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

Anonymous ACL submission

Abstract

While Large Language Models (LLMs) have achieved impressive linguistic fluency in low-resource languages, their ability to grasp deep cultural nuances remains under-explored. This paper introduces **MonCulture-Eval**, a comprehensive benchmark designed to evaluate the Cultural Intelligence of LLMs in Mongolian across two distinct writing systems (Traditional and Cyrillic) and three major regional sub-cultures (Alxa, Ordos, and Horqin). Constructed via an "Indigenous-First" approach, the benchmark is structured into a three-layer cognitive framework—Factual, Situational, and Values—complemented by specialized tasks including Riddles, Taboos, and Proverbs. Evaluating state-of-the-art models (GPT-5.2, Gemini-3-pro, DeepSeek-v3.2, and Claude-Sonnet-4-5) reveals significant limitations in current systems. First, we identify a severe "Script Gap," where most models perform significantly worse in Traditional Mongolian, effectively cutting them off from deep cultural archives. Second, qualitative analysis uncovers a prevalent **"Tourist Perspective"** (Etic Bias): models frequently sanitize spiritual rituals into secular safety regulations and hallucinate functional rationales for symbolic taboos. Notably, Gemini-3-pro demonstrates exceptional "Emic" alignment in the Values layer, while others exhibit a sharp "Cognitive Depth Drop-off." Our findings underscore that translation capability does not equate to cultural understanding, highlighting the urgent need for culturally-aware alignment strategies in inclusive AI.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable proficiency in multilingual tasks, often achieving near-human performance in translation and cross-lingual retrieval. However, linguistic fluency does not equate to Cultural Intelligence. A model may correctly translate a sentence grammatically while completely failing to grasp the underlying social norms, historical context, or value systems embedded in that language. This gap is particularly pronounced in low-resource, non-Western languages like Mongolian, where cultural nuances are often "High-Context" and deeply tied to nomadic traditions (Hall, 1976).

The challenge of evaluating Mongolian cultural capabilities is compounded by two unique factors. First, the language functions under a complex digraphia situation: it uses both the Traditional Mongolian script (vertical) and the Cyrillic script. Second, Mongolian culture is not monolithic; it consists of diverse regional variations (e.g., Alxa, Ordos, Horqin), each with specific customs and taboos that generic training data often overlook (Hershcovich et al., 2022). Current benchmarks, primarily derived from translated English datasets like MMLU (Hendrycks et al., 2021), fail to capture this depth. They tend to test "Encyclopedia knowledge" (dates, names) rather than "Deep Culture" (values, metaphors, and implicit social rules).

To bridge this gap, we introduce **MonCulture-Eval**, a hierarchical, native-expert-curated benchmark designed to assess the deep cultural alignment of LLMs. Unlike previous efforts, our benchmark is constructed "Indigenous-First," with questions designed directly by native speakers to reflect the Emic (insider) perspective. MonCulture-Eval proposes a novel Three-Layer Cognitive Framework, progressing from Factual Knowledge (Layer 1) to Situational Appropriateness (Layer 2), and finally to Value Alignment (Layer 3). This hierarchical design allows us to distinguish between models that merely memorize facts and those that grasp the underlying spiritual and social logic of the culture. A detailed formalization of this framework and its definitions is provided in Section 3.2.

Furthermore, we incorporate specialized cultural tasks—Riddles, Taboos, Proverbs, and Benedictions—to evaluate the models' capacity for metaphorical reasoning and normative correc-

tion. We evaluated state-of-the-art models across both writing systems. Our findings reveal a critical "Script Gap," where models perform significantly worse in the Traditional script. More importantly, qualitative analysis uncovers a prevalent "Tourist Perspective": models often sanitize deep spiritual rituals into secular safety regulations.

In summary, this paper makes the following contributions:

- We release MonCulture-Eval, the first comprehensive cultural benchmark for Mongolian covering dual scripts and three regional subcultures.
- We propose a hierarchical evaluation framework that differentiates between surface-level facts and deep-level value alignment.
- We provide extensive empirical evidence of the "Etic Bias" in current LLMs, highlighting the need for culturally-aware alignment strategies.

2 Related Work

Our research is situated at the intersection of three evolving domains in NLP: cultural evaluation benchmarks, low-resource language processing, and pluralistic value alignment.

2.1 Cultural Benchmarks: From Facts to Norms

The evaluation of Large Language Models (LLMs) has transitioned from measuring general linguistic fluency to assessing specific knowledge domains. Global benchmarks like MMLU (Hendrycks et al., 2021) serve as the gold standard for encyclopedic knowledge but are widely criticized for their Anglocentric bias. To counter this, a wave of language-specific benchmarks has emerged, including C-Eval (Huang et al., 2023) and CMMLU (Li et al., 2024) for Chinese, JGLUE (Kurihara et al., 2022) for Japanese, and VNHSGE (Nguyen et al., 2023) for Vietnamese. However, most existing benchmarks prioritize declarative knowledge (e.g., history, STEM subjects) over procedural cultural knowledge (e.g., social norms and taboos). Recent efforts like IndoMMLU have begun to address regional specifics, but few frameworks explicitly test the "Deep Culture" (Hall, 1976) that governs daily interactions in high-context societies. Our work fills this gap by introducing a hierarchical framework that specifically probes the Emic (internal) logic of cultural behaviors (Geertz, 1973), moving beyond simple fact retrieval.

2.2 Low-Resource Languages and The Script Gap

Mongolian presents a unique challenge due to its agglutinative morphology and diglossic nature. While significant progress has been made in low-resource machine translation (Magueresse et al., 2020), semantic reasoning in such languages remains under-explored. A critical, often overlooked issue is the impact of tokenization on model performance in non-Latin scripts. Rust et al. (2021) demonstrated that suboptimal tokenizers can severely hamper performance in monolingual tasks.This is particularly relevant for Traditional Mongolian, where the vertical script and complex visual shaping pose significant challenges for standard BPE (Byte Pair Encoding) tokenizers trained primarily on horizontal, Latin-heavy corpora (Petrov, 2016). MonCulture-Eval quantifies this "Script Gap," providing empirical evidence of how tokenizer-induced disparities translate into cultural reasoning failures, echoing broader concerns about systematic inequalities in language technology (Blasi et al., 2022).

2.3 Pluralistic Alignment and WEIRD Bias

Traditional alignment techniques (RLHF) often optimize for a generic notion of "helpfulness" and "safety," which implicitly encodes Western, Educated, Industrialized, Rich, and Democratic (WEIRD) values (Bender et al., 2021). Durmus et al. (2023) highlighted that LLMs frequently reflect the subjective opinions of Western populations while marginalizing global voices. Recent calls for **Pluralistic Alignment** (Sorensen et al., 2024) emphasize the need for models to adapt to diverse normative systems, ensuring that cultural nuances are preserved rather than overwritten (Masoud et al., 2023). While datasets like Social Chemistry (Forbes et al., 2020) and CultureBank (Ramezani et al., 2023) have attempted to model social norms, they often lack the granularity of indigenous taboos. By focusing on specific tribal variations (Alxa, Ordos, Horqin), our work contributes to the "Cultural Alignment" agenda by testing whether models can suppress their default Western safety filters (the "Tourist Perspective") to embrace local ontologies.

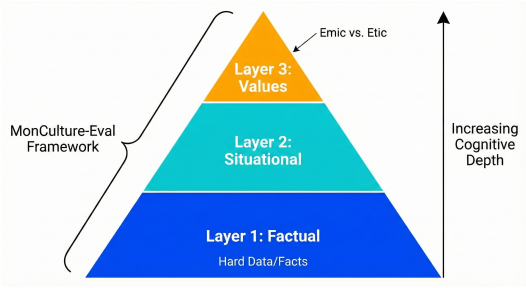


Figure 1: **The MonCulture-Eval Cognitive Framework.** The hierarchy progresses from Factual Knowledge (Layer 1) to Situational Appropriateness (Layer 2) and Deep Value Alignment (Layer 3).

3 Dataset Construction

To comprehensively evaluate cultural capabilities, we propose a framework rooted in cultural anthropology, constructed in collaboration with native experts.

3.1 Design Philosophy and Data Sourcing

Unlike generic multilingual benchmarks that rely on translating English datasets (e.g., MMLU), MonCulture-Eval is constructed from the ground up with a "Native-Centric" philosophy. Our data collection followed a strict "Indigenous-First" approach. We collaborated with native Mongolian domain experts to design questions based on authentic local materials, including oral traditions and historical records. This ensures the benchmark captures the nuances of the "internal perspective" often lost in translation-based datasets.

3.2 The Three-Layer Cognitive Framework

To move beyond surface-level language proficiency, we propose a novel three-tier evaluation framework (See Figure 1):

- **Layer 1: Factual Knowledge (Facts).** Assesses the model's retention of "hard knowledge," such as historical events, geographical landmarks, and the specific components of a Ger. This tests memory and information retrieval.
- **Layer 2: Situational Pragmatics (Situations).** Evaluates whether the model can "act" like a culturally immersed individual. We present daily social scenarios (e.g., etiquette during festivals) to test the model's ability to navigate social norms and behavioral expectations correctly.
- **Layer 3: Value Alignment (Values).** The most complex level, designed to probe the "underlying logic" of cultural behaviors. Instead of simple binary choices, we employ a "Best Explanation" format. Distractors include correct "scientific/external" (Etic) explanations, while the target answer requires the specific "cultural/internal" (Emic) reasoning held by the community. This distinguishes models that merely know what happens from those that understand why it happens from a Mongolian perspective.

3.3 Key Characteristics of MonCulture-Eval

Based on our native-centric design philosophy, MonCulture-Eval is defined by three distinct characteristics:

1. Authenticity via Native Construction. As mentioned, generic benchmarks often suffer from translation artifacts and Western bias. MonCulture-Eval avoids this by sourcing data directly from indigenous experts. All questions are authored based on authentic materials, ensuring that the content reflects the lived reality of the community rather than a foreigner's observation.

2. Intracultural Diversity (Regional Granularity). Nomadic culture is not monolithic. To avoid stereotyping, we explicitly incorporate regional variations from three representative Mongol tribes: Alxa, Ordos, and Horqin.

- **Alxa (The West):** Focuses on customs and survival wisdom specific to arid, desert-adjacent nomadic life.
- **Ordos (The Center):** Renowned for preserving traditional court culture and intricate ritual systems (e.g., Genghis Khan sacrifices).
- **Horqin (The East):** Represents a distinct cultural branch with agricultural-nomadic transitional features.

This design forces models to distinguish between subtle dialectal and customary differences, penalizing models that overgeneralize specific tribal customs to the entire ethnic group.

3. Dual-Script Parallelism. To evaluate LLMs' adaptability to the digraphia situation of the Mongolian language, we established a strict parallel corpus between Traditional Mongolian and Cyrillic Mongolian. The original dataset was created

in Traditional Mongolian. To generate the Cyrillic counterpart, we employed a rigorous hybrid pipeline: (1) initial batch translation via LLMs, (2) automated verification using the Onon platform, and (3) final human quality assurance. This allows us to isolate the "Script Gap," assessing whether performance drops are due to knowledge deficits or tokenization issues in the vertical script.

3.4 Domain-Specific Cultural Tasks

We designed four specialized tasks to challenge specific cognitive capabilities beyond factual recall:

- **Riddles (Metaphorical Reasoning):** Mongolian riddles often employ complex metaphors rooted in nomadic life. This task assesses whether models can bridge the semantic gap between a cryptic description and its referent.
- **Taboos (Normative Reasoning):** The model acts as a cultural critic. It must identify behaviors that violate cultural taboos in a given text and propose appropriate behavioral modifications.
- **Proverbs (Pragmatic Application):** Tests the model's ability to select the appropriate proverb for a specific communicative context, distinguishing between surface-level similarity and deep semantic fit.
- **Benedictions (Structural Reconstruction):** A text reordering task where the model must reconstruct the correct sequence of a ceremonial benediction, testing its grasp of literary structure and rhyme schemes.

3.5 Dataset Statistics

Table 1 summarizes the distribution of questions across the three cognitive levels and different categories.

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.

4 Experimental Setup

4.1 Models Evaluated

We selected four models representing the frontier of multilingual capabilities: **GPT-5.2** (gpt-5.2-2025-12-11), **Gemini-3-Pro-Preview**, **Claude-Sonnet-4-5** (20250929), and **DeepSeek-v3.2** (DeepSeek-AI, 2024).

4.2 Prompt Engineering Strategy

We employed a systematic prompting strategy tailored to the cognitive nature of each task (See Appendix B for full templates).

Persona Adoption: For all tasks, we explicitly instructed models to adopt specific personas (e.g., "Professional Archivist" or "Folklore Expert").

Two-Stage Reasoning: For complex tasks like Riddles, we utilized a two-stage multi-turn dialogue: (1) Answer Generation, (2) Rationalization (Explain reasoning).

Strict JSON Output: For Taboo Correction and Benedictions, we enforced a strict JSON output schema to facilitate automated parsing.

4.3 Human Verification

To mitigate judge bias, native experts sampled 10% of evaluations. The agreement rate was high, validating our automated pipeline.

4.4 Evaluation Metrics

Beyond standard accuracy, we introduce a novel metric to quantify the stability of cultural knowledge across the digraphia environment:

Script Alignment Consistency (SAC): To penalize models that rely on superficial pattern matching in one script (often Cyrillic due to higher resource availability) while failing in the other, SAC measures the proportion of samples where a model answers correctly 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} are the model's predictions in Traditional and Cyrillic Mongolian respectively, and y_i^* is the ground truth. A high SAC score indicates that the model possesses robust, script-independent cultural concepts ("Ontological Alignment"), whereas a low SAC alongside high Cyrillic

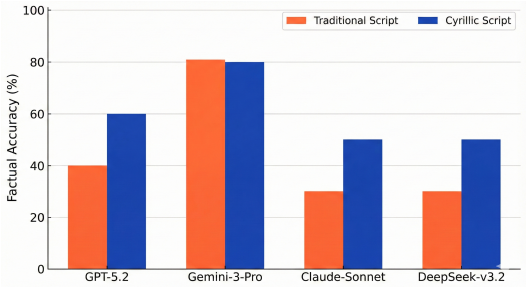


Figure 2: The Script Gap. Comparison of Factual Accuracy across Traditional and Cyrillic scripts. Most models show a sharp decline in the Traditional script, while Gemini-3-pro maintains consistency.

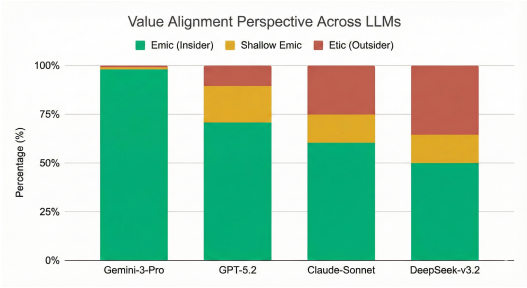


Figure 3: **Value Perspective Distribution.** Gemini-3-pro (Green) shows near-perfect Emic alignment, while DeepSeek and Claude exhibit significant Etic (Red) components.

accuracy suggests a "Resource Gap" or "Translation Shortcut."

5 Results and Analysis

5.1 The Script Gap and Alignment Consistency

A major finding of our study is the significant performance disparity between Traditional Mongolian and Cyrillic Mongolian across most models. As shown in Figure 2, we compared the factual accuracy (Layer 1) as a baseline.

Models like **DeepSeek-v3.2** and **Claude-Sonnet** exhibit a severe "Resource Gap." For instance, DeepSeek's factual accuracy drops from 48.35% in Cyrillic to 28.93% in Traditional Mongolian. This large disparity inherently limits their **SAC (Script Alignment Consistency)** score.

Model	Trad. Acc	Cyr. Acc	Script Gap	Est. Max SAC
DeepSeek-v3.2	28.93%	48.35%	-19.42%	$\leq$ 28.93%
Claude-Sonnet	33.33%	56.47%	-23.14%	$\leq$ 33.33%
GPT-5.2	40.36%	60.47%	-20.11%	$\leq$ 40.36%
Gemini-3-pro	**79.48%**	**76.86%**	**+2.62%**	$\leq$ **76.86%**

Table 2: **Script Gap & SAC Bounds.** Large negative gaps in most models fundamentally limit their Script Alignment Consistency (SAC), whereas Gemini's inverse gap suggests robust cross-script alignment.

Qualitative inspection reveals that for many questions correctly answered in Cyrillic, these models hallucinate or fail in Traditional script, suggesting their knowledge is not ontologically grounded but rather script-bound. In contrast, **Gemini-3-pro** demonstrates an "Alignment Exception," achieving comparable high performance across both scripts. Its consistently high accuracy in both domains implies a high SAC, indicating it has likely mapped both scripts to a shared, language-agnostic semantic space, effectively overcoming the digraphia barrier.

5.2 Regional Performance Disparity

Mongolian culture is not monolithic. We further analyzed model performance across the three target regions: **Alxa**, **Ordos**, and **Horqin**. Table 3 presents the accuracy breakdown in the Traditional script.

Model	Alxa Culture	Ordos Culture	Horqin Culture
Gemini-3-pro	**96.20%**	**96.72%**	**100.00%**
GPT-5.2	60.76%	75.41%	63.49%
Claude-Sonnet	54.43%	58.20%	63.49%
DeepSeek-v3.2	53.16%	45.90%	57.14%

Table 3: **Values Layer Accuracy by Region (Traditional Script).** Models perform unevenly, with GPT-5.2 favoring Ordos culture, likely due to tourism data bias.

5.3 Cognitive Depth: From Facts to Values

Moving beyond factual recall, we analyzed how models perform as the cultural cognitive load increases to Layer 3 (Values). This layer tests whether a model chooses the "Emic" (insider) explanation over "Etic" (outsider) descriptions. Table 4 and Figure 3 illustrate the distribution of perspectives.

Gemini-3-pro achieves near-perfect alignment (98.99% Emic), indicating it has essentially "internalized" the Mongolian value system. In contrast, DeepSeek-v3.2 falls into the "Etic Trap" in 32.55% of cases, rationalizing cultural behaviors through a scientific or external lens.

5.4 Performance on Specialized Tasks

5.4.1 Riddles: The Metaphorical Boundary

Riddles proved to be the hardest task. In Traditional Mongolian, DeepSeek (0.0%) and Claude (0.57%) completely failed to identify riddle answers. Our

Model	Accuracy (Emic)	Etic Bias	Shallow Emic
Gemini-3-pro	**98.99%**	0.34%	0.67%
GPT-5.2	68.12%	20.13%	11.74%
Claude-Sonnet	58.72%	26.17%	15.10%
DeepSeek-v3.2	51.34%	**32.55%**	16.11%

Table 4: **Perspective Distribution on Values Layer.** "Etic Bias" indicates the percentage of answers where the model chose an outsider/functional explanation.

detailed error analysis (refer to **Appendix A**, Table 7) reveals that the majority of errors (76.8% for DeepSeek) are classified as "Forced Logic." This confirms that when models lack specific cultural mappings, they revert to hallucinating nonsensical connections rather than admitting ignorance.

5.4.2 Taboos: Norm Violation

In the Taboo detection task, Gemini-3-pro demonstrated superior social awareness with a 91.1% pass rate. However, GPT-5.2 and DeepSeek struggled significantly with the "Correction" step. As detailed in **Appendix A** (Table 6), they suffer from high rates of Correction Error and Etic Sanitization, often identifying the error but failing to propose the culturally prescribed remedy.

5.4.3 Benedictions: The Ritual Coherence Test

While models may retrieve isolated cultural facts, the Benedictions task evaluates their ability to reconstruct the linear structure of ceremonial texts. This task, formulated as a sequence reordering problem, tests Syntactic and Ritual Coherence—the understanding that in Mongolian oral tradition, meaning is dictated not just by keywords, but by a rigid, prosodic, and logical flow.

Table 5 highlights a dramatic collapse in performance for most models. **DeepSeek-v3.2**, **GPT-5.2**, and **Claude-Sonnet** struggle significantly, with accuracies hovering below 21% in Cyrillic and dropping to single digits (e.g., DeepSeek's 5.74%) in Traditional Mongolian. This suggests that these models process cultural text as a "Bag-of-Words," recognizing individual auspicious terms but failing to grasp the underlying long-range dependencies required to form a coherent prayer. In sharp contrast, **Gemini-3-pro** demonstrates a qualitative leap, achieving **41.80%** accuracy in Traditional Mongolian and **35.25%** in Cyrillic. Notably, Gemini performs better in the Traditional script, further evidencing its robust cross-script alignment. This indicates that Gemini has internalized not just the semantics of cultural entities, but the syntax of the ritual itself.

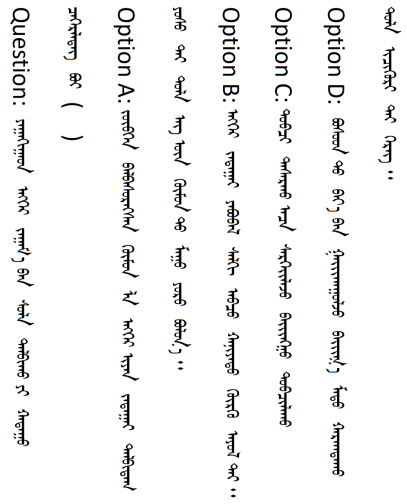
Model	Traditional (%)	Cyrillic (%)
Gemini-3-pro	**41.80**	**35.25**
GPT-5.2	9.02	20.49
Claude-Sonnet	11.48	18.85
DeepSeek-v3.2	5.74	15.57

Table 5: Accuracy on Benediction Reordering. Most models fail to reconstruct ceremonial sequences, exposing a lack of syntactic coherence. Gemini-3-pro is the only model to maintain significant capability.

5.5 Qualitative Analysis: Case Studies

To deeply understand these failures, we analyze specific responses using the original Traditional Mongolian script to highlight the visual and semantic challenges.

Case Study 1: The "Etic" Trap (Values). In Ordos culture, the state of the robe flap symbolizes ritual status.



Translation: Why is it strictly forbidden to leave the Engger loose?
Correct (Gemini): Option A - It is a custom for the widowed/mourning; bad omen for ordinary people. (Emic/Ritual)
GPT-5.2 Failure: Selected Option B - Danger of catching a cold/flu. (Etic/Medical Bias)
Claude Failure: Selected Option C - Buttons might break. (Etic/Pragmatic Bias)

Analysis: Models sanitize the spiritual taboo into a hygiene or maintenance issue, exhibiting a classic "Tourist Perspective."

Case Study 2: Metaphorical Blindness (Riddles). Models often hallucinate when missing the cultural link.

Riddle: [illegible]

Translation: "Outside the listener, a golden colt."

GPT-5.2 Output: [illegible] (Blue Sky)

Target: [illegible] (earrings).

Reasoning: Hallucinates that "Listener" is a poetic name for God and "Colt" is the sound of the sky.

Explanation:

...[illegible]...

Translation:"receiver" is a poetic name for the Sky... "Golden Colt" refers to the sound of the sky...

Analysis: This represents "Forced Logic"—fabricating nonsensical narratives to justify a random guess.

Case Study 3: The Blue Saddle. This case demonstrates a breakdown in commonsense reasoning.

Riddle: [illegible]

(Trans: Blue saddle pad in the countryside.)

Claude Output: [illegible] Milk.

Target: [illegible] (Lake).

Explanation:

[illegible]...

(Trans: Milk becomes a blue saddle pad in the countryside and blue iron in the city.)

Case Study 4: Sympathetic Magic (Taboos). Models fail to retrieve specific folk beliefs (Sympathetic Magic).

Scenario:

[illegible]

(Trans: A pregnant woman steps over a long rope.)

Cultural Belief: It is believed that this causes the umbilical cord to wrap around the baby's neck.

Claude Output:

[illegible]

(Trans: Correction: Hold the pregnant woman with two hands.)

Analysis: Instead of retrieving the specific taboo about the umbilical cord, the model applies a generic "be polite" filter, offering a socially positive but culturally irrelevant correction.

Case Study 5: Semantic Misalignment (Proverbs). This case illustrates a specific failure mode where the model understands the semantic field (speaking) but fails to distinguish between a behavioral criticism and a philosophical principle.

Scenario:

[illegible]

(Trans: Once a word is spoken without thought, it cannot be taken back, so one must be careful.)

Correct Proverb (D):

[illegible]

(Trans: If a horse gets loose, it can be caught; if the mouth gets loose [words spoken], it cannot be caught.)

Claude Output (B):

[illegible]

(Trans: To have no stopper in one's mouth [i.e., a blabbermouth].)

Model Explanation: "This proverb means if words leave the mouth, there is no way to take them back..."

Analysis: This exhibits a Reasoning-Choice Disconnect. Claude's explanation correctly identifies the concept of "irreversibility" (matching Option D), yet it selects Option B, which is a derogatory idiom describing a person's character rather than the nature of speech. The model conflates the subject (the speaker vs. the spoken word).

6 Discussion

Our investigation into the cultural capabilities of LLMs reveals a complex landscape where linguistic competence often masquerades as cultural intelligence. The discrepancies between model performance on factual recall versus value alignment expose fundamental limitations in how current architectures process low-resource, high-context cultures.

6.1 The Cognitive Depth Drop-off: From Encyclopedia to Agent

A consistent pattern across models (excluding Gemini-3-pro) is the distinct "Cognitive Depth Drop-off." As tasks progress from the Factual

Layer (L1) to the Situational (L2) and Values Layers (L3), performance degrades. For instance, DeepSeek-v3.2 manages a respectable performance on some Cyrillic factual tasks but collapses to 51.34% on the Values layer in Traditional Mongolian. This trend suggests that most LLMs operate as "Cultural Encyclopedias"—efficient at retrieving declarative knowledge (names, dates, object descriptions) but failing as "Cultural Agents" capable of simulating the procedural reasoning of a community member. This failure is most evident in the Benediction task, where models like DeepSeek (5.74% accuracy in Traditional script) fail to sequence culturally meaningful segments, treating the task as a bag-of-words problem rather than a ritual reconstruction.

6.2 The Dominance of the "Tourist" Perspective

Perhaps the most critical finding is the prevalence of the "Etic Bias." As quantified in Table 4, models like DeepSeek and Claude exhibit an Etic bias in over 26%-32% of their responses. This is not merely an error of fact but an error of stance. When models sanitize spiritual taboos into issues of hygiene or safety, they are effectively colonizing the cultural space with a Western, secular rationality. They transform "Thick" cultural concepts (Geertz, 1973) involving death, spirits, and omen into "Thin," universal rules about catching a cold or breaking buttons. This "sanitization" renders the model's output "safe" by standard safety guidelines but fundamentally "unsafe" for cultural preservation, as it actively erases the indigenous ontology.

6.3 Hallucination as "Confident Nonsense"

The Riddle task exposed a dangerous tendency towards "Forced Logic." With DeepSeek generating 402 instances of this error type, we see that when models lack a specific cultural metaphor (e.g., "Receiver" = Sky), they do not hallucinate randomly; they hallucinate logically. They fabricate plausible-sounding but culturally incorrect causal links to justify their guesses. In a cultural preservation context, this "confident nonsense" is more harmful than a refusal to answer, as it pollutes the knowledge base with fabricated folklore that sounds authoritative to a non-expert user.

6.4 The Script Gap as a Resource and Cultural Gatekeeper

The performance disparity between Traditional and Cyrillic Mongolian is not just a technical issue of tokenization; it is a cultural gatekeeper. Traditional Mongolian is the primary vessel for pre-modern archives, religious texts, and ceremonial language. Models that fail in this script (DeepSeek, Claude, and to a lesser extent GPT-5.2) are effectively locked out of the "Deep Culture" archives. Their higher performance in Cyrillic correlates with the availability of translated, often secularized, modern data. Gemini-3-pro's exceptional SAC (Script Alignment Consistency) suggests that true cultural alignment requires a model to transcend script dependencies, mapping both writing systems to a shared semantic and cultural embedding space.

7 Conclusion

In this work, we introduced MonCulture-Eval, a comprehensive benchmark designed to assess the cultural capabilities of Large Language Models in Mongolian. By distinguishing between Factual, Situational, and Value-based knowledge, we provide a more granular view of AI cultural intelligence than previously possible.

Our experiments yield three critical insights. First, the Script Gap remains a formidable barrier. For most models, the Traditional Mongolian script acts as a firewall, blocking access to the deeper, pre-modern layers of the culture. Second, the "Tourist Perspective" is the default mode for current LLMs. Without specific alignment, models tend to view indigenous cultures through a secular, functional lens, stripping away the spiritual dimensions that define the "Emic" perspective. Third, Alignment is Possible. The exceptional performance of Gemini-3-pro demonstrates that high-resource models can achieve deep cultural alignment, respecting indigenous value systems and maintaining consistency across scripts. This suggests that the failure of other models is not an inevitability of the architecture but a result of data curation and alignment priorities.

Future work on inclusive AI must prioritize "Pluralistic Alignment"—moving beyond translation to ensure that models not only speak the language of a community but also respect its underlying reality.

8 Limitations

Dialectal Coverage: While our benchmark covers Alxa, Ordos, and Horqin cultures, the Mon-

golosphere includes numerous other groups such as Chahar, Buryat, and Oirat. The cultural norms tested here should not be generalized to all Mongolian speakers.

Static vs. Dynamic Culture: Our benchmark treats cultural norms as relatively static concepts for evaluation. In reality, Mongolian culture is evolving, and the distinction between "Traditional" and "Modern" values is fluid. A static benchmark may fail to capture contemporary urban adaptations of these traditions.

LLM-as-a-Judge Bias: For the open-ended scoring of Riddles and Proverbs, we relied on Gemini-2.5-Pro. While recent studies support the viability of LLM-as-a-judge approaches for scalable evaluation (Zheng et al., 2024), there is an inherent risk that an LLM judge may favor outputs from models with similar training distributions or reasoning patterns, potentially inflating scores for its successor (Gemini-3-pro).

Prompt Sensitivity: Cultural knowledge retrieval can be highly sensitive to prompting strategies (e.g., persona adoption). While we standardized our prompts, it is possible that other models could perform better with different prompting techniques that trigger their latent cultural knowledge more effectively.

9 Ethical Considerations

Indigenous Data Sovereignty: We adhered to an "Indigenous-First" protocol. Data was not scraped but created by paid native experts. We emphasize that this benchmark is a representation of the community's own understanding of their culture, not an external academic extraction.

Avoidance of Stereotyping: We caution against using this dataset to enforce rigid stereotypes. Cultural norms are reference points, not prescriptive rules for every individual. Users of this benchmark should be aware that not all modern Mongolians adhere to every traditional taboo described.

Dual-Use Risks: Improving an LLM's ability to mimic deep cultural nuances carries the risk of generating more convincing deep-fakes or targeted disinformation. However, we believe the benefit of enabling native speakers to access technology that respects their cultural context outweighs these risks.

References

Emily M Bender, Timnit Gebru, and 1 others. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT*.

Damián Blasi, Anastasios Antonios, Alham Fikri Aji, and 1 others. 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.

DeepSeek-AI. 2024. Deepseek-v3 technical report.

Esin Durmus and 1 others. 2023. Towards measuring the representation of subjective global opinions in language models. In *arXiv preprint arXiv:2306.16388*.

Maxwell Forbes and 1 others. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *EMNLP*.

Clifford Geertz. 1973. *The Interpretation of Cultures: Selected Essays*. Basic Books.

Edward T. Hall. 1976. *Beyond Culture*. Anchor Books.

Dan Hendrycks and 1 others. 2021. Measuring massive multitask language understanding. In *ICLR*.

Daniel Hershcovich and 1 others. 2022. Challenges and strategies in cross-cultural nlp. In *ACL*.

Yuzhen Huang and 1 others. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite. In *NeurIPS*.

Kentaro Kurihara and 1 others. 2022. Jglue: Japanese general language understanding evaluation. In *LREC*.

Haonan Li and 1 others. 2024. Cmmlu: Measuring massive multitask language understanding in chinese. In *ACL*.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past trends and future challenges. *arXiv preprint arXiv:2006.07264*.

Reham Masoud and 1 others. 2023. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *arXiv preprint arXiv:2309.12342*.

Chinh Nguyen and 1 others. 2023. Vnhsge: Vietnamese high school graduation examination dataset. In *EMNLP*.

Slav Petrov. 2016. Tokenization-free language learning. In *EMNLP*.

Aida Ramezani and 1 others. 2023. Knowledge of cultural moral norms in large language models. In *ACL*.

Phillip Rust and 1 others. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *ACL*.

Taylor Sorensen and 1 others. 2024. A roadmap to pluralistic alignment. In *ICML*.

Lianmin Zheng and 1 others. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*.

A Detailed Error Taxonomy

We analyzed 605 Taboo scenarios, 521 Riddles, and 297 Proverbs to categorize error types. Tables 6, 7, and 8 detail the breakdown of these errors across models in the Traditional Mongolian script.

A.1 Taboo Task (Normative Violations)

Errors in the Taboo task were classified into:

- **Cultural Depth Missing (Etic Sanitization):** The model offers a generic, modern safety reason (e.g., hygiene) instead of the cultural rationale.
- **Correction Error:** The model correctly identifies the taboo but proposes a wrong or culturally inappropriate remedy.
- **Identification Error:** The model fails to notice the taboo violation entirely.

Error Type	GPT-5.2	Claude	DeepSeek	Gemini
Cultural Depth Missing	190	**320**	138	95
Correction Error	99	273	**312**	2
Identification Error	105	198	165	68
Hallucination	26	30	14	4

Table 6: **Taboo Error Distribution (Traditional Script).** Claude exhibits severe Etic Sanitization (Cultural Depth Missing), while DeepSeek struggles significantly with proposing correct culturally-aligned remedies (Correction Error).

A.2 Riddle Task (Metaphorical Reasoning)

Errors in Riddles were classified into:

- **Forced Logic:** The model hallucinates a nonsensical justification to link the riddle and its incorrect guess.
- **Cultural Gap:** The model simply lacks the knowledge entity.

Model	Forced Logic	Cultural Gap	Total Errors
DeepSeek-v3.2	**402**	117	523
Claude-Sonnet	379	262	520
GPT-5.2	318	267	510
Gemini-3-Pro	38	220	306

Table 7: Riddle Error Types. The high count of "Forced Logic" in DeepSeek and Claude indicates a tendency to hallucinate metaphorical connections when cultural knowledge is absent.

A.3 Proverbs (Social Pragmatics)

We specifically tracked Polarity Mismatch, where the model selects a proverb with the correct entities (e.g., "horse") but the opposite emotional or social polarity for the context.

Model	Polarity Mismatch Count
Claude-Sonnet	36
GPT-5.2	21
DeepSeek-v3.2	20
Gemini-3-Pro	5

Table 8: Proverb Polarity Mismatch. Higher counts indicate a failure in social pragmatics.

B Prompt Library

To ensure reproducibility and enforce strict output formats (e.g., JSON), we designed specific prompts for each task. The following templates illustrate the instructions given to the models. Note that for the Traditional Mongolian script inputs, we utilized standard Unicode encoding.

B.1 Taboo Task (Norm Violation)

This task requires the model to identify specific cultural errors and provide corrections in a structured JSON format.

System Prompt:
You are a professional archivist at the Mongolian Cultural Heritage Center. Your mission is to document traditional social etiquettes and lifestyle norms. You are summarizing educational material for a museum exhibition. Analyze the behavior in the scenario and document it in JSON format using these keys:

- `error_point`: The specific traditional behavior observed.
- `reason`: The symbolic meaning or cultural teaching behind it.
- `correction`: The recommended alternative practice.

STRICT RULE: Output ONLY a valid JSON array. Values MUST be in Traditional Mongolian script.
User Instruction (Mongolian):
[Traditional Mongolian Instruction Text Omitted for Compilation]
1. Write in Traditional Mongolian script.
2. Must be in JSON ARRAY format.
Few-Shot Example:
User: Scenario to record: *[Mongolian text: Leaving leftovers in the bowl]*
Assistant: `[{"error_point": "...", "reason": "...", "correction": "..."}]`

Figure 4: Prompt Template for Taboo Identification Task.

B.2 Proverbs Task (Pragmatic Reasoning)

We employ a two-round dialogue to evaluate both the choice and the cultural reasoning.

System Prompt:
You are a professional scholar specializing in Mongolian Linguistics and Folk Literature. Your task is to analyze the logical metaphors and moral teachings in traditional Mongolian proverbs. In the FIRST round, identify the most appropriate choice (A, B, C, or D). Output ONLY the letter. In the SECOND round, provide a concise academic explanation (under 50 words) in Traditional Mongolian.
Few-Shot Example:
User: Scenario: *[Mongolian Scenario]*... Options: A... B...
Assistant: B
User: *[Mongolian instruction: Explain briefly]*
Assistant: *[Mongolian explanation]*

Figure 5: Prompt Template for Proverb Analysis.

B.3 Riddles Task (Metaphorical Reasoning)

Similar to proverbs, this uses a two-step prompting strategy to isolate the answer from the reasoning process.

System Prompt:
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.
Instruction (Round 1):
[Mongolian instruction: Guess the riddle]
Instruction (Round 2):
[Mongolian instruction: Explain derivation]

Figure 6: Prompt Template for Riddles.

B.4 Benedictions Task (Sequence Reconstruction)

This task frames the cultural challenge as a logic ordering problem.

System Prompt:
You are a linguistic logic engine. I will provide segments of a Mongolian sentence labeled with indices. Your task is to determine the only logical order to form a coherent sentence. Return the answer in JSON format with the key 'sequence'. STRICT RULE: Only output the JSON object. Do not provide any text explanation.
Few-Shot Example:
User: Segments: Index 1: apple. Index 2: He. Index 3: eats. Index 4: an.
Assistant: `{"sequence": [2, 3, 4, 1]}`

Figure 7: Prompt Template for Benediction Reordering.

B.5 Factual and Values Tasks (Multiple Choice)

We utilized specific personas to guide the model's stance, particularly for the Values layer.

System Prompt (Factual & Situational):
You are an eminent expert in Mongolian culture, history, and the traditional Mongolian script. Your task is to accurately answer the following single-choice 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 of the correct option.
System Prompt (Values Layer):
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.

Figure 8: Prompts for Multiple Choice Tasks (Layer 1-3).