NeCCo: Nepali Cultural Commonsense Benchmark for Large Language Model Evaluation

Sanket Shrestha*, **Raunak Regmi,*** **Sadikshya Ghimire,***
Satyam Rana, Supriya Khadka
Sunway College Kathmandu, Birmingham City University, Kathmandu, Nepal
{sanket_24128469, raunak_regmi_a25, sadikshya_23140745}@sunway.edu.np
{satyam_rana_a25, supriya}@sunway.edu.np

Abstract

Large language models perform strongly on standard evaluations, yet these benchmarks prioritize high-resource languages and culturally dominant knowledge, leaving culture-specific commonsense underexamined. In low-resource languages such as Nepali, everyday communication depends on culturally embedded cues, including kinship hierarchies, ritual practices, food systems, idioms, and honorific distinctions that literal translation often fails to capture. As a result, models that appear competent on global metrics can perform poorly in local contexts. To address this gap, we introduce **NeCCo**, a curated multiple-choice benchmark for culturally situated reasoning across five domains: kinship and social hierarchy; festivals, rituals, and geography; idioms, proverbs, and metaphors; commonsense and daily life; and gastronomy, agriculture, and nature. The dataset was created through structured authoring, cross-review, and normalization, and is released in Devanagari, English, and Romanized formats. We evaluate multiple state-of-the-art LLMs using standardized prompting and controlled decoding. Results show substantial variation: models perform better on globally documented knowledge such as geography, but struggle with relational and linguistically implicit tasks, including extended kinship reasoning and proverb interpretation. The most culturally dense categories expose brittleness and increased hallucination. These findings suggest that multilingual competence requires more than translation coverage and highlight the need for culturally grounded benchmarks and training signals.

Keywords: Cultural Common sense, Low-Resource Languages, Multilingual Evaluation, Benchmark Construction, LLM Evaluation, Cross-Cultural Reasoning.

1. Introduction

Nepali is an Indo-Aryan language spoken by over 17 million people and serves as the official language of Nepal. Despite its widespread use, it is considered a low-resource language in NLP due to its limited presence in large-scale training corpora and evaluation benchmarks (Shahi and Sitaula, 2022). This under-representation leads to insufficient modeling of linguistic variability and culturally grounded nuances in large language models (LLMs) (Nyachhyon et al., 2025). Many multilingual benchmarks either focus on high-resource languages or test translation-equivalent skills, which can obscure whether models actually understand locally grounded meanings in languages such as Nepali (Hu et al., 2020; Liang et al., 2020; Nyachhyon et al., 2025).

A key source of difficulty is cultural common sense: knowledge that speakers routinely use but rarely state explicitly. In Nepali, interpretation often depends on kinship structure and social hierarchy, honorific choice, ritual etiquette, and idiomatic language that does not survive literal translation (Krause, 1980). For example, kinship terms encode distinctions such as paternal versus maternal relations and elder versus younger relatives, and these distinctions shape appropriate address forms and expected behavior. Such cues are sparsely documented in globally available text and are therefore weakly represented in the data that dominates LLM training. As a result, a model can appear strong on general multilingual benchmarks while still failing on culturally specific reasoning required in everyday Nepali use (Thakur, 2025).

This paper targets that evaluation blind spot. We introduce NeCCo, a multiple-choice benchmark designed to test culturally anchored commonsense reasoning in Nepali across five domains: kinship and social hierarchy; festivals, rituals, and geography; idioms, proverbs, and metaphors; commonsense and daily life; and gastronomy, agriculture, and nature. We adopt a multiple-choice format to support controlled comparisons across models and prompt settings, reduce sensitivity to open-ended generation style, and enable fine-grained analysis of systematic confusions. At the same time, the format keeps the evaluation focused on selecting culturally appropriate inferences rather than on fluency or long-form explanation quality.

NeCCo is authored by native speakers and cross-reviewed to reduce ambiguity, with standardized formatting to support consistent evaluation. Using this benchmark, we evaluate a set of state-

*Equal contribution

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 154–168
16 May 2026.

of-the-art LLMs under standardized prompting and decoding, and analyze performance by category to identify where culturally grounded reasoning breaks down (Thapa et al., 2025). We then use these findings to highlight gaps in current multilingual evaluation practice and to motivate more culturally grounded training and assessment.

Our core contributions are

- We introduce **NeCCo**, a culturally anchored multiple-choice benchmark for evaluating Nepali commonsense reasoning across five domains.
- We conduct a systematic evaluation of contemporary LLMs on culturally embedded reasoning tasks, highlighting failures that standard multilingual benchmarks can miss.
- We provide recommendations for future dataset construction and model training that incorporate culturally grounded signals.

2. Related Work

2.1. Multilingual and Cross-Lingual Evaluation of LLMs

A significant portion of LLM evaluation continues to rely on English-centric benchmarks. In these settings, multilingual capability is frequently inferred through translation or limited cross-lingual transfer rather than direct evaluation in the target languages (Clark et al., 2020). Broad multilingual benchmarks, such as XTREME-style evaluation suites and information-seeking datasets like TyDi QA (Clark et al., 2020), were introduced to assess generalization across typologically diverse languages. These evaluations consistently show significant performance degradation outside high-resource settings.

More recent evaluation frameworks, including holistic taxonomies and large collaborative benchmarks like BIG-bench (Kazemi et al., 2025), emphasize the multi-dimensional nature of model quality. In these frameworks, aggregate accuracy can easily obscure failures in robustness, calibration, fairness, and cultural coverage. Multilingual extensions of complex reasoning benchmarks demonstrate that performance drops persist even under few-shot prompting. This indicates that multilinguality is not a uniform capability. Instead, it is a fragile interaction between training data distribution, script variation, and socio-cultural grounding (Clark et al., 2020).

2.2. Commonsense Reasoning Benchmarks

Traditionally, commonsense reasoning benchmarks have focused on general-purpose inference, encompassing causal, temporal, physical, and social dimensions. However, widely used datasets such as CommonsenseQA, SocialIQA, and PIQA (Sap et al., 2019) are primarily constructed in English and implicitly encode Western or globally dominant knowledge priors (Toroghi et al., 2025). Cross-lingual efforts like XCOPA (Ponti et al., 2020) reveal that transferring commonsense tasks across languages exposes distinct linguistic and cultural limitations. In many cases, translation-based approaches actually outperform direct multilingual reasoning (Parashar et al., 2025).

This highlights a major limitation: most existing benchmarks evaluate generalized commonsense rather than culturally situated reasoning. Key dimensions like kinship structures, honorific systems, ritual practices, and idiomatic expressions remain starkly underrepresented. This creates a critical evaluation gap where models appear competent in abstract reasoning but fail in culturally grounded contexts.

2.3. Cultural and Socio-Cultural Bias in LLMs

A parallel body of work investigates cultural and socio-cultural bias in LLMs, demonstrating that models inherently encode and reproduce patterns present in their training data. Foundational benchmarks such as CrowS-Pairs (Nangia et al., 2020) and StereoSet (Nadeem et al., 2021) measure stereotypical associations, BBQ (Parrish et al., 2022) evaluates bias in question answering, and HolisticBias (Zake, 2023) expands this coverage across multiple demographic dimensions. Although essential for assessing fairness, these tools are largely centered on English-speaking or Western contexts. This geographical and linguistic concentration leaves significant gaps in understanding localized harms, such as the specific mechanisms of gender bias in under-represented languages like Nepali (Khadka and Bhattarai, 2025). Emerging work on culturally diverse evaluation highlights that true cultural competence extends beyond simple bias mitigation. It requires a nuanced understanding of implicit norms, localized knowledge systems, and culturally bounded meanings (Piao et al.). When confronted with unfamiliar cultural inputs in these settings, LLMs frequently exhibit hallucination, over-generalization, or default to globally dominant interpretations (Peters et al., 2022).

2.4. Low-Resource and Nepali NLP

Low-resource languages present a unique set of challenges characterized by limited high-quality corpora, sparse benchmark availability, and severe under-representation in large-scale training pipelines. Nepali exemplifies this pattern: despite being spoken by millions, it remains critically under-resourced in terms of standardized evaluation frameworks (Cahyawijaya, 2024). Existing efforts in Nepali NLP have primarily focused on foundational tasks, including part-of-speech tagging (Pradhan and Yajnik, 2024), named entity recognition (Singh et al., 2019), machine translation (Shahi and Sitaula, 2022), text classification (Prabha et al., 2018) and text-to-speech (Khadka et al., 2023). Benchmarks such as Nep-gLUE (Timilsina et al., 2022) and subsequent Nepali NLU datasets (Nyachhyon et al., 2025) have contributed to task-level standardization. Concurrently, monolingual models like NepBERTa have demonstrated improved performance over multilingual baselines in specific settings (Timilsina et al., 2022). Additionally, resources such as FLoRes provide reliable evaluation datasets for Nepali in machine translation (Guzmán et al., 2019). However, these efforts largely target conventional syntactic or semantic tasks. Much of Nepali reasoning is transmitted through oral tradition and lived experience rather than formal corpora, making it particularly difficult for models to learn through these standard, foundational training datasets.

2.5. Toward Culturally Grounded Evaluation

Recent work has increasingly highlighted the limitations of large language models in capturing culturally grounded reasoning, particularly beyond high-resource and globally dominant contexts. While several benchmarks have been proposed for South Asian languages, most focus on broader Indic settings or emphasize translation-based evaluation rather than culturally embedded understanding. Datasets such as SANSKRITI (Maji et al., 2025) and MILU (Verma et al., 2025) evaluate LLMs across multiple Indic languages and domains, demonstrating that model performance drops in culturally rich categories such as rituals, social norms, and traditional knowledge (Joshi et al., 2025). Similarly, large-scale multilingual evaluations like PARIKSHA reveal inconsistencies and reduced reliability of LLMs in low-resource languages, particularly when tasks require implicit reasoning or contextual awareness. However, these efforts largely aggregate across languages and do not account for the unique sociocultural structures of individual languages such as Nepali, as cultural common-sense in Nepali remains underrepresented in existing benchmarks. (Chiu et al., 2025).

3. Methodology

3.1. Dataset Creation

Dataset Overview The Nepali Cultural & Commonsense Benchmark (**NeCCo**) is a curated dataset of 1,295 four-option multiple-choice questions (A to D) designed to evaluate culturally grounded reasoning in Nepali.[1] Each question has a single unambiguous correct answer. The dataset emphasizes implicit cultural knowledge that cannot be inferred through direct translation or surface-level reasoning, and the distractors are written to be culturally plausible without requiring external tools. Each instance in the dataset follows a structured tabular format containing a unique ID, category, question prompt, four options, and the ground truth answer to facilitate systematic evaluation.

Linguistic Representations To study representation sensitivity and translation effects, each question is released in three aligned forms: native Devanagari Nepali, Romanized Nepali, and English translation. Romanized Nepali refers to the phonetic transcription of the native Devanagari script into the Latin alphabet, a format widely used in informal digital communication (Madhani et al., 2023). Unlike standard Devanagari, Romanized text often lacks rigid spelling conventions, introducing natural linguistic variability. For example, the Devanagari word राम्रो (good) is commonly romanized as *ramro*, and a phrase like के छ? (What's up?) might appear variously as *ke cha*, *k cha*, or *k chha*.

We provide English translations for categories where translation does not substantially distort the intended reasoning signal. Accordingly, the *Kinship & Social Hierarchy* and *Idioms, Proverbs & Metaphors* categories intentionally omit English translations to preserve semantic nuance. In total, NeCCo contains 3,460 released items: 1,295 (Devanagari) + 1,295 (Romanized) + 870 (English).[2]

Category Design and Distribution NeCCo is organized into five categories chosen to reflect core dimensions of Nepali life, social organization, ritual practices, and subsistence patterns. The

[1] NeCCo is publicly available at https://github.com/Mezamenta-Nakama09/NeCCo.

[2] The 870 English items correspond to the 1,295 total questions minus the 425 questions in the two non-translated categories.

dataset is approximately balanced, with each category contributing roughly 18 to 22% of the total questions (see Table 1).

Category	Questions
Kinship & Social Hierarchy	235
Festivals, Rituals & Geography	382
Idioms, Proverbs & Metaphors	190
Commonsense & Daily Life	200
Gastronomy, Agriculture & Nature	288

Table 1: Number of questions per category in NeCCo.

- **Kinship & Social Hierarchy:** Tests understanding of intricate familial relations, hierarchical distinctions, honorific usage, and lineage (maternal vs. paternal) reflecting Nepal's social structure.
- **Festivals, Rituals & Geography:** Covers religious observances (e.g., Dashain, Tihar), local jatras, agricultural calendars, ritual objects, and region-specific events shaped by Nepal's terrain.
- **Idioms, Proverbs & Metaphors:** Focuses on culturally embedded expressions whose meanings cannot be derived from literal translation, requiring contextual interpretation.
- **Commonsense & Daily Life:** Captures implicit norms in daily interactions, including etiquette, hospitality, purity concepts, and socially appropriate household customs.
- **Gastronomy, Agriculture & Nature:** Encompasses culinary practices, ingredient pairings, agricultural planting cycles, common tools, and culturally significant flora and fauna.

This categorization creates a natural gradient of difficulty. Categories such as Kinship and Idioms are inherently more challenging due to their reliance on relational reasoning and non-literal meaning, whereas Festivals and Geography include a mix of factual and contextual knowledge. Representative examples from each category are provided in Appendix A.

Data Collection and Curation Pipeline The dataset was collaboratively developed by four native Nepali speakers. Each contributor led specific categories, generating over 250 questions to ensure both domain specialization and diverse coverage. Questions were sourced from a combination of public references (e.g., Wikipedia, books, grammar materials) alongside rich oral knowledge from community elders and lived cultural experience, ensuring authenticity across regional and ethnic contexts.

To ensure dataset quality, maintain consistency, and minimize annotation bias, we adopted a multi-stage pipeline:

- **Initial Authoring:** Native speakers generated culturally grounded MCQs within their assigned domains.
- **Cross-Review:** Question sets were exchanged for peer review. Reviewers labeled each item as *Keep*, *Modify*, or *Delete* based on cultural accuracy, ambiguity, and distractor plausibility.
- **Consensus Revision:** Items marked *Modify* were iteratively refined, and disputed cases were resolved through discussion until consensus was reached.
- **Translation and Alignment:** Initial English translations were generated using machine translation tools (e.g., Google Translate[3]) and subsequently refined manually by bilingual native speakers to ensure semantic fidelity. Romanization was also standardized.
- **Post-processing and Quality Control:** We conducted manual inspections to correct spelling and grammatical issues. A custom automated cleaning script detected duplicate entries, missing values, CSV formatting inconsistencies, and sequencing errors.
- **Final Cross-Check:** A final review ensured alignment across the Devanagari, Romanized, and English variants.

3.2. Experimental Setup

3.2.1. LLMs Used

We evaluate nine recent LLMs chosen to span a range of architectures, release types (proprietary vs. open-weight), and multilingual capabilities. Model selection was informed by contemporaneous standings on OpenRouter and other public leaderboards, with an emphasis on general instruction following and multilingual reasoning performance at the time of evaluation.[4]

The evaluated models are: GPT-4o-mini, GPT-5.2, and GPT-OSS-120B (OpenAI) (Hesham and Hamdy, 2024; Ma et al., 2026; Agarwal et al., 2025); Gemini 3 Flash Preview (Google DeepMind) (Zhang et al., 2026); DeepSeek V3.2 (Liu et al., 2025); Grok 4.1 Fast (xAI) (Ma et al., 2026); GLM-5 (Tsinghua University and Zhipu AI) (Zeng et al., 2026); Minimax M2.5 (Ko, 2026); and Trinity Large Preview (OpenRouter, n.d.). We include

[3] https://translate.google.com/
[4] https://openrouter.ai/rankings

Prompt Format (2-shot MCQ)

System/Instruction: You are a helpful assistant that answers multiple choice questions about Nepali common sense and cultural etiquette. Your task is to provide ONLY the letter (A, B, C, or D) corresponding to the correct option. Do NOT provide any reasoning, explanation, or additional text. Just the single letter.

Example 1:
Question: तिमी आफन्तको घरमा खाना खाँदै छौ। काँक्रोको अचार एकदम पिरो छ। चलन अनुसार तिमीले के गर्छौ?
A: मुखबाट 'आफ्, आफ्' आवाज निकाल्दै पानी माग्छौ।
B: थोरै भात वा चिउरासँग मुछेर चुपचाप निल्छौ।
C: अचारलाई थालको छेउमा थुक्छौ।
D: पाहुनालाई गाली गर्छौ।
Answer: B

Example 2:
Question: तिमी मन्दिर भित्र छौ। देउतालाई ढोग्दा तिम्रो टाउकोले केमा छुने कोसिस गर्छौ?
A: मन्दिरको सिँढीमा।
B: देउताको पाउ वा तोकिएको चरण-पादुकामा।
C: घण्टीमा।
D: पुजारीको हातमा।
Answer: B

Evaluation Instance Template:
Question: {question}
A: {A}
B: {B}
C: {C}
D: {D}
Answer:

Figure 1: Prompt Template

Trinity Large Preview in particular because it accounts for the highest usage share for Nepali on OpenRouter (see Appendix C), making it a highly relevant, community-preferred baseline for culturally grounded Nepali evaluation.

3.2.2. Evaluation Procedure

We use a standardized few-shot prompting setup (2-shot) for all models. Two fixed example question-answer pairs are prepended to every evaluation instance across all categories. Keeping the demonstrations and instructions constant ensures that performance differences primarily reflect the model's reasoning, rather than variation in prompt framing. The prompt template and demonstrations are shown Figure 1.

All models are evaluated on the same four-way MCQ task and are instructed to output *only* the option letter. We use a uniform system prompt and identical input formatting for all models to minimize differences due to verbosity, output style, or prompt interpretation, and to focus the evaluation on answer selection.

Inputs are evaluated in three representations: Devanagari Nepali, Romanized Nepali, and English translation (except for the two categories where English is intentionally omitted). All models are accessed through a unified API interface. Decoding settings are kept consistent across models by using a single set of parameters, with no model-specific tuning. This standardized configuration ensures a fair evaluation, isolating performance differences to the models' inherent capabilities rather than external optimization. Each question is submitted exactly once per model and representation, with no resampling or response averaging.

3.2.3. Analysis Protocol

We evaluate model performance using **accuracy**, measured as the proportion of questions for which a model selects the correct option. Beyond overall accuracy, we report results **by category** to better capture domain-specific strengths and weaknesses, since aggregate scores can mask substantial variation across cultural domains.

We further stratify performance across the three input representations: **Devanagari Nepali**, **Romanized Nepali**, and **English translation**. This breakdown enables a direct assessment of how sensitive models are to changes in script and language form. For culturally dense categories such as *Kinship & Social Hierarchy* and *Idioms, Proverbs & Metaphors*, we do not include English-based analysis because English translations were intentionally omitted to avoid loss of meaning and evaluation ambiguity.

Finally, we compare performance across models, categories, and representations to identify recurring error patterns and question types that are systematically challenging. Although our primary metric is accuracy, these structured slices provide a clearer view of model limitations and highlight where cultural knowledge and language representation most strongly influence performance.

4. Results

In this section, we summarize overall trends in accuracy to highlight where current LLMs succeed and where they systematically fail on Nepali cultural commonsense. A complete breakdown of per-model results, including accuracy, precision, recall, and F1 across categories and representations, is provided in the Appendix B.

Table 2: Cross-Model Comparative Accuracy across Categories (Bold indicates highest in each category)

Model	Agriculture	Cultural Etiquette	Festivals	Geography	Kinship & Social Hierarchy	Idioms/Proverbs	Overall Avg
Gemini 3 Flash	0.447	**0.935**	0.821	0.570	**0.821**	**0.866**	**0.743**
GPT-5.2	0.448	0.930	**0.865**	**0.689**	0.694	0.539	0.694
GLM-5	**0.512**	0.886	0.740	0.517	0.706	0.779	0.690
Grok-4.1 Fast	0.508	0.879	0.742	0.520	0.702	0.768	0.687
GPT-4o Mini	0.419	0.824	0.592	0.504	0.534	0.621	0.582
MiniMax-M2.5	0.349	0.826	0.635	0.465	0.532	0.579	0.564
DeepSeek-V3.2	0.331	0.836	0.686	0.385	0.551	0.524	0.552
GPT-OSS-120B	0.357	0.774	0.581	0.544	0.517	0.539	0.552
Trinity large	0.297	0.866	0.572	0.270	0.455	0.421	0.480

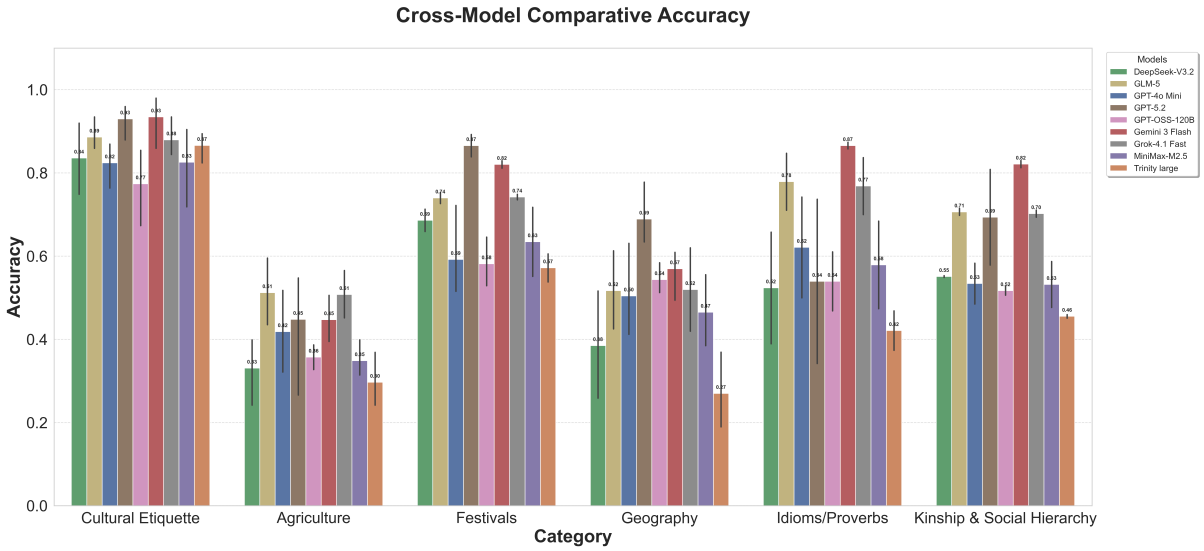


Figure 2: Comparative Accuracy across Categories

Comparative Accuracy across Categories. Figure 2 and Table 2 illustrate a significant imbalance in how models capture cultural knowledge across different domains. Models perform exceptionally well on widely documented domains. For instance, Cultural Etiquette is the best performing category, with both GPT-5.2 and Gemini 3 Flash achieving peak accuracies of 0.93. Festivals also demonstrate strong cross-model performance, with top models consistently scoring above 0.80.

Conversely, performance drops sharply on localized, community-specific domains like Agriculture and Geography. Agriculture is the most universally challenging category. Even the highest-performing models struggle in this area, with scores hovering between 0.30 and 0.51. This suggests a widespread deficit in implicit or rural localized knowledge. Furthermore, categories like Idioms/Proverbs and Kinship & Social Hierarchy are highly model-dependent. Gemini 3 Flash excels in Idioms/Proverbs (0.87) and Kinship (0.82), whereas models like Trinity large and DeepSeek-V3.2 struggle significantly in these exact same categories (scoring between 0.42 and 0.55). This variance indicates that specific training mixtures allow certain models to grasp complex linguistic and social nuances that others miss.

Performance Comparison: English vs. Nepali vs. Romanized. Figure 3 isolates the role of script and language form by evaluating each question across three formats. While aggregate trends favor native script, individual models exhibit fascinating divergence in their linguistic handling. Models such as Gemini 3 Flash and GLM-5 perform best in native Devanagari (0.83 and 0.78 respectively), outperforming their own English baselines. However, this is not a universal rule. GPT-5.2 demonstrates an inverted proficiency, performing best in English (0.78), followed by Romanized Nepali (0.75), and performing worst in native Devanagari (0.66). A similar trend is visible with GPT-4o Mini. This highlights that multilingual alignment varies heavily by model architecture, and language representation choices materially affect how reliably cultural commonsense is retrieved.

Overall Language Performance Across All Models. Figure 4a consolidates the linguistic results across all models to confirm a stable over-

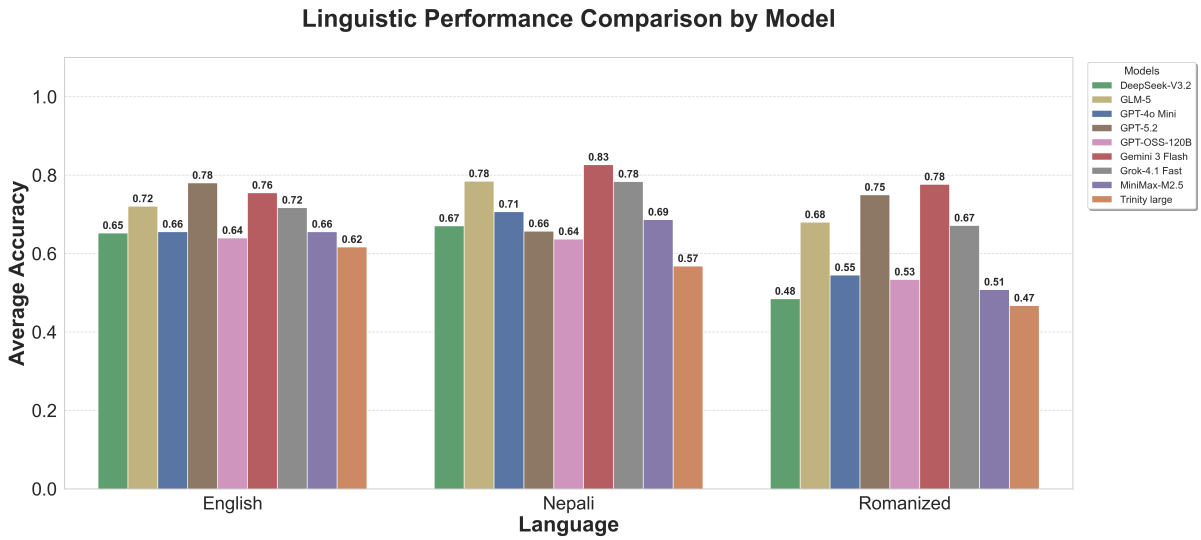
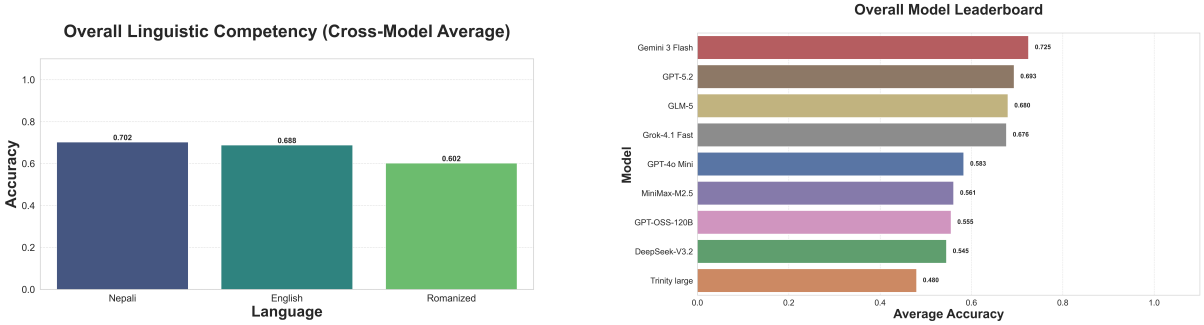


Figure 3: Performance Comparison: English vs. Nepali vs. Romanized



(a) Overall Language Performance Across All Models

(b) Overall Model Ranking (Mean Accuracy Across All Categories)

Figure 4: Aggregate results across input representations and models.

all ordering. Native Devanagari Nepali achieves the highest aggregate accuracy (0.702), followed closely by English (0.688). Romanized Nepali represents a significant step down (0.602). This aggregate pattern suggests that non-standard transliteration meaningfully hinders an LLM's access to its culturally grounded knowledge base.

Overall Model Ranking. Finally, Figure 4b presents the overall leaderboard, revealing a stratification in model capabilities. The top tier of models demonstrates a clear advantage, with Gemini 3 Flash leading the benchmark at an average accuracy of 0.725. This is followed closely by GPT-5.2 (0.693), GLM-5 (0.680), and Grok-4.1 Fast (0.676). There is a notable drop-off to the second tier of models, beginning with GPT-4o Mini (0.583) and cascading down to Trinity large (0.480). This distribution indicates that larger parameter counts or general popularity do not automatically guarantee superior cultural reasoning, as highly optimized models often outperform larger alternatives.

5. Discussion

The NeCCo evaluation offers a detailed view of how well current LLMs capture Nepali cultural intelligence. While the top-performing systems show strong general capabilities, we observe recurring failure modes that expose the limits of today's models in culturally grounded commonsense reasoning. In many cases, models handle surface-level facts and linguistic form reliably but struggle when correct answers depend on implicit norms, context-specific practices, or culturally encoded meanings that native speakers take for granted.

5.1. The Importance of Cultural Benchmarking

Most standard NLP benchmarks emphasize linguistic competence, general knowledge, or skills that transfer well across languages. They rarely test the everyday commonsense and social expectations that are obvious within a community but

opaque to outsiders. NeCCo illustrates why culturally targeted benchmarks are necessary: models can appear highly capable on global tasks while still failing in realistic local scenarios where appropriate behavior, social roles, and cultural interpretations matter.

This has practical implications for deployment. A model may produce fluent Nepali and handle translation accurately, yet remain unreliable when advising users in culturally sensitive contexts such as kinship address terms, social hierarchy, ritual practices, or community norms. Evaluating these dimensions helps prevent a form of cultural erasure in which systems are treated as "Nepali-capable" based primarily on language form, even though they lack key cultural competencies required for safe and useful real-world interaction.

Our results also underscore that cultural understanding does not automatically transfer through translation. The performance gap between English and Devanagari Nepali suggests that knowing a concept in English is not the same as recognizing its lived meaning, associated norms, and situational usage in Nepali contexts. Cultural competence therefore requires more than mapping words across languages. It requires interpreting meaning within local practices, expectations, and history.

5.2. Key Findings and Observations

Linguistic Representation and Script Sensitivity A consistent pattern in our findings is the clear advantage of native-script inputs. For several top-tier models, Devanagari Nepali performs as well as or better than English. This indicates that culturally grounded associations are often more accessible in the original language form, as native script provides stable token representations that map directly to culturally specific semantic spaces in the model's pre-training data.

In contrast, Romanized Nepali severely underperforms across models, showing a substantial drop in accuracy. A primary mechanical reason for this is high token fertility: standard LLM tokenizers often fail to recognize Romanized Indic words, breaking them down into excessively small, semantically disjointed character fragments. Furthermore, informal Romanization introduces significant orthographic noise and spelling variations in real-world use. This fragmentation and inconsistency make the model's conceptual retrieval highly brittle, weakening its ability to connect transliterated forms to stable cultural concepts.

We also observe a clear gap in figurative and culturally embedded language. Performance is notably low on Proverbs/Idioms and other idiomatic content, where meaning is rarely recoverable from literal interpretation. Correct answers often require familiarity with social context, implied intent, irony, and shared cultural references. These results suggest that current LLMs still struggle when the task requires inferring non-literal meaning that is culturally conventional rather than textually explicit.

Model Scaling and Local Knowledge Gaps Another important takeaway is that model scale alone does not guarantee better cultural reasoning. In our experiments, optimized proprietary models often outperform substantially larger open-weight models. This suggests that data quality, local coverage, and the presence of culturally representative training signals matter far more than raw parameter counts for low-resource cultural competence.

Finally, the weakest categories point to a severe local knowledge void. Low scores in areas such as agriculture, geography, and kinship relations indicate that many models lack grounded knowledge about Nepali seasonal routines, tools, local topography, and family relationship systems. These domains are often under-represented in widely used training corpora, which tend to over-index on globally dominant and frequently documented contexts. NeCCo makes these gaps visible and provides a concrete target for improving cultural coverage in future model training and evaluation.

5.3. Future Implications

NeCCo points toward a concrete direction for the next generation of Nepali-centered AI. Our findings suggest that improving cultural understanding is less about scaling model size and more about deliberately incorporating local knowledge and sustained feedback from native speakers throughout data curation, training, and evaluation. Future work should systematically investigate the effects of parameter tuning, prompt design, and targeted fine-tuning on Nepali cultural resources, including oral narratives, proverbs, and ethnographic texts. This domain-adaptive approach may help mitigate the severe performance gaps observed in categories like agriculture and kinship. Increasing parameter counts or relying strictly on translation pathways cannot compensate for missing culturally grounded signals.

Without such intentional effort, the digital divide risks expanding beyond access into a divide in representation and understanding. Communities may be able to use AI systems yet still be misunderstood when those systems interpret their language, traditions, and everyday practices through an incomplete cultural lens. Progress in AI for low-resource languages therefore requires inclusion: deeper cultural coverage, community-informed development, and evaluation methods that measure what truly matters in real use.

6. Conclusion

We introduced NeCCo, a culturally grounded benchmark for evaluating Nepali commonsense reasoning across domains shaped by social structure, language, and lived practice. Our evaluation demonstrates that strong performance on global benchmarks does not reliably translate to culturally grounded competence. Models consistently struggle with implicit and relational knowledge, particularly in domains like kinship, proverbs, and localized agriculture. Furthermore, results vary substantially based on linguistic representation. Native Devanagari generally outperforms English, while Romanized Nepali consistently degrades model performance. Overall, NeCCo highlights that advancing cultural reasoning depends heavily on high-quality, culturally representative data and targeted evaluation rather than simply increasing model scale. This benchmark provides a critical foundation for more inclusive and accurate assessments of LLMs in low-resource settings.

7. Limitations

While NeCCo introduces a novel framework for cultural evaluation, it has several limitations. First, the dataset is relatively small; despite careful curation, a larger corpus would support more granular statistical analysis and better capture Nepal's diverse subcultures. Second, the use of fixed multiple-choice questions enables standardized evaluation but constrains the assessment of open-ended cultural reasoning and may oversimplify complex lived knowledge. The benchmark also focuses on a limited set of foundational domains. Future work should expand into areas such as indigenous folklore, regional histories, and contemporary digital culture for a more comprehensive evaluation. Additionally, annotations were conducted solely by the authors, who are native Nepali speakers. While this ensures linguistic fluency, it introduces potential subjectivity and lacks input from domain experts such as sociologists or anthropologists. The absence of inter-annotator agreement metrics (e.g., Cohen's Kappa or Krippendorff's Alpha) further limits the ability to assess annotation reliability.

Furthermore, our evaluation relies on standardized default inference settings and a fixed 2-shot prompt. Model performance on culturally grounded reasoning can be sensitive to temperature adjustments or advanced prompt engineering, such as providing dynamic, domain-specific cultural demonstrations. Finally, the benchmark does not randomize answer option order across runs, leaving room for positional bias in model responses. Prior work suggests that LLMs may favor certain option positions independent of content, introducing evaluation noise (Pezeshkpour and Hruschka, 2024). Future iterations will incorporate option randomization and repeated evaluations to better reflect true reasoning performance.

8. Ethical Consideration

NeCCo is constructed using culturally grounded knowledge sourced from native Nepali speakers, which inherently introduces challenges related to representation, bias, and coverage. While we employed a multi-stage validation process to ensure clarity and cultural accuracy, the dataset may still reflect localized perspectives and might not fully capture the vast diversity of Nepal's socio-cultural landscape. Additionally, reducing culturally specific practices to fixed-answer formats risks flattening context-dependent knowledge. This benchmark is intended strictly for evaluation and research purposes; it should not be interpreted as a complete, definitive, or monolithic representation of Nepali culture. We strongly encourage responsible use, transparency, and future expansion through broader community involvement to continuously improve the benchmark's inclusivity and cultural fidelity.

9. Data/Code Availability Statement

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

10. References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Xiaoxuan Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott

McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, et al. 2025. gpt-oss-120b and gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Samuel Cahyawijaya. 2024. *Llm for everyone: Representing the underrepresented in large language models*. Hong Kong University of Science and Technology (Hong Kong).

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. Culturalbench: A robust, diverse, and challenging cultural benchmark by human-ai culturalteaming.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The flores evaluation datasets for low-resource machine translation: Nepali–english and sinhala–english. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 6098–6111.

Alaa Hesham and Abeer Hamdy. 2024. Fine-tuning GPT-4o-Mini for programming questions generation. In *Proceedings of the 2024 International Conference on Computer and Applications (ICCA)*, pages 1–6.

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.

Neha Joshi, Pamir Gogoi, AasimBaig Mirza, Aayush Jansari, Aditya Yadavalli, Ayushi Pandey, Arunima Shukla, Deepthi Sudharsan, Kalika Bali, and Vivek Seshadri. 2025. Elr-1000: A community-generated dataset for endangered indic indigenous languages. 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 2441–2457.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K Jain, Virginia Aglietti, Disha Jindal, Yuanzhu Peter Chen, et al. 2025. Big-bench extra hard. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26473–26501.

Supriya Khadka and Bijayan Bhattarai. 2025. Gender bias in nepali-english machine translation: A comparison of llms and existing mt systems. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 75–82.

Supriya Khadka, Ranju G.C., Prabin Paudel, Rahul Shah, and Basanta Joshi. 2023. Nepali text-to-speech synthesis using tacotron2 for mel-spectrogram generation. In *SIGUL 2023, 2nd Annual Meeting of the Special Interest Group on Under-resourced Languages: a Satellite Workshop of Interspeech 2023*.

Ju-Chun Ko. 2026. Bilingual bias in large language models: A taiwan sovereignty benchmark study. *arXiv preprint arXiv:2602.06371*.

Britt Krause. 1980. Kinship, hierarchy and equality in north western nepal. *Contributions to Indian Sociology*, 14(2):169–194.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Xingjun Ma, Yixu Wang, Hengyuan Xu, Yutao Wu, Yifan Ding, Yunhan Zhao, Zilong Wang, Jiabin

Hua, Ming Wen, Jianan Liu, Ranjie Duan, Yifeng Gao, Yingshui Tan, Yunhao Chen, Hui Xue, Xin Wang, Wei Cheng, Jingjing Chen, Zuxuan Wu, Bo Li, and Yu-Gang Jiang. 2026. A safety report on GPT-5.2, Gemini 3 Pro, Qwen3-VL, Grok 4.1 Fast, Nano Banana Pro, and Seedream 4.5. *arXiv preprint arXiv:2601.10527*.

Yash Madhani, Sushane Parthan, Priyanka Bedekar, Gokul Nc, Ruchi Khapra, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. 2023. Aksharantar: Open indic-language transliteration datasets and models for the next billion users. In *Findings of the association for computational linguistics: Emnlp 2023*, pages 40–57.

Arijit Maji, Raghvendra Kumar, Akash Ghosh, Anushka Anushka, and Sriparna Saha. 2025. Sanskriti: A comprehensive benchmark for evaluating language models' knowledge of indian culture. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4434–4451.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 5356–5371.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1953–1967.

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.

OpenRouter. n.d. Llm rankings. https://openrouter.ai/rankings. Accessed: 2026-02-21.

Shubham Parashar, Blake Olson, Sambhav Khurana, Eric Li, Hongyi Ling, James Caverlee, and Shuiwang Ji. 2025. Inference-time computations for llm reasoning and planning: A benchmark and insights. *arXiv preprint arXiv:2502.12521*.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105.

Uwe Peters, Alexander Krauss, and Oliver Braganza. 2022. Generalization bias in science. *Cognitive science*, 46(9):e13188.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017.

Mengyao Piao, Lingqi Miao, Yilun Liu, Minggui He, Hongxia Ma, Li Zhang, Daimeng Wei, and Shimin Tao. Beyond the west: A survey of cultural datasets for culturally-grounded llms.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.

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

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

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 4463–4473.

Tej Bahadur Shahi and Chiranjibi Sitaula. 2022. Natural language processing for nepali text: A review. *Artificial Intelligence Review*, 55(4):3401–3429.

Oyesh Mann Singh, Ankur Padia, and Anupam Joshi. 2019. Named entity recognition for nepali

language. In *2019 IEEE 5th international conference on collaboration and internet computing (cic)*, pages 184–190. IEEE.

Madhavendra Thakur. 2025. Culturally-grounded chain-of-thought (cg-cot): Enhancing llm performance on culturally-specific tasks in low-resource languages. *arXiv preprint arXiv:2506.01190*.

Surendrabikram Thapa, Kritesh Rauniyar, Hariram Veeramani, Surabhi Adhikari, Imran Razzak, and Usman Naseem. 2025. Probing the limits of multilingual language understanding: Low-resource language proverbs as llm benchmark for ai wisdom. In *Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025)*, pages 120–129.

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.

Armin Toroghi, Ali Pesaranghader, Tanmana Sadhu, and Scott Sanner. 2025. Llm-based typed hyperresolution for commonsense reasoning with knowledge bases. In *The Thirteenth International Conference on Learning Representations*.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. Milu: A multi-task indic language understanding benchmark. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10076–10132.

Ieva Zake. 2023. Holistic bias in sociology: Contemporary trends. In *The Palgrave Handbook of Methodological Individualism: Volume II*, pages 403–421. Springer.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chengxing Xie, Cunxiang Wang, et al. 2026. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*.

P. Zhang, J. Wang, X. Hu, X. Wang, X. Fan, W. Chi, and W. Yang. 2026. Comparative performance of GPT-4, GPT-o3, GPT-5, Gemini-3-Flash, and DeepSeek-R1 in ophthalmology question answering. *Frontiers in Cell and Developmental Biology*, 14:1744389.

11. Appendix

Appendix A: Representative Examples from NeCCo

This appendix presents representative NeCCo examples in Devanagari Nepali, Romanized Nepali, and English (Table 3). For categories without official English translations (Kinship & Social Hierarchy; Idioms, Proverbs & Metaphors), we provide an English gloss for readability; this is not part of the benchmark.

Appendix B: Complete Results Across All Categories

This appendix reports per-model results across all NeCCo categories and input types. For *Proverbs/Idioms* and *Kinship & Social Hierarchy*, English translations are omitted to preserve meaning; corresponding columns are marked 0 (not evaluated). Across all tables (Table 4 – Table 9), **A** = accuracy, **P** = precision, **R** = recall, and **F1** = F1 score.

Appendix C: OpenRouter Usage Snapshot for Nepali

Figure 5 presents a snapshot of model usage statistics for the Nepali language on the OpenRouter platform, captured over a 30-day period ending in late March 2026. The data highlights Trinity Large Preview as the most frequently utilized model for Nepali prompts during this timeframe, accounting for 46.5% of the total volume.

Figure 5: Model usage share and token volume for the Nepali language on OpenRouter.

Category	Variant	Question	Options (A, B, C, D)	Ans.
Kinship & Social Hierarchy	Devanagari	बुवा को भिनाजु के हुन्छ?	A: आमाजू B: सालो C: छोरा D: फुपाजु	D
	Romanized	buwa ko bhinaju ke hunchha?	A: aamaju B: salo C: chhora D: phupaju	D
	English (gloss)	What do we call “father’s sister’s husband”?	A: aunt B: brother-in-law (wife’s brother) C: son D: paternal aunt’s husband	D
Festivals, Rituals & Geography	Devanagari	इन्द्रजात्रा मुख्यतः कहाँ मनाइन्छ?	A: पोखरा B: काठमाडौं C: भक्तपुर D: पाटन	B
	Romanized	Indrajatra mukhyata kaha manaincha?	A: Pokhara B: Kathmandu C: Bhaktapur D: Patan	B
	English	Where is Indra Jatra primarily celebrated?	A: Pokhara B: Kathmandu C: Bhaktapur D: Patan	B
Gastronomy, Agriculture & Nature	Devanagari	मास को दाल मा के झानिन्छ?	A: कागती B: मुला C: प्याज D: जिम्बु	D
	Romanized	maas ko daal ma ke jhanincha?	A: kagati B: mula C: pyaj D: jimbu	D
	English	What is typically tempered in black gram (maas) lentil soup?	A: lemon B: radish C: onion D: jimbu	D
Idioms, Proverbs & Metaphors	Devanagari	लुटको धन फुपुको श्राद्ध	A: पुण्य कमाउनु B: अरूको कमाइ खर्च गर्नु C: श्राद्ध गर्नु D: कमाउनु	B
	Romanized	lutko dhan phupuko shraddha	A: punya kamaunu B: aruko kamai kharcha garnu C: shraddha garnu D: kamaunu	B
	English (gloss)	Idiom meaning closest to:	A: earn merit B: spend someone else’s earnings C: perform a death ritual D: earn	B

Table 3: Representative NeCCo examples shown in parallel variants. English lines for Kinship and Idioms are explanatory glosses and are not part of the benchmark release.

Dataset	English				Devanagari				Romanized			
Model	A	P	R	F1	A	P	R	F1	A	P	R	F1
GPT-4o Mini	0.462	0.494	0.417	0.406	0.521	0.564	0.521	0.631	0.321	0.388	0.321	0.410
Trinity Large	0.280	0.282	0.280	0.320	0.369	0.435	0.39	0.422	242	0.287	0.242	0.215
DeepSeek-V3.2	0.351	0.370	0.351	0.466	0.399	0.422	0.399	0.000	0.242	0.276	0.242	0.000
Gemini 3 Flash	0.509	0.525	0.509	0.512	0.441	0.448	0.441	0.000	0.395	0.426	0.395	0.000
MiniMax-M2.5	0.333	0.365	0.333	0.000	0.399	0.412	0.366	0.000	0.322	0.370	0.322	0.000
GPT-5.2	0.548	0.568	0.548	0.000	0.530	0.563	0.530	0.000	0.266	0.327	0.266	0.202
GPT-OSS-120B	0.331	0.356	0.331	0.000	0.387	0.410	0.387	0.000	0.364	0.392	0.364	0.000
Grok-4.1 Fast	0.506	0.508	0.506	0.000	0.571	0.592	0.571	0.534	0.452	0.474	0.452	0.456
GLM-5	0.5952	0.614	0.595	0.000	0.595	0.614	0.595	0.000	0.436	0.488	0.436	0.000

Table 4: Performance comparison on the **Agriculture** category.

Dataset	English				Devanagari				Romanized			
Model	A	P	R	F1	A	P	R	F1	A	P	R	F1
GPT-4o Mini	0.840	0.842	0.840	0.824	0.870	0.920	0.870	0.887	0.764	0.871	0.764	0.803
Trinity Large	0.879	0.882	0.879	0.854	0.905	0.905	0.905	0.904	0.824	0.882	0.824	0.851
DeepSeek-V3.2	0.839	0.831	0.839	0.825	0.920	0.952	0.920	0.932	0.749	0.860	0.749	0.791
Gemini 3 Flash	0.864	0.868	0.864	0.847	0.980	0.980	0.980	0.980	0.968	0.976	0.970	0.972
MiniMax-M2.5	0.854	0.865	0.854	0.841	0.905	0.923	0.905	0.911	0.719	0.873	0.719	0.769
GPT-5.2	0.879	0.873	0.879	0.857	0.950	0.958	0.950	0.952	0.960	0.969	0.960	0.963
GPT-OSS-120B	0.819	0.836	0.819	0.816	0.867	0.938	0.867	0.893	0.677	0.835	0.677	0.737
Grok-4.1 Fast	0.844	0.854	0.844	0.823	0.935	0.949	0.935	0.940	0.859	0.916	0.859	0.882
GLM-5	0.859	0.845	.859	0.837	0.935	0.946	0.935	0.940	0.864	0.902	0.864	0.879

Table 5: Performance comparison on the **Cultural Etiquette** category.

Dataset	English				Devanagari				Romanized			
Model	A	P	R	F1	A	P	R	F1	A	P	R	F1
GPT-4o Mini					0.586	0.628	0.586	0.582	0.491	0.530	0.431	0.476
Trinity Large	0.000	0.000	0.000	0.000	0.453	0.485	0.453	0.423	0.464	0.494	0.464	0.4449
DeepSeek-V3.2	0.000	0.000	0.000	0.000	0.556	0.561	0.556	0.557	0.551	0.552	0.551	0.551
Gemini 3 Flash	0.000	0.000	0.000	0.000	0.827	0.831	0.827	0.828	0.833	0.837	0.833	0.834
MiniMax-M2.5	0.000	0.000	0.000	0.000	0.590	0.596	0.590	0.592	0.483	0.423	0.483	0.483
GPT-5.2	0.000	0.000	0.000	0.000	0.586	0.639	0.586	0.595	0.812	0.812	0.812	0.812
GPT-OSS-120B	0.000	0.000	0.000	0.000	0.532	0.542	0.532	0.535	0.513	0.516	0.513	0.513
Grok-4.1 Fast	0.000	0.000	0.000	0.000	0.714	0.726	0.714	0.715	0.697	0.710	0.697	0.697
GLM-5	0.000	0.000	0.000	0.000	0.701	0.707	0.701	0.701	0.718	0.722	0.718	0.718

Table 6: Performance comparison on the **Kinship & Social Hierarchy** category.

Dataset	English				Devanagari				Romanized			
Model	A	P	R	F1	A	P	R	F1	A	P	R	F1
GPT-4o Mini	0.000	0.000	0.000	0.000	0.742	0.786	0.742	0.759	0.500	0.619	0.500	0.533
Trinity Large	0.000	0.000	0.000	0.000	0.468	0.717	0.468	0.460	0.374	0.714	0.374	0.363
DeepSeek-V3.2	0.000	0.000	0.000	0.000	0.658	0.798	0.658	0.704	0.390	0.608	0.390	0.434
Gemini 3 Flash	0.000	0.000	0.000	0.000	0.878	0.903	0.878	0.888	0.862	0.877	0.862	0.869
MiniMax-M2.5	0.000	0.000	0.000	0.000	0.684	0.778	0.684	0.709	0.474	0.671	0.474	0.529
GPT-5.2	0.000	0.000	0.000	0.000	0.342	0.563	0.342	0.407	0.768	0.799	0.737	0.752
GPT-OSS-120B	0.000	0.000	0.000	0.000	0.614	0.715	0.614	0.647	0.471	0.656	0.471	0.517
Grok-4.1 Fast	0.000	0.000	0.000	0.000	0.837	0.882	0.837	0.851	0.700	0.796	0.700	0.739
GLM-5	0.000	0.000	0.000	0.000	0.847	0.885	0.847	0.860	0.711	0.817	0.711	0.755

Table 7: Performance comparison on the **Proverbs/Idioms** category.

Dataset	English				Devanagari				Romanized			
Model	A	P	R	F1	A	P	R	F1	A	P	R	F1
GPT-4o Mini	0.516	0.480	0.516	0.474	0.722	0.722	0.722	0.720	0.538	0.51	0.538	0.503
Trinity Large	0.538	0.581	0.538	0.503	0.605	0.600	0.605	0.571	0.000	0.000	0.000	0.000
DeepSeek-V3.2	0.659	0.659	0.659	0.656	0.713	0.708	0.713	0.707	0.000	0.000	0.000	0.000
Gemini 3 Flash	0.842	0.842	0.842	0.841	0.837	0.838	0.837	0.836	0.000	0.000	0.000	0.000
MiniMax-M2.5	0.552	0.540	0.552	0.531	0.718	0.714	0.718	0.708	0.000	0.000	0.000	0.000
GPT-5.2	0.839	0.839	0.839	0.838	0.892	0.893	0.832	0.891	0.000	0.000	0.000	0.000
GPT-OSS-120B	0.572	0.550	0.572	0.547	0.649	0.623	0.649	0.619	0.534	0.505	0.534	0.000
Grok-4.1 Fast	0.7400	0.719	0.7400	0.710	0.749	0.738	0.749	0.000	0.000	0.000	0.000	0.000
GLM-5	0.753	0.745	0.753	0.000	0.753	0.745	0.753	0.000	0.000	0.000	0.000	0.000

Table 8: Performance comparison on the **Festivals** category.

Dataset	English				Devanagari				Romanized			
Model	A	P	R	F1	A	P	R	F1	A	P	R	F1
GPT-4o Mini	0.470	0.475	0.470	0.000	0.633	0.628	0.633	0.627	0.412	0.422	0.412	0.413
Trinity Large	0.251	0.521	0.261	0.228	0.369	0.588	0.369	0.362	0.190	0.455	0.180	0.154
DeepSeek-V3.2	0.378	0.470	0.380	0.401	0.516	0.599	0.516	0.536	0.259	0.447	0.259	0.282
Gemini 3 Flash	0.612	0.632	0.612	0.617	0.610	0.614	0.608	0.610	0.497	0.496	0.497	0.494
MiniMax-M2.5	0.456	0.497	0.455	0.470	0.556	0.608	0.556	0.572	0.387	0.484	0.387	0.413
GPT-5.2	0.634	0.671	0.634	0.646	0.778	0.784	0.778	0.766	0.655	0.636	0.655	0.613
GPT-OSS-120B	0.514	0.550	0.514	0.526	0.586	0.640	0.586	0.602	0.540	0.575	0.540	0.551
Grok-4.1 Fast	0.520	0.546	0.520	0.527	0.620	0.635	0.620	0.625	0.420	0.472	0.420	0.438
GLM-5	0.617	0.636	0.617	0.623	0.617	0.636	0.617	0.623	0.425	0471	0.425	0.440

Table 9: Performance comparison on the **Geography** category.