behavior and Relationships', `Nature and Environment', `Fate and Caution'} Proverb: _____. Provide only the category name that best fits the meaning of the given proverb. No explanation is needed.

Example of Zero-shot Prompt in Nepali for Proverb Category

तल दिएको नेपाली उखानलाई पाँचमध्ये कुनै एक वर्गमा वर्गीकृत गर्नुहोस्: 'ज्ञान र बुद्धि', 'मेहनत र धैर्यता', 'सामाजिक व्यवहार र सम्बन्धहरू', 'प्रकृति र वातावरण', 'भाग्य र सावधानी'. उखान: _____ उखानको अर्थसँग सबैभन्दा राम्रोसँग मिल्ने वर्गको नाम मात्र प्रदान गर्नुहोस्। कुनै स्पष्टीकरण आवश्यक छैन।

Example of Few-shot Prompt in English for Proverb Category

Classify the following Nepali proverb into one of the five categories: `Wisdom and Knowledge', `Hard Work and Perseverance', `Social Behavior and Relationships', `Nature and Environment', `Fate and Caution'. Provide only the category name that best fits the meaning of the given proverb. Example 1: Proverb: पाप धुरीबाट कराउँछ Category: Fate and Caution ; Example 2: Proverb: आधा गाग्रो छचल्किन्छ Category: Wisdom and Knowledge; Example 3: Proverb: बाहिरका ठूला, भित्रका लुला Category: Social Behavior and Relationships; Example 4: Proverb: अल्छे तिघ्रो, स्वादे जिब्रो Category: Hard Work and Perseverance Example 5: Proverb: वनको चरो वनैमा रमाउँछ Category: Nature and Environment Now, classify the following proverb: Proverb: _____

Example of Few-shot Prompt in Nepali for Proverb Category

तल दिएको नेपाली उखानलाई पाँचमध्ये कुनै एक वर्गमा वर्गीकृत गर्नुहोस्: 'ज्ञान र बुद्धि', 'मेहनत र धैर्यता', 'सामाजिक व्यवहार र सम्बन्धहरू', 'प्रकृति र वातावरण', 'भाग्य र सावधानी'. उखानको अर्थसँग सबैभन्दा राम्रोसँग मिल्ने वर्गको नाम मात्र प्रदान गर्नुहोस्। उदाहरण १: उखान: पाप धुरीबाट कराउँछ वर्ग: भाग्य र सावधानी उदाहरण २: उखान: आधा गाग्रो छचल्किन्छ वर्ग: ज्ञान र बुद्धि उदाहरण ३: उखान: बाहिरका ठूला, भित्रका लुला वर्ग: सामाजिक व्यवहार र सम्बन्धहरू उदाहरण ४: उखान: अल्छे तिघ्रो, स्वादे जिब्रो वर्ग: मेहनत र धैर्यता उदाहरण ५: उखान: वनको चरो वनैमा रमाउँछ वर्ग: प्रकृति र वातावरण। अब, तलको उखानलाई वर्गीकृत गर्नुहोस्: उखान: _____