MonCulture-Eval: A Hierarchical Benchmark for Evaluating Mongolian Cultural Capabilities of Large Language Models across Scripts and Regions

Quulgan Minggad[♣] **Xiao Zinan**[♣] **Yuan Sun**[♣♡†]
[♣]Minzu University of China, Beijing, China
[♡]National Language Resource Monitoring & Research Center Minority Languages Branch
Institute of National Security, Minzu University of China
sunyuan@muc.edu.cn

[†]Corresponding author.

Abstract

While Large Language Models (LLMs) have achieved impressive linguistic fluency in low-resource languages, their capacity to process deep cultural nuances remains insufficiently quantified. This paper introduces MonCulture-Eval, a benchmark designed to assess the cultural intelligence of LLMs in the Mongolian context across two writing systems (Traditional and Cyrillic) and three regional subcultures (Alxa, Ordos, and Horqin). Curated entirely from primary, non-digitized archives to inherently prevent data contamination, the benchmark employs a three-layer cognitive hierarchy—Factual, Situational, and Values—supplemented by specialized tasks including Riddles, Taboos, and Proverbs. Evaluation of frontier models, including GPT-5.2, Gemini-3-Pro-Preview, Claude-Sonnet-4-5, DeepSeek-v3.2, and the regionally optimized Qwen3-Max, reveals distinct structural limitations. First, we observe a severe "Script Gap," where most models experience a sharp performance decline in the Traditional script, effectively restricting their access to historical cultural archives. Second, qualitative analysis identifies a systematic "Tourist Perspective" (Etic Bias), wherein models sanitize spiritual rituals into secular functional norms. While Gemini-3-Pro-preview maintains robust cross-script alignment and Emic consistency, the broader results demonstrate that linguistic translation capability does not guarantee cultural value alignment. These findings provide empirical baselines for advancing culturally grounded AI systems. Our data and code are publicly available at https://github.com/Ayakades/MonCulture-Eval.

1 Introduction

While Large Language Models (LLMs) excel at multilingual translation and instruction-following in low-resource settings like Tibetan (Zhuang et al., 2025a), they frequently suffer from severe cultural misalignment. Current models systematically default to Western-centric priors, revealing a critical gap where linguistic fluency masks an inability to process underlying social norms and value systems (Masoud et al., 2024; Sorensen et al., 2024; Xu et al., 2025). This disconnect is especially acute in low-resource, non-Western languages such as Mongolian, whose cultural paradigms are highly contextual and deeply rooted in nomadic traditions (Hall, 1976).

The rigorous evaluation of Mongolian cultural capabilities presents two distinct challenges. First, the language operates under a complex digraphia system: it utilizes both the Traditional Mongolian script (vertical) and the Cyrillic script. Second, Mongolian culture consists of diverse regional variations (e.g., Alxa, Ordos, Horqin), each maintaining specific customary logic that generic pre-training data often filter out (Hershcovich et al., 2022). While global benchmarks assess declarative knowledge, they fail to operationalize "Deep Culture"—the dynamic intersection of shared knowledge, implicit values, and social practice. Existing frameworks typically evaluate factual retrieval (e.g., dates, figures) rather than the procedural reasoning required to adapt to indigenous taboos or metaphorical constructs.

To address this structural gap, we propose **MonCulture-Eval**, a hierarchical benchmark specifically engineered to quantify the cultural alignment of LLMs. Moving away from translation-based dataset construction, MonCulture-Eval is developed through a strict "Indigenous-First" protocol. Native domain experts curated the dataset directly from primary, non-digitized archives, establishing a natural firewall against pre-training data contamination.

Drawing upon the seminal cultural levels theory of Schein (1985),The benchmark implements a Three-Layer Cognitive Framework, structuring evaluation from Factual Knowledge (Layer 1) to Situational Appropriateness (Layer 2), and concluding with Value Alignment (Layer 3). This pro-

Findings of the Association for Computational Linguistics: ACL 2026, pages 28997–29014

gression isolates models capable of surface-level memorization from those exhibiting true ontological alignment with the target culture's spiritual and social logic. Furthermore, the dataset incorporates culturally specific tasks—Riddles, Taboos, Proverbs, and Benedictions—to test metaphorical reasoning and normative correction.

We evaluated a suite of frontier models, expanding our analysis to include global systems (GPT-5.2, Gemini-3-Pro, Claude-Sonnet-4-5, DeepSeek-v3.2) and the regionally optimized Qwen3-Max. The empirical results expose a critical "Script Gap," indicating that most models process the Traditional script primarily through translation shortcuts rather than embedded cultural understanding. Additionally, the evaluations reveal a pervasive "Tourist Perspective," where models systematically default to functional, secular explanations (Etic Bias) rather than retrieving the authentic internal rationale (Emic perspective) of a cultural practice.

In summary, this paper presents the following contributions:

- **Native-Centric Curation with Contamination Insulation:** We release MonCulture-Eval, an expert-curated cultural benchmark covering dual scripts and three regional subcultures, built entirely from primary sources to prevent web-crawled data contamination.
- **Hierarchical Probing Framework:** We propose a three-layer cognitive evaluation architecture that systematically differentiates factual retrieval from deep value alignment and situational pragmatics.
- **Pluralistic Evaluation Findings:** We provide robust empirical evidence, validated via cross-judge verification, detailing the systematic "Script Gap" and "Etic Bias" present in current LLMs, including both global and regionally optimized architectures.

2 Background: Regional Archetypes in Mongolian Culture

To operationalize cultural intelligence in the Mongolian context, it is necessary to establish clear regional archetypes. Generic pre-training corpora often homogenize marginalized cultures, stripping away the localized norms that dictate daily social interactions. MonCulture-Eval constructs its tasks around three distinct regional cultures, each representing a specific ecological and historical trajectory:

- Alxa Culture: Characterized by an arid desert-steppe environment, preserving ancient nomadic survival strategies and specific taboos regarding livestock and water conservation.
- Ordos Culture: Historically serving as the guardian of Genghis Khan's mausoleum, this region maintains highly formalized rituals and central-court sacrificial protocols.
- Horqin Culture: Situated in an agricultural-pastoral ecotone, integrating nomadic traditions with agrarian social conventions.

Delineating these archetypes prevents the benchmark from treating Mongolian culture as a monolith, enabling a granular evaluation of an LLM's capacity to adapt to specific situational constraints.

3 Related Work

3.1 Multilingual LLMs and Pluralistic Alignment

Initial language model evaluation focused heavily on declarative knowledge retrieval through benchmarks like MMLU (Hendrycks et al., 2021) and C-Eval (Huang et al., 2023). While subsequent frameworks expanded to multilingual capacities (Blasi et al., 2022; Koto et al., 2023; Kurihara et al., 2022), linguistic proficiency frequently fails to equate to cultural alignment. Recent scholarship emphasizes pluralistic alignment, demonstrating that models trained primarily on Western corpora default to Western ontological norms and struggle to evaluate non-Western societies accurately (Masoud et al., 2024; Sorensen et al., 2024; Xu et al., 2025). Our benchmark extends this trajectory by shifting the evaluation target from surface-level facts to implicit cultural values.

3.2 The Script Gap in Low-Resource Settings

Evaluating traditional, non-Latin scripts presents unique computational barriers. Structural disparities in tokenization heavily penalize morphologically rich and low-resource languages (Rust et al., 2021). While models may effectively process the Cyrillic alphabet due to parameter transfer from Russian corpora, the vertical Traditional Mongolian script is frequently subjected to fragmentation and mistranslation. This script gap effectively limits a model's access to historical and indigenous knowledge bases.

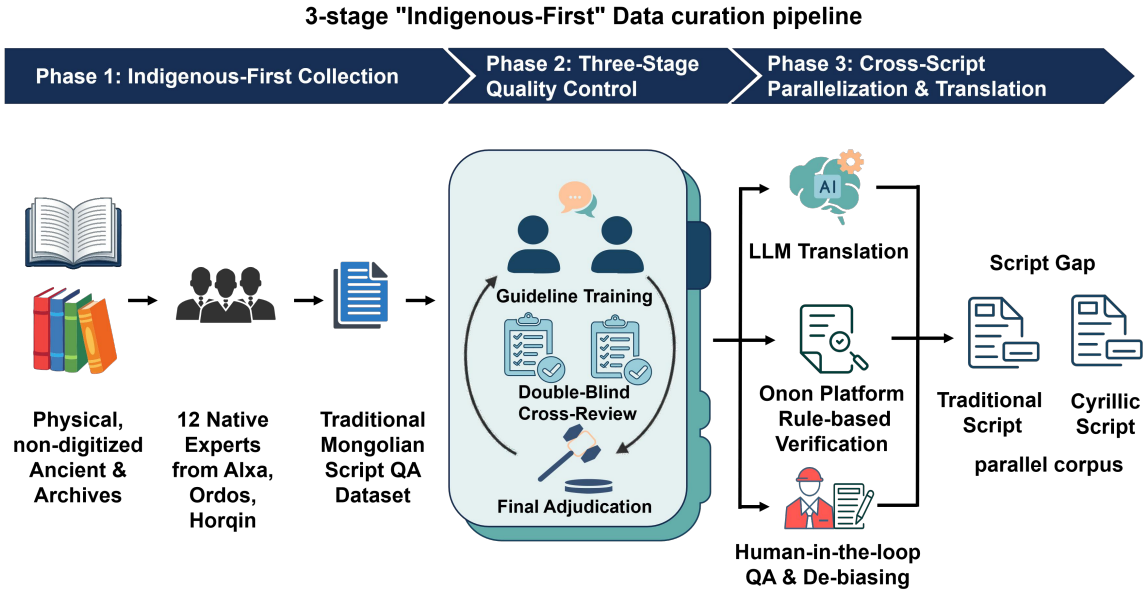

Figure 1: The Indigenous-First curation pipeline. The three-stage process incorporates primary archival sourcing, double-blind cross-review, and human-in-the-loop QA via the Onon platform to ensure strict cross-script semantic equivalence and data contamination insulation.

3.3 Quantifying Cultural Norms

Previous efforts to quantify social norms, such as Social Chemistry (Forbes et al., 2020), largely reflect Western demographics. While recent cultural datasets attempt to address this, they typically rely on automated generation, translation, or web scraping, inadvertently introducing pre-training data contamination. In contrast, MonCulture-Eval adopts a strict indigenous-first curation protocol, drawing directly from primary offline archives to ensure authentic representation and prevent data leakage.

4 Dataset Construction

4.1 The Indigenous-First Curation Pipeline

Recent studies emphasize that web-scraped benchmarks are highly susceptible to pre-training data contamination and representational biases, artificially inflating perceived model capabilities (Bender et al., 2021). To guarantee the benchmark's integrity, MonCulture-Eval is constructed entirely from primary, non-digitized archives. This methodological choice creates a natural firewall against data leakage. We implemented a rigorous three-stage curation pipeline (illustrated in Figure 1) to ensure high annotation transparency and cross-judge consistency.

The pipeline operates in three phases. Phase 1 involves native experts sourcing material directly from physical, non-digitized archives to prevent pre-training corpora intersection. Phase 2 introduces a double-blind cross-review mechanism among experts to resolve ambiguities and enforce standards. Finally, Phase 3 employs a human-in-the-loop QA process via the Onon platform, meticulously parallelizing Cyrillic and Traditional scripts to filter out machine-translation artifacts and guarantee absolute semantic equivalence. This rigorous, native-expert-driven curation strategy aligns with recent best practices in developing high-fidelity quality assurance and reading comprehension datasets for other low-resource languages, such as Tibetan (Dan and Sun, 2024).

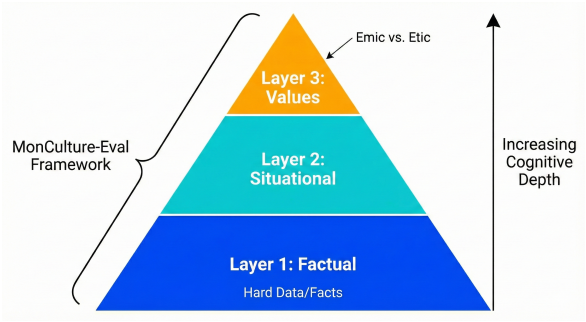

Figure 2: The hierarchical structure of MonCulture-Eval, comprising the three-layer cognitive framework and specialized cultural tasks.

4.2 Taxonomy and Task Formulation

To systematically isolate factual memorization from deep cultural alignment, the dataset follows a hierarchical cognitive architecture, illustrated in Figure 2. The four specialized cultural tasks are seamlessly integrated into this hierarchy based on their evaluation objectives.

- Layer 1: Factual Knowledge. Evaluates explicit declarative knowledge regarding geographical concepts, historical events, and traditional artifacts (e.g., components of a Mongolian yurt). This layer utilizes objective single-choice and multiple-choice questions. The data distribution encompasses both specific regional cultures (Alxa, Horqin, Ordos) and pan-Mongolian domains, specifically General History and Ritual practices.

- Layer 2: Situational Appropriateness. Assesses pragmatic social behaviors and the model's capacity to conform to indigenous norms. Evaluating LLMs on such cultural and moral norms has become increasingly critical (Ramezani and Xu, 2023). Beyond standard situational single-choice questions, this layer incorporates the Taboos task. Formatted as open-ended generation, the Taboos task requires models to process a daily scenario, explicitly identify embedded normative violations, explain the consequences, and propose the correct behavior.

- Layer 3: Values Layer. Probes the deep ontological logic driving cultural practices. The foundational structure relies on "best-explanation" single-choice questions, where models must distinguish between secular functional logic (Etic/outsider), surface-level norms (Shallow Emic), and authentic spiritual reasoning (Emic/insider). Furthermore, this layer governs three specialized tasks requiring complex cultural mapping:

  – Riddles (Metaphorical QA): Evaluates the capacity to map abstract metaphors to physical entities based on nomadic ecological knowledge. As highlighted by recent benchmarking efforts in other complex languages (Liu et al., 2025), the ability to process non-literal, metaphorical logic is a critical indicator of deep semantic alignment in LLMs.
  – Proverbs (Contextual Matching): A single-choice task requiring the alignment of traditional proverbs with complex social scenarios, rigorously testing polarity consistency.
  – Benedictions (Sequence Sorting): Assesses the pragmatic and syntactic mastery of traditional blessing formulas through LLM-evaluated sequence reconstruction and coherence scoring.

4.3 Data Statistics

The final dataset is structured to rigorously evaluate the models across the three cognitive layers and the specialized generative tasks. All questions are formatted symmetrically for both Traditional and Cyrillic Mongolian scripts. The comprehensive statistical distribution of the benchmark is detailed in Table 1.

Domain	L1: Facts	L2: Situations	L3: Values	Total
Regional Culture				
Alxa	205	62	79	346
Horqin	183	47	63	293
Ordos	125	86	122	333
General Culture				
History	123	-	-	123
Culture & Hist.	172	-	-	172
Rituals	180	-	34	214
Special Tasks				
Riddles	-	-	523	523
Taboos	-	610	-	610
Proverbs	-	-	297	297
Benedictions	-	-	122	122
Total	**988**	**805**	**1,118**	**2,911**

Table 1: Statistics of the MonCulture-Eval Benchmark.

5 Experimental Setup

5.1 Models Evaluated

To establish a comprehensive baseline of cultural capabilities, we evaluated five models. We se-

Model	Script	Factual Layer		Situational Layer		Values Layer			
		Single (%)	Multi (%)	Single (%)	Taboos (Gem/GPT)	Single (%)	Benedictions (Gem/GPT)	Riddles (Gem/GPT)	Proverbs (Gem/GPT)
Gemini-3-Pro-Preview	Trad.	**79.48**	**23.28**	**84.62**	**8.87 / 7.84**	**98.99**	**8.28 / 8.31**	**6.54 / 7.84**	**9.15 / 9.02**
	Cyr.	**76.72**	29.01	**85.13**	**9.12 / 8.27**	**97.65**	**9.51 / 9.42**	**7.00 / 5.58**	**9.38 / 8.92**
Qwen3-Max	Trad.	44.63	15.27	55.38	3.52 / 7.47	85.57	2.93 / 3.66	2.82 / 1.57	2.36 / 7.07
	Cyr.	58.82	**35.88**	67.69	6.01 / 7.35	92.95	6.01 / 5.98	2.82 / 2.22	8.25 / 8.03
GPT-5.2	Trad.	40.36	9.16	40.51	6.09 / 7.49	68.12	4.89 / 5.63	1.97 / 1.84	6.78 / 6.88
	Cyr.	60.47	30.53	65.13	8.29 / 7.95	94.97	6.93 / 6.70	3.75 / 3.75	8.97 / 8.76
Claude-Sonnet-4.5	Trad.	33.33	2.29	31.28	3.30 / 6.37	58.72	5.20 / 4.89	0.97 / 1.56	5.56 / 6.46
	Cyr.	56.47	9.92	62.56	7.12 / 4.35	95.64	6.39 / 6.49	4.06 / 3.00	8.85 / 8.44
DeepSeek-v3.2	Trad.	28.93	0.76	33.85	1.53 / 4.51	51.34	2.77 / 3.02	0.20 / 1.26	2.45 / 3.86
	Cyr.	48.35	6.87	55.38	6.03 / 4.57	89.93	5.53 / 5.44	2.32 / 1.96	8.68 / 8.14

Table 2: Comprehensive performance across cognitive layers. Objective tasks use accuracy, while generative tasks report average scores (out of 10) evaluated by Gemini-2.5-Pro and GPT-5.2.

lected GPT-5.2 (OpenAI, 2026), Gemini-3-Pro-Preview (Gemini Team, 2025), Claude-Sonnet-4.5 (Anthropic, 2025), and DeepSeek-v3.2 (DeepSeek-AI, 2024) to represent the frontier of global multilingual systems. Additionally, we included the regionally optimized Qwen3-Max (Yang et al., 2025) to provide a robust comparison baseline tuned specifically for Asian and inner-Asian linguistic contexts.

5.2 Prompt Engineering Strategy

We employed a systematic prompting strategy tailored to the cognitive nature of each task (full templates are provided in Appendix A). For all tasks, we explicitly instructed models to adopt specific professional personas (e.g., "Professional Archivist"). For complex reasoning tasks such as Riddles, inspired by recent advancements in eliciting multi-step reasoning from large models (Wei et al., 2023; Kojima et al., 2023), we utilized a two-stage dialogue process isolating answer generation from rationalization. For generative evaluation tasks (Taboos and Benedictions), strict JSON output schemas were enforced to guarantee parsing reliability.

5.3 Evaluation Protocol and Cross-Judge Validation

For objective multiple-choice questions, accuracy was computed directly. For open-ended generative tasks, we adopted an LLM-as-a-judge paradigm (Zheng et al., 2023), utilizing Gemini-2.5-Pro as the primary automated evaluator scoring outputs on a 10-point scale based on expert rubrics.

To mitigate potential family-lineage bias from a single AI evaluator and validate our pipeline, we implemented a rigorous dual-judge validation protocol. Rather than relying solely on one model, we evaluated all open-ended outputs using both Gemini-2.5-Pro and GPT-5.2 as parallel, independent judges.The Pearson correlation (r) between the rankings of the two judges exceeded 0.85, confirming a high consensus on relative model performance (see Table 2 for detailed distributions). Notably, native experts manually sampled 10% of the automated evaluations, achieving a strict agreement rate of 72.1% and a tolerance-based (± 1 point) agreement of 85.4% with the primary AI judge. Comprehensive adjudication metrics and quality control procedures are detailed in Appendix B.

5.4 Evaluation Metrics: The Motivation for SAC

Evaluating languages characterized by complex digraphia (the use of two different writing systems for the same language) requires specialized metrics beyond standard accuracy. As demonstrated by recent studies on systemic inequalities in NLP (Blasi et al., 2022) and the downstream performance impact of multilingual tokenizers (Rust et al., 2021), standard evaluation often masks a model's reliance on higher-resource scripts due to cross-lingual tokenization penalties in traditional orthographies. To quantify structural robustness, we introduce **Script Alignment Consistency (SAC)**. SAC measures the proportion of instances where a model successfully answers the identical question in both writing systems simultaneously:

$$SAC = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(y_i^{trad} = y_i^* \wedge y_i^{cyr} = y_i^*) \quad (1)$$

where y_i^{trad} and y_i^{cyr} represent predictions in Traditional and Cyrillic scripts, and y_i^* is the ground truth. Specifically, for objective tasks, the condition $y = y^*$ is met if the model selects the correct option. For subjective generative tasks, this condition is satisfied if the judge's score $S \geq 8$ for the given response. A high SAC score demonstrates "Ontological Alignment"—indicating the model pro-

cesses stable cultural concepts independent of the visual script surface.

6 Results and Analysis

6.1 Regional Knowledge Disparities

A critical yet often overlooked dimension of cultural intelligence is regional representation. Mongolian culture is not monolithic; it comprises distinct regional archetypes. To avoid treating the culture as a homogeneous entity, MonCulture-Eval evaluates models across three representative tribes: Alxa (Western desert-steppe), Ordos (Central ceremonial), and Horqin (Eastern agrarian-pastoral).

Table 3 disaggregates model performance on objective Factual tasks across these specific regional subcultures alongside general cultural knowledge. A common assumption in evaluating low-resource languages is that models severely marginalize peripheral tribes in favor of a central dialect. However, our empirical data reveals a more nuanced reality. As shown in Table 3, while models generally exhibit a slight factual preference for the more heavily documented Horqin culture, the performance variance across the three regions remains relatively stable. For example, Qwen3-Max's accuracy in the Traditional script spans from 31.25% (Ordos) to 44.85% (Horqin)—an observable gap, but not a cliff-like comprehension failure. This indicates that regionality influences factual retrieval, but does not create an absolute epistemological barrier among different tribal archetypes.

More importantly, a cross-layer analysis (comparing Table 2 and Table 3) reveals a profound architectural insight: models exhibit significantly higher robustness in regional Values than in regional Facts. For instance, while Qwen3-Max achieves only 38.29% on Alxa factual questions, its performance on Alxa values layer reasoning surges to 82.28%. A similar trend is observed in GPT-5.2, where Ordos factual accuracy sits at 35.00%, yet its Ordos values layer reaches 75.41%.

This systemic disparity suggests that while granular, localized facts (e.g., specific tribal history, artifact components) are highly vulnerable to pretraining data scarcity, the underlying abstract cultural logic and spiritual values form a more generalized, resilient representation in the LLMs' latent space. Consequently, models can successfully align with regional nomadic values even when they fail to retrieve the specific surface-level facts of that exact region.

6.2 Comprehensive Task Performance

Table 2 presents the global performance landscape of all evaluated models, systematically organized by our three-layer cognitive framework and their respective sub-tasks. Objective tasks are reported in accuracy percentages, while generative tasks (Taboos, Riddles, Proverbs, Benedictions) report the average judge scores (out of 10) from both Gemini-2.5-Pro and GPT-5.2, denoted as (Gem/GPT).

The data reveals a stark cognitive hierarchy. Gemini-3-Pro maintains a dominant performance across all layers, demonstrating deep cultural internalization. Conversely, other frontier models exhibit a systemic collapse as tasks transition from Factual memorization to Values and generation. Notably, tasks requiring deep metaphorical mapping, such as Riddles, expose severe limitations across all models in the Traditional script, highlighting the boundary of current multimodal and multilingual capabilities.

6.3 The Script Gap and Alignment Consistency

Beyond task and regional complexity, models suffer from a severe Script Gap. As observed in Table 2, almost all models default to higher accuracy and scores in the Cyrillic script, likely due to parameter transfer from Russian corpora and higher resource availability on the modern web. Overcoming this script-bound limitation and facilitating genuine cross-lingual transfer in low-resource orthographies remains a significant frontier in current LLM alignment research (Zhuang et al., 2025b; Lu et al., 2025). To comprehensively visualize this structural deficit, Figure 3 illustrates the Script Alignment Consistency (SAC) across various objective tasks. SAC measures the proportion of instances where a model correctly answers a question in both writing systems simultaneously. A negative performance gap mathematically bounds their maximum possible SAC, proving that their cultural knowledge is script-bound rather than ontologically grounded.

As depicted in Figure 3, Gemini-3-Pro-preview stands as the primary exception, achieving a commanding SAC across all objective dimensions and demonstrating that it maps both writing systems to a shared semantic embedding space. Qwen3-Max, benefiting from regional linguistic optimization, exhibits moderate robustness, notably achieving

Model	Script	Alxa Culture (%)	Horqin Culture (%)	Ordos Culture (%)	General Culture (%)
Gemini-3-Pro	Trad.	**77.14**	**77.21**	**72.50**	**83.28**
	Cyr.	**72.57**	**73.53**	**72.50**	**81.49**
Qwen3-Max	Trad.	38.29	44.85	31.25	51.04
	Cyr.	52.00	60.29	46.25	64.78
GPT-5.2	Trad.	34.86	42.65	35.00	43.58
	Cyr.	49.14	60.29	58.75	66.87
Claude-Sonnet	Trad.	28.57	41.91	26.25	34.03
	Cyr.	45.71	66.18	40.00	52.24
DeepSeek-v3.2	Trad.	29.71	30.88	22.50	29.25
	Cyr.	39.43	50.74	35.00	51.94

Table 3: Task performance breakdown by region. The cross-layer analysis indicates that while models experience factual variance across tribes, their value-alignment across different regions remains highly resilient.

a highly competitive SAC in the Values Single Choice task. Notably, we observe that for top-tier models like Gemini-3-Pro-preview and Qwen3-Max, the SAC often increases as the cognitive demand moves from Factual (L1) to Values (L3). This suggests that while surface-level facts are sensitive to script-specific training data, the underlying cultural logic (Values) achieves a more robust ontological alignment across different writing systems. However, a pervasive trend across the broader model suite is the precipitous degradation of SAC as the cognitive demand shifts from straightforward Factual retrieval to complex Situational pragmatics.

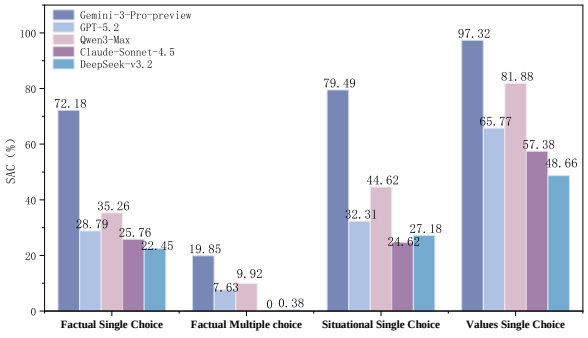

Figure 3: Script Alignment Consistency (SAC) across objective tasks (Factual, Situational, and Values layers). A high SAC score indicates true ontological understanding that is independent of the visual script surface.

To further validate these findings against potential evaluation biases in generative tasks, we introduce a Cross-Judge SAC analysis, illustrated in Figure 4. This framework evaluates the alignment of model outputs across four subjective cultural tasks (Taboos, Riddles, Proverbs, and Benedictions) utilizing both Gemini-2.5-Pro and GPT-5.2 as independent arbiters. Crucially, the relative performance hierarchy—where Gemini-3-Pro-preview consistently outperforms GPT-5.2, followed by Qwen3-Max, Claude-Sonnet-4.5, and DeepSeek-v3.2—remains remarkably stable regardless of the judge employed. For instance, in the complex Taboos task, the SAC score parity between the two distinct AI judges provides robust empirical evidence. It confirms that the superior performance of leading models is rooted in profound cultural depth and objective ontological alignment, effectively refuting concerns regarding LLM-as-a-judge lineage bias or evaluation circularity.

6.4 Cognitive Depth Drop-off and Etic Bias

To systematically quantify how models process cultural reasoning across different writing systems, Table 4 presents the distribution of perspective alignments (Emic, Shallow Emic, and Etic) for all evaluated models in both the Traditional and Cyrillic Mongolian scripts within the Values Layer.

Model	Emic (%)		S. Emic (%)		Etic (%)	
	Trad.	Cyr.	Trad.	Cyr.	Trad.	Cyr.
Gemini-3-Pro-Preview	96.97	97.65	1.01	1.01	2.01	1.34
Qwen3-Max	85.57	92.95	6.71	3.36	7.72	3.69
GPT-5.2	68.12	95.93	11.74	2.37	20.13	1.69
Claude-Sonnet-4.5	58.72	95.64	15.10	2.01	26.17	2.35
DeepSeek-v3.2	51.34	89.93	16.11	4.70	32.55	5.37

Table 4: Distribution of cultural perspectives within the Values Layer across Traditional (Trad.) and Cyrillic (Cyr.) scripts. Models are ordered descendingly by their Emic alignment in the Traditional script, highlighting the severe Etic Bias inflation in low-resource settings.

The empirical data reveals a stark cross-model divergence in cognitive depth that is heavily exacerbated by the script gap. When processing the higher-resource Cyrillic script, most models achieve robust Emic alignment, with GPT-5.2 and

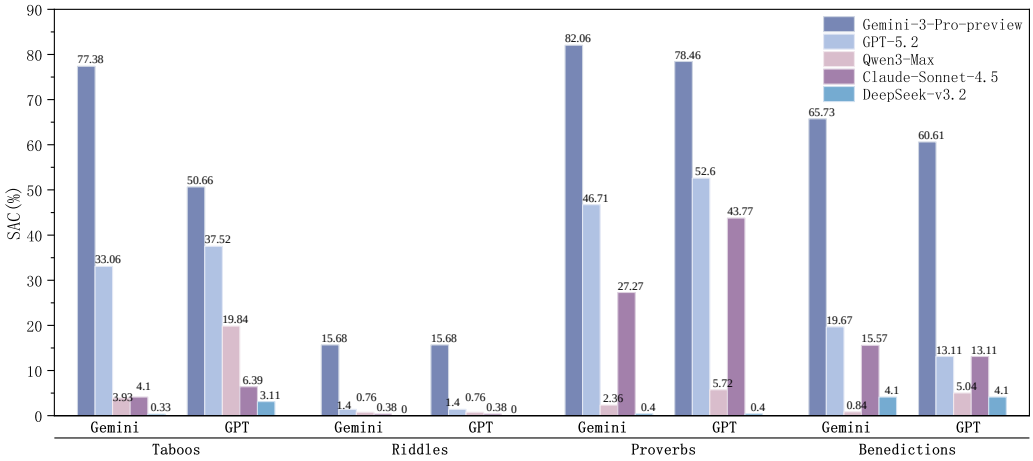

Figure 4: Cross-Judge Script Alignment Consistency (SAC) for subjective generative tasks. The consistent model ranking across both Gemini-2.5-Pro and GPT-5.2 judges validates the robustness of the evaluation and mitigates concerns regarding lineage bias. Note: "GPT" and "Gemini" in the legend refer to GPT-5.2 and Gemini-3-Pro-preview, respectively.

Claude-Sonnet-4.5 reaching 95.93% and 95.64% respectively. However, a transition to the Traditional script triggers a severe cognitive drop-off. While Gemini-3-Pro-Preview maintains cross-script resilience (96.97% Emic) and the regionally optimized Qwen3-Max retains a strong 85.57% Emic rate due to its targeted inner-Asian linguistic exposure, models trained primarily on generic global corpora exhibit a pervasive Etic Bias. In the Traditional script, GPT-5.2's Etic selection surges to 20.13%, despite maintaining moderate accuracy on standard factual tasks. This structural deficit is even more pronounced in Claude-Sonnet-4.5 and DeepSeek-v3.2, which yield secular Etic explanations in 26.17% and 32.55% of cases, respectively.

This systemic default to the "Tourist Perspective" in low-resource orthographies exposes a profound architectural vulnerability. Faced with the semantic opacity of traditional paradigms, models retreat to majority-culture logic or generic functionalism. By substituting sacred rationales with universally palatable secular habits, these systems systematically impose cross-cultural normative priors onto marginalized knowledge structures, a bias recognized in recent global evaluations (Durmus et al., 2024; Xu et al., 2025).

Furthermore, in their attempt to bridge the gap between an unfamiliar indigenous prompt and external secular priors, models frequently generate authoritative hallucinations in low-resource metaphors. They construct logically coherent, yet ontologically vacant, rationalizations for unique cultural behaviors (Ji et al., 2023). While this functionalist reduction may appear harmless in generic translation contexts, it is highly detrimental to cultural fidelity for preservation-oriented applications. The failure to align with the authentic Emic perspective inadvertently dilutes the structural integrity of the very culture the AI attempts to represent.To further contextualize these statistical performance gaps, detailed qualitative case studies illustrating epistemological reduction and contextual mapping failures are provided in Appendix C.

7 Discussion

The empirical results presented in Section 6 demonstrate a systematic divergence between linguistic fluency and cultural alignment. This section examines the mechanisms driving these failures, specifically focusing on logical collapses, normative biases, and the persistence of architectural bottlenecks.

7.1 Beyond Surface Translation: The Ritualized Reasoning Collapse

The performance metrics indicate that models appearing fluent in low-resource scripts frequently fail at the level of ritualized logic. As demonstrated in Table 2, while the regionally optimized Qwen3-Max achieves a 44.63% accuracy on Factual Single Choice questions in the Traditional script, its performance drops significantly in the Benedictions task, scoring only 2.93 (Gemini judge) and 3.66 (GPT judge). This empirical gap suggests that the difficulty lies not merely in lexical retrieval, but in

the preservation of procedural logic inherent in cultural rituals. When ritual sequences are disrupted, the models default to generic semantic associations, resulting in a structural fracture where the cultural form is maintained while the functional essence is lost.

7.2 Imposing Cross-Cultural Normative Priors

Our analysis of perspective alignments in Table 4 reveals a consistent empirical trend: models frequently default to functional or secular explanations when processing traditional contexts. In the Traditional script, frontier models such as GPT-5.2, Claude-Sonnet-4.5, and DeepSeek-v3.2 exhibit a pervasive reliance on Etic selection, peaking at over a third of the evaluations. This provides an empirical observation of the models' reliance on majority-culture logical priors when faced with indigenous paradigms.

From a normative standpoint, this systematic default is detrimental to cultural fidelity for preservation-oriented applications. This systemic reliance on etic reasoning demonstrates how LLMs act as normative filters. Rather than preserving the ontological distinctiveness of indigenous practices, the models reconstruct them to align with majority-culture utility (Durmus et al., 2024; Talat et al., 2022; Xu et al., 2025). This perspective prioritizes universalist explanations over localized ritual meanings, subtly recalibrating local knowledge to fit paradigms favored in large-scale training corpora.

7.3 Authoritative Hallucinations in Low-Resource Metaphors

A distinct failure mode observed in tasks requiring deep metaphorical mapping is the generation of authoritative hallucinations in low-resource metaphors. As detailed in Appendix Table 6, when models lack specific cultural context in the Riddle task, they tend to construct plausible yet false causal links.DeepSeek-v3.2 generated 402 instances of "Forced Logic" errors, and Claude-Sonnet produced 379. These hallucinations are particularly deceptive because the models maintain a high degree of confidence and utilize the correct target script, masking the underlying logical void (Ji et al., 2023).

7.4 The Persistence of the Script Gap and Cross-Judge Validation

The structural decay observed across writing systems confirms that the "Script Gap" remains a persistent architectural bottleneck. As illustrated by the Script Alignment Consistency (SAC) scores in Figure 3, models bounded by negative performance gaps fail to map both writing systems to a shared semantic embedding space. For instance, Qwen3-Max experiences a -14.19% performance drop between Cyrillic and Traditional scripts.

To empirically refute potential LLM-as-a-judge evaluation circularity or lineage bias, we introduced a Cross-Judge SAC analysis (Figure 4). The relative performance hierarchy remains highly stable whether evaluated by Gemini-2.5-Pro or GPT-5.2. Gemini-3-Pro-preview consistently outperforms other models across both independent arbiters, validating that its superior performance is rooted in objective ontological alignment rather than family-lineage preference.Bridging this gap requires advancing beyond mere data scaling toward script-agnostic representation learning and robust cross-lingual knowledge transfer architectures (Zhuang et al., 2025b; Zhuang and Sun, 2025).

8 Conclusion

This study provides an empirical evaluation of LLMs regarding the cultural and logical complexities of Mongolian heritage.Through the implementation of a native-centric, "Indigenous-First" curation pipeline and a three-layer cognitive framework, our findings highlight three structural barriers in current AI systems:(1) an architectural Script Gap that restricts access to traditional archives, (2) a systemic Etic Bias that leads to imposing cross-cultural normative priors onto localized knowledge, and (3) a high frequency of authoritative hallucinations in low-resource metaphors.

While models like Gemini-3-Pro demonstrate the capacity for deep cultural alignment—validated by robust cross-judge consistency—the broader results indicate that linguistic translation proficiency does not guarantee cultural value alignment.Future research must transition toward Pluralistic Alignment, ensuring that foundational models integrate diverse ontological frameworks and serve as faithful custodians of global cultural diversity.

9 Limitations

LLM-as-a-Judge Lineage Bias The evaluation of subjective generative tasks (e.g., Riddles, Proverbs) relies primarily on Gemini-2.5-Pro as the automated adjudicator. his introduces a potential lineage bias, where the judge might favor outputs from models sharing similar training distributions or reinforcement learning pipelines (Zheng et al., 2023). To empirically mitigate this, we introduced the Cross-Judge Script Alignment Consistency (SAC) analysis detailed in Figure 4. Furthermore, the human-verification phase established a 72.1% strict Inter-Annotator Agreement (IAA) and an 85.4% tolerance-based agreement with the AI judge, confirming the alignment between automated scoring and native expert intuition.

Instruction Following and Prompt Sensitivity The cultural knowledge retrieval in this study is elicited using standardized persona-based prompts (e.g., assigning a "Professional Archivist" role). Models exhibit varying degrees of sensitivity to specific instruction formats. Proprietary and open-weight models may require different prompt engineering strategies to optimally trigger their latent cultural knowledge. Consequently, the observed performance gaps reflect a combination of cultural alignment and base instruction-following capabilities.

Static Evaluation of Dynamic Cultural Evolution MonCulture-Eval evaluates cultural intelligence using a static set of historical and traditional norms. In practice, Mongolian culture is continuously evolving, with traditional nomadic practices adapting to modern, urbanized environments. The benchmark does not capture these contemporary cultural shifts or hybrid social norms.

10 Ethical Considerations

Indigenous Data Sovereignty and Epistemological Diversity This benchmark was constructed following an "Indigenous-First" protocol, relying exclusively on paid native domain experts and primary offline archives rather than web-scraped corpora. This methodology ensures data contamination insulation while respecting indigenous data sovereignty. By structurally avoiding reliance on majority-culture web data, this approach protects the epistemological diversity of the Mongolian community and prevents the extraction of cultural artifacts without proper contextual grounding.

Avoidance of Cultural Stereotyping The dataset is designed to measure a model's capacity to process localized social logic, not to establish rigid, prescriptive stereotypes of modern Mongolian individuals. The deliberate inclusion of distinct regional sub-cultures (Alxa, Ordos, and Horqin), as analyzed in Table 3, serves to reflect the internal diversity of the Mongolosphere. Users must recognize that cultural norms operate as contextual reference points rather than universal behavioral laws.

Dual-Use Risks in Cultural Alignment Enhancing a model's capacity to generate culturally authentic, emic-aligned text introduces inherent dual-use risks. Systems possessing deep cultural fidelity could theoretically be deployed to generate highly persuasive, culturally targeted disinformation. However, we assess that the benefits of developing language technologies that accurately reflect and serve marginalized linguistic communities outweigh the theoretical risks of misuse.

Acknowledgements

This work is supported by Science and Technology Strategic Consulting Project of the Chinese Academy of Engineering (2025-XZ-16-06), Project of the China Tibetology Literature and Resources Data Center (2025SJ003), the National Social Science Foundation (22&ZD035).

References

Anthropic. 2025. Introducing Claude Sonnet 4.5.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Zhengcuo Dan and Yuan Sun. 2024. Tibetanqa2.0: Dataset with unanswerable questions for tibetan machine reading comprehension. *DATA INTELLIGENCE*, 6(4):1158–1167.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. *Preprint*, arXiv:2306.16388.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Gemini Team. 2025. A new era of intelligence with gemini 3.

E.T. Hall. 1976. *Beyond Culture*. Knopf Doubleday Publishing Group.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in indonesia: A comprehensive test on indommlu. *Preprint*, arXiv:2310.04928.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966, Marseille, France. European Language Resources Association.

Yan Liu, Renren Jin, Tianhao Shen, and Deyi Xiong. 2025. Cmgbench: Benchmarking chinese metaphor generation for large language models. *DATA INTELLIGENCE*, 7(4):1270–1290.

Kaiwen Lu, Yating Yang, Fengyi Yang, Rui Dong, Bo Ma, Aihetamujiang Aihemaiti, Abibilla Atawulla, Lei Wang, and Xi Zhou. 2025. Low-resource language expansion and translation capacity enhancement for LLM: A study on the Uyghur. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8360–8373, Abu Dhabi, UAE. Association for Computational Linguistics.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2024. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *Preprint*, arXiv:2309.12342.

OpenAI. 2026. OpenAI GPT-5 system card. *CoRR*, abs/2601.03267.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual 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 3118–3135, Online. Association for Computational Linguistics.

E.H. Schein. 1985. *Organizational Culture and Leadership*. A Joint publication in the Jossey-Bass management series and the Jossey-Bass social and behavioral science series. Jossey-Bass Publishers.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A roadmap to pluralistic alignment. *Preprint*, arXiv:2402.05070.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2025. Self-pluralising culture alignment for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 6859–6877. Association for Computational Linguistics.

Qwen Team: An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Wenhao Zhuang, Dawa Cairen, and Yuan Sun. 2025a. Tifd: Tibetan instruction-following dataset for large language models supervised fine-tuning. *DATA INTELLIGENCE*, 7(3):776–785.

Wenhao Zhuang and Yuan Sun. 2025. CUTE: A multilingual dataset for enhancing cross-lingual knowledge transfer in low-resource languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10037–10046, Abu Dhabi, UAE. Association for Computational Linguistics.

Wenhao Zhuang, Yuan Sun, and Xiaobing Zhao. 2025b. Enhancing cross-lingual transfer through reversible transliteration: A huffman-based approach for low-resource languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16299–16313.

A Prompt Library

To ensure reproducibility, enforce strict output structures (e.g., JSON schemas), and mitigate instruction-following variance across models, we designed specialized prompt templates for each cognitive layer and task. The following templates illustrate the exact English instructions provided to the models.

Cross-Script Alignment Note: To guarantee strict experimental alignment between the two writing systems, the prompt templates used for both Cyrillic and Traditional Mongolian are entirely identical. The only variation introduced is the explicit constraint regarding the target output script (e.g., requesting responses in "Traditional Mongolian" versus "Cyrillic Mongolian"). All task-specific data (questions, scenarios, and options) were injected using their respective scripts.

A.1 Objective Tasks (Multiple Choice)

For standard factual and situational objective tasks, we utilized a professional expert persona to ground the model in the appropriate cultural context. For the Values layer, we added a specific instruction to enforce emic alignment.

System Prompt: Objective Tasks (Layers 1 & 2)

```
You are an eminent expert in Mongolian culture, history, and the traditional Mongolian script. Your task is to accurately answer the following questions based on your profound knowledge.

Always stand on the standpoint of a Mongolian person when solving problems. The questions and options are provided in traditional Mongolian.

Please output ONLY the letter(s) of the correct option(s) (e.g., A, B, C, or D). Do not provide any explanations, punctuation, or additional text.
```

System Prompt: Values Layer (Layer 3)

You are an eminent expert in Mongolian culture, history, and traditional social values.

Always stand on the standpoint of a Mongolian person when solving problems. Your task is to analyze the following question and identify the option that best reflects authentic Mongolian cultural values.

Please output ONLY the letter of the correct option. Do not provide any explanations, punctuation, or additional text.

A.2 Taboo Task (Situational Pragmatics)

This task requires the model to identify specific normative violations in a given scenario and output a structured JSON array for automated evaluation.

System Prompt: Taboo Answer Generation

You are a professional archivist at the Mongolian Cultural Heritage Center. Your mission is to document traditional social etiquettes and lifestyle norms to preserve them for future generations. You are summarizing educational material for a museum exhibition.

Analyze the behavior in the scenario and document it in JSON format using these keys:
1. error_point: The specific traditional behavior observed.
2. reason: The symbolic meaning or cultural teaching behind it.
3. correction: The recommended alternative practice.

STRICT RULE: Output ONLY a valid JSON array. Values MUST be in Traditional Mongolian script.

LLM-as-a-Judge Prompt: Taboo Evaluation

You are a senior expert in Mongolian customs and cultural taboos. Your task is to evaluate a model's analysis of a cultural scenario.

SCORING LOGIC:
1. **Identify Error Points (The Foundation):** Check if the model identified the same or semantically similar error points as the Reference.
2. **Cultural Reasoning (The Depth):** Does the reason mention specific Mongolian cultural concepts (e.g., fortune/Buyan, bad omens, respect for elders/nature)? Deduct points if generic.
3. **Correction (The Practicality):** Is the suggested correction consistent with traditional Mongolian etiquette?

SCORING RUBRIC (0-10):
- [9-10]: Correct error points + Accurate cultural reasons + Correct traditional corrections.
- [7-8]: Correct error points + Reasonable but slightly generic reasons or minor flaws in correction.
- [4-6]: Partially correct/generic + Reason lacks deep Mongolian cultural mapping.
- [0-3]: Wrong error points identified + Incorrect/irrelevant reasoning.

Output Format:
Reasoning: [Briefly analyze point-by-point comparison]
Final Score: [0-10]

A.3 Riddles Task (Metaphorical Reasoning)

To prevent models from hallucinating justifications before committing to an answer, we employed a strict two-round dialogue prompt.

System Prompt: Riddles Answer Generation

You are an eminent expert in Mongolian culture and traditional riddles. Please provide your responses in Traditional Mongolian (Mongol Bichig).

Rule: Provide ONLY the answer in the first round. The explanation must be under 50 words in the second round.

User (Round 1): [Mongolian instruction: Guess the riddle]
User (Round 2): [Mongolian instruction: Explain derivation]

LLM-as-a-Judge Prompt: Riddles Evaluation

You are a professional judge for Mongolian riddles. Your task is to score a model's riddle answer based on a standard answer and a reference explanation.

Evaluation Criteria:
- 9-10: The answer is correct and the explanation is logical.
- 6-8: The answer is correct, but the explanation is weak or slightly irrelevant.
- 3-5: The answer is incorrect but very close/reasonable.
- 0-2: Both answer and explanation are wrong.

Task: First, briefly analyze the logic, then provide the final score. You MUST end your response with the format: "Final Score: X" (where X is 0-10).

A.4 Proverbs Task (Contextual Matching)

For this complex reasoning task, we required the models to output a Chain-of-Thought (CoT) process within a strict JSON schema, processing the information natively.

System Prompt: Proverbs Answer Generation

You are an AI natively fluent in Traditional Mongolian and an absolute expert in Mongolian folklore, nomadic culture, and proverbs. You process and analyze information directly in Mongolian without translating it into English or Chinese.

Workflow & Chain of Thought
1. [Contextual Immersion]: Grasp the core conflict and underlying moral of the Mongolian scenario. Focus on cultural nuance.
2. [Native Proverb Analysis]: Evaluate options A, B, C, and D individually. Extract the direct metaphorical meaning.
3. [Logical Synthesis & Elimination]: Identify why three proverbs represent a semantic mismatch. Confirm why the remaining proverb is the exact, culturally appropriate match.
4. [Final Output]: Conclude with the correct option letter.

Please output your response strictly in JSON format:
{ "thought_process": { "scenario_analysis": "...", "options_analysis": { "A": "...", "B": "...", "C": "...", "D": "..." }, "alignment_conclusion": "..." }, "final_answer": "A/B/C/D" }

LLM-as-a-Judge Prompt: Proverbs Evaluation

You are an impartial, highly rigorous AI Evaluator and an expert in Traditional Mongolian linguistics, nomadic folklore, and AI benchmarking. Your task is to evaluate the quality of another AI model's response to a multiple-choice reading comprehension question based on Traditional Mongolian proverbs.

Evaluation Criteria (10-Point Scale)
* [10 points] - Perfect: Final answer is CORRECT. Thought process shows flawless, native-level understanding. Distractors eliminated logically. No hallucinations.
* [8-9 points] - Excellent: Final answer is CORRECT. Strong comprehension overall, but minor nuances in a distractor were glossed over. Core logic is sound.
* [6-7 points] - Good/Lucky: Final answer is CORRECT, BUT reasoning is shallow (e.g., keyword matching) OR contains noticeable translation errors.
* [4-5 points] - Mediocre: Final answer is INCORRECT, BUT the thought process demonstrates a genuine partial understanding.
* [2-3 points] - Poor: Final answer is INCORRECT. Severe hallucinations, failed to understand the scenario's core conflict.
* [0-1 points] - Fail: Completely irrelevant response, gibberish, or extreme hallucination.

A.5 Benedictions Task (Sequence Reconstruction)

This task assesses grammatical and ritualistic sequencing capabilities.

System Prompt: Benedictions Answer Generation

You are a scholar of classical Mongolian poetic rhetoric and folk culture, highly proficient in the traditional Mongolian script and the structural logic of Mongolian blessings and praise poetry (Iruge-l).

You will receive an array of shuffled traditional Mongolian poetic lines. Your task is to restore these lines to their exact original sequence.

CRITICAL ENGINEERING RULE: The length of your 'correct_order' array MUST exactly match the number of lines provided in the 'shuffled_lines' array. Output ONLY a strict JSON object.

LLM-as-a-Judge Prompt: Benedictions Evaluation

```
You are an expert evaluator specialized in assessing the logical reasoning, zero-shot capabilities, and cultural alignment of Large Language Models on Traditional Mongolian.

Evaluate the tested model's performance on a 10-Point Scale.
Dimension 1: Sequence Accuracy (Order Score) - [Max: 5 Points]
* 5 Points: Exact match. / 3 Points: Grouped correct rhyming pairs but swapped two semantically parallel lines. / 1 Point: Mostly wrong, but identified the starting or ending line. / 0 Points: Completely wrong.
Dimension 2: Reasoning Quality (Reasoning Score) - [Max: 5 Points]
* 5 Points: Accurately identifies head rhyme and semantic progression. / 4 Points: Explains logic well but misses head rhyme, or vice versa. / 2-3 Points: Vague reasoning. / 1 Point: Severe Hallucination. / 0 Points: Fail.

Output ONLY a valid JSON object.
```

B Quality Control & Adjudication Metrics

To ensure the high reliability and "Emic" authenticity of MonCulture-Eval, we implemented a rigorous, multi-stage quality control pipeline. This section details our expert recruitment criteria, annotation workflow, and quantitative adjudication metrics.

B.1 Expert Recruitment and Annotation Workflow

We assembled a professional team of 12 native Mongolian domain experts. To ensure diverse and authentic regional representation, recruitment required: (1) a minimum of 15 years of residency in the target regions (Alxa, Ordos, or Horqin); and (2) at least a Bachelor's or Master's degree in Mongolian Linguistics, Folklore, or Cultural History. The team was evenly balanced with 4 experts per region.

Data construction was guided by over a dozen authentic local volumes (e.g., *Culture of Ordos*, *Culture of Horqin*, *Overview of Mongolian History*). Reflecting task complexity, experts were compensated at 30 RMB/hour for choice-based items and 50 RMB/hour for specialized generative tasks (Riddles, Taboos, etc.). Every item underwent a double-blind peer review by a second regional expert to verify cultural accuracy. In cases of disagreement (which occurred in less than 5% of items), experts from the corresponding region voted collectively, with seniority serving as the tie-breaker.

B.2 Cross-Script Semantic Fidelity

The Cyrillic version of the benchmark serves as a parallel corpus to strictly isolate the "Script Gap." To guarantee semantic equivalence and prevent translation-induced biases:

- **Back-Translation Verification:** We utilized the Onon platform to back-translate the Cyrillic text into the Traditional script.
- **Human-in-the-Loop QA:** Experts manually compared the back-translation with the gold-standard original text, correcting semantic or morphological discrepancies in 14.2% of the segments. A double-blind verification on a 20% sample of these refined segments yielded a Cohen's Kappa of 0.82, demonstrating high reliability in preserving semantic fidelity across scripts.

B.3 Quantitative Adjudication Metrics

To provide a numerical assessment of our evaluation pipeline and validate the reliability of our primary AI Judge (Gemini-2.5-Pro), we conducted a rigorous Inter-Annotator Agreement (IAA) analysis. We compared the AI judge's 10-point scoring logic against native expert evaluations. As shown in Table 5, the high strict and tolerance-based (± 1 point) agreement rates confirm that the automated scoring remains highly consistent with native expert intuition.

Task Type	Cognitive Layer	Strict Agreement	Agreement (w/ ± 1 tol.)	Cohen's Kappa (κ)
Choice Questions	Layers 1, 2, 3	92.5%	-	0.84
Taboo Tasks	Layer 2 (Situational)	76.4%	88.2%	0.77
Special Tasks (Riddles, etc.)	Layer 3 (Values)	72.1%	85.4%	0.75

Table 5: Inter-Annotator Agreement (IAA) metrics between the primary AI Judge (Gemini-2.5-Pro) and native Mongolian experts.

C Qualitative Case Studies

To contextualize the statistical performance gaps, this section examines representative failure modes alongside successful retrievals. The qualitative analysis isolates instances where models impose cross-cultural normative priors (Etic bias) or exhibit failures in contextual mapping within situational pragmatics. Note that due to traditional

Mongolian script rendering constraints in the current LaTeX environment, only English translations of the case content are provided. Furthermore, the perspective labels (e.g., Emic, Etic) assigned to options in the first case were not visible to the models during inference; they are included here solely to facilitate reader interpretation.

C.1 Case 1: Epistemological Reduction (The "Tourist Perspective")

The first failure mode involves the reduction of emic ontology to etic functionalism. In Mongolian culture, specific customs are rooted in deep socio-spiritual frameworks rather than secular utility.

Task: Values Layer (Ordos Culture)
Question: Why is it strictly taboo to leave the front flap of the robe unbuttoned?
Options:
[A] (Emic, Ground Truth): By custom, only those in mourning or bereaved leave the flap open; for ordinary people, it constitutes a bad omen.
[B] (Etic): Leaving it open risks catching a cold from the wind.
[C] (Etic): It should remain buttoned to prevent the buttons from falling off.
[D] (Shallow Emic): It is considered shameful as it appears as if one is showing off their body to others.

Correct Model (Gemini-3-Pro): Option A.
Incorrect Model (GPT-5.2): Option B.
Incorrect Model (Claude-Sonnet): Option C.

Figure 5: Comparison between emic alignment and etic functionalist logic in values-based reasoning.

As illustrated in Figure 5, while models like Gemini-3-Pro successfully align with the emic perspective, others such as GPT-5.2 and Claude reject the authentic spiritual explanation (Option A), which links the unbuttoned flap to funerary states. Instead, they prioritize functional, health-centric (Option B) or material-preservation (Option C) rationales. Drawing on **?**, this behavior indicates that certain models compress "thick" indigenous beliefs into "thin," universally palatable hygiene habits. Empirically, this constitutes an epistemological reduction. Rather than making an absolute normative claim, we frame this as a potential risk: such technical biases can be detrimental to cultural fidelity for preservation-oriented applications, as they inadvertently dilute the original ontological weight of marginalized practices.

C.2 Case 2: Failures in Contextual Mapping (Proverbs)

The second failure mode emerges in situational pragmatics, where models must map complex social behaviors to culturally specific idioms. This task tests whether a model possesses deep procedural knowledge of cultural metaphors or relies on surface-level logical associations.

Task: Proverbs (Values Layer)
Scenario: A woman harshly criticizes her neighbor's child's faults to outsiders, yet remains silent or defensive when her own child causes significant trouble.
Options:
[A]: Not seeing the camel on one's own head, but seeing the hair on another's head.
[B]: Other people's food is tasty; other people's wives are beautiful.
[C]: Many words lead to error; a long thread leads to a break.
[D] (Ground Truth): A brooding bird when speaking of itself; a hungry tiger when speaking of others.

Correct Model (Claude-Sonnet): Option D.
Incorrect Model (GPT-5.2): Option A.
Incorrect Model (Qwen3-max): Option C.

Figure 6: Contrast between deep cultural alignment and surface-level or irrelevant mapping in proverbs.

Figure 6 demonstrates the nuances of cultural alignment. While Claude-Sonnet identifies the precise idiom (Option D) that captures the behavioral contrast between protective self-treatment and aggressive external criticism, other models falter. GPT-5.2 defaults to a more generalized "blindness to self" metaphor (Option A), which, while logically adjacent, lacks the specific situational precision of the brooding bird versus hungry tiger imagery. More critically, Qwen3-max selects Option C, a proverb regarding verbal discretion that is contextually irrelevant to the scenario's theme of double standards. This confirms that even models with high linguistic fluency often lack the cultural depth required to distinguish between general logic and specific indigenous semantic structures.

C.3 Case 3: Metaphorical Hallucinations (Riddles)

Riddles represent a profound test of a model's metaphorical reasoning and cultural abstraction in low-resource environments.

As shown in Figure 7, Gemini-3-Pro and GPT-5.2 successfully decode the anthropomorphic mapping of body parts to furniture components. In contrast, Claude-Sonnet exhibits confident halluci-

Task: Riddles (Values Layer)
Riddle: Has a face but no mouth, has legs but no hands, has four legs but cannot move.
Ground Truth: A Table.

Correct Model (Gemini-3-Pro & GPT-5.2): Answered "Table". Both correctly mapped the "face" to the tabletop and the "legs" to the supports.
Incorrect Model (Claude-Sonnet 4.5): Answered "Green Temple". The model attempted to justify this through circular logic, stating a temple metaphorically has a face and legs, demonstrating severe hallucination.
Incorrect Model (DeepSeek-V3.2): Answered "Moving". The model extracted a verb from the prompt instead of reasoning about a physical entity, failing to abstract the object entirely.

Figure 7: Model performances on Mongolian metaphorical abstractions.

nation, proposing a "Green Temple" and forcing a nonsensical justification. DeepSeek-V3.2 collapses completely, extracting a literal word ("moving") from the negation in the prompt rather than synthesizing the clues. This highlights the fragility of LLMs when navigating abstract indigenous metaphors, where literal translation falls short of capturing the intended ontological mapping.

C.4 Case 4: Normative Reasoning in Situational Pragmatics (Taboos)

The Taboo task evaluates a model's capability to identify specific normative violations within daily social scenarios and provide culturally grounded corrections.

Task: Taboos (Situational Layer)
Scenario: A guest enters the house of a postpartum woman carrying a whip, and the mother uses a knife to eat meat.
Ground Truth Error Points: 1. Entering with a whip. 2. Postpartum mother using a knife.

Correct Model (Gemini-3-Pro): Identifies both errors. Reasoning: "A whip symbolizes aggression and disturbs the peace. For the mother, touching a sharp object is believed to cause her breast milk to dry up."
Incorrect Model (Qwen3-max): Identifies the errors but provides flawed corrections. For the knife, it hallucinates a nonsensical cultural remedy: "Use Genghis Khan's measure to eat meat."

Figure 8: Identifying taboos and providing authentic cultural corrections.

Gemini-3-Pro demonstrates profound cultural reasoning, not only identifying the physical violations but accurately attributing them to specific indigenous beliefs (e.g., sharp objects causing milk to dry up). Qwen3-max, while capable of identifying the surface-level violations, hallucinates bizarre cultural corrections ("Genghis Khan's measure"). This reveals a critical gap: recognizing a taboo grammatically does not equate to understanding its authentic practical resolution within the cultural matrix.

C.5 Case 5: Ritualistic Sequencing (Benedictions)

Benedictions (Iruge-l) are highly structured poetic praises. Reconstructing them requires implicit knowledge of traditional Mongolian alliteration (head rhyme) and logical semantic progression.

Task: Benedictions (Values Layer)
Scenario: Reconstruct the shuffled lines of the "Hunter's Blessing".
Shuffled Lines:
1. Blessing the masses.
2. Preparing the bow .
3. Blessing the family .
4. Mounting the wind-swift horse.
Ground Truth Sequence: $2 \rightarrow 4 \rightarrow 3 \rightarrow 1$.

Evaluation Logic: The sequence follows a strict phonetic head rhyme pairing ("S" for lines 2 and 4, "T" for lines 3 and 1) and a semantic expansion from the individual hunter's preparation to the broader society (Family $\rightarrow$ Masses).

Figure 9: Reconstruction logic of traditional Mongolian praise poetry.

The Hunter's Blessing (Figure 9) illustrates how structural linguistics and cultural philosophy intertwine. Correct sequencing relies heavily on recognizing the phonetic alliteration pairs and the conceptual hierarchy of nomadic blessings (progressing from the micro-level of mounting a horse to the macro-level of societal peace). Models that fail this task invariably ignore the rigid poetic structures of Traditional Mongolian or misinterpret the "small-to-large" progression inherent in indigenous well-wishing.

D Detailed Error Taxonomy

In Section 7.3 of the main text, we discussed the phenomenon of "Authoritative Hallucinations" and the systematic failure of models in handling low-resource cultural metaphors. To provide a transparent and comprehensive view of these failure modes, we systematically categorized the qualitative evaluation logs across all evaluated frontier models.

Scope of Analysis: In this appendix, we focus

our detailed taxonomy specifically on the **Traditional Mongolian** script. This decision is informed by our primary findings, which indicate that models exhibit a significantly higher frequency and more complex variety of errors in the Traditional script compared to the Cyrillic script. By analyzing the "worst-case" scenarios in the Traditional script, we can better isolate the structural and cultural blind spots of these models. Furthermore, since the qualitative justifications provided by our two AI judges (Gemini-2.5-Pro and GPT-5.2) were found to be nearly identical in reasoning and classification, we present only the adjudication results from the primary judge, Gemini-2.5-Pro, to ensure clarity and avoid redundancy.

D.1 Error Taxonomy Framework

Based on a rigorous semantic analysis of the evaluation rubrics and model outputs, we identified the following dominant error categories:

- **Cultural Gap:** The model correctly parses the grammar but lacks the specific indigenous knowledge required to resolve the metaphor or identify the deep cultural reason. It defaults to general, often secular, logic.
- **Forced Logic & Hallucination:** The model confidently generates an incorrect answer and invents a highly convoluted, illogical, or historically inaccurate justification to support it (e.g., the "Green Temple" case in Appendix C).
- **Reasonable Error (Riddles Only):** The model's answer is technically incorrect according to the standard Mongolian folklore corpus, but the logic provided is highly plausible and adjacent to the ground truth.
- **Irrelevant / Non-sequitur (Riddles Only):** The model's response is entirely disconnected from the prompt, or it extracts a literal word from the instruction rather than synthesizing a coherent entity.
- **Identification Failure (Taboos Only):** The model completely fails to detect the normative violation embedded in the scenario or identifies an innocent action as the taboo.
- **Correction Error (Taboos Only):** The model successfully identifies the taboo but provides an inappropriate, impractical, or culturally misaligned behavioral correction.
- **Weak Reasoning (Taboos Only):** The model identifies the taboo and the correction, but struggles to articulate the authentic cultural or spiritual reason behind the practice.

D.2 Quantitative Error Distribution

Table 6 presents the distribution of these error types for the highly abstract Riddles task, revealing the prevalence of "Forced Logic." Table 7 illustrates the error distribution for the Taboos task.

Model	Cultural Gap	Forced Logic	Reasonable Error	Irrelevant
GPT-5.2	267	318	42	88
Gemini-3-Pro-Preview	220	38	63	10
Claude-Sonnet-4.5	262	379	38	94
DeepSeek-V3.2	117	402	14	118
Qwen3-Max	245	311	21	92

Table 6: Error type distribution for the Riddles Task. Counts are derived from the Gemini-2.5-Pro evaluation logs.

Model	Cultural Gap	Identification	Correction	Hallucination	Weak Reason
GPT-5.2	190	105	99	26	24
Gemini-3-Pro-Preview	95	68	2	4	1
Claude-Sonnet-4.5	320	198	273	30	87
DeepSeek-V3.2	138	165	312	14	30
Qwen3-Max	247	189	241	37	54

Table 7: Error type distribution for the Taboos Task. Demonstrates a high frequency of correction failures.