From Guanyin, UFOs to Paradise: Capturing Cultural Variation in Dream Interpretation

Anonymous ACL submission

Abstract

Humans have long explored dreams, from predicting fortune and future to reflecting the subconscious. This curiosity now extends to large language models (LLMs). Commercial LLMs exhibit preliminary dream interpretation abilities, while open-source research remains limited to monolingual, western-centric datasets, with evaluations largely confined to classifications. We address these gaps by introducing a bilingual dataset of 31,877 unique dream-interpretation pairs across three cultural contexts: China, the Islamic and the West in English and Arabic. Fewer than 22% dream symbols overlap across cultures. Chinese symbols emphasize scenario-based activities and figures like *Guanyin*, Islamic references religion and concepts (*paradise*, *fasting*), while the West draws on technology like *UFOs*. We evaluated 17 models. New state-of-the-art models integrating general-purpose and reasoning modes into one model perform best in reasoning mode, while earlier models separating chat and reasoning favor chat settings. While language is not a bottleneck for SOTA models, capturing cultural nuances of under-represented regions e.g., the Islamic remains challenging. Fine-tuning of six LLMs shows that LoRA benefits larger models, while full-parameter is better for smaller ones. Although SFT equips models with cultural knowledge, post-training knowledge is less stable than pre-training, exhibiting sensitivity to training settings. Data and code are available at `URL Withheld`.

1 Introduction

Dreams have long fascinated humans (Harris-McCoy, 2012). Freud proposed that dreams express repressed desires and relieve internal tension (Freud, 1900). Later studies explored their psychological and neurological relevance (Wamsley and Stickgold, 2011; Wamsley, 2014; Zadra and Stickgold, 2021), connection to memory and consciousness (Siclari et al., 2017), and analyzing dream narratives (Domhoff and Schneider, 2008; Laureano and Calvo, 2024). Early analyses relied on human specialists (Elce et al., 2021), followed by NLP methods to analyze narrative structure and content. Recent work investigates dream analysis with LLMs (Niederhoffer et al., 2017; McNamara et al., 2019; Juncker, 2023; Laureano and Calvo, 2024). More literature review in Appendix A.

While these efforts have advanced dream understanding, little attention has been devoted to *dream interpretation*, which seeks to derive symbolic, cultural, and contextual meaning from dream content. Most publicly available datasets and studies are centered on English and Western cultures and adopt linguistic, emotional, psychological or biological views rather than cultural symbolism. Thus, they capture only a subset of interpretive traditions.

Our case study shows that <22% of dream symbols overlap across Western, Chinese and Islamic cultures (Section 3). Even when symbols coincide, their interpretations diverge. In Figure 1, all three cultures associate *Water* with positive meanings, whereas *Fire* varies: threat, betrayal or temptation in Western culture, wealth and prosperity in Chinese, and both danger and wealth in Islamic.

To mitigate this gap, we collect a bilingual dream interpretation dataset from three culture sources — West, China and Middle East, resulting in a total of 31,877 unique entries (a dream symbol + its interpretation) in Arabic and English. We used these entries to formulate four tasks: free-form question answering (QA), multi-choice question answering (MCQ), is it a *good or bad* dream (GB), and is this interpretation *true or false* (TF). The dataset supports both training and evaluation.

Based on our dataset, we analyzed dream symbols and interpretations across cultures. Evaluations across 17 LLMs show that new state-of-the-art models with two modes (general-purpose and reasoning) in one model outperform earlier models separating these two modes. Current LLMs have

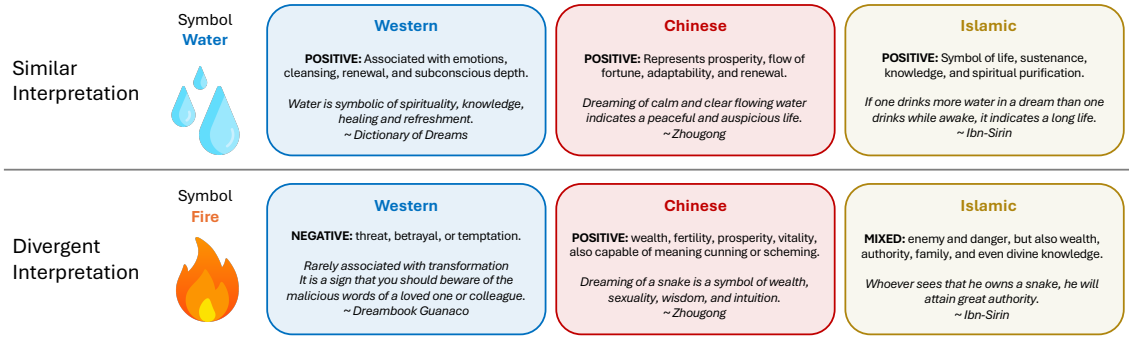

Figure 1: Dream Interpretations of symbols *Water* and *Fire* cross three cultures.

overcome language barriers while cultural nuances remain room to improve. Further fine-tuning of six LLMs shows that LoRA benefits larger models, while full-parameter tuning is better for smaller ones. Our contributions are summarized as follows:

- We curated a large-scale bilingual dream interpretation dataset for training and evaluation, covering Western, Chinese and Islamic cultural contexts.
- We analyze cross-cultural similarities and differences in interpreting the same dream symbols, revealing culturally grounded divergence in symbolic meanings.
- We evaluate 17 LLMs and fine-tune six models, revealing the advantage of *one model, two modes*, and the optimal SFT settings for cross-cultural dream interpretation.

2 Dataset

We curated the dataset by collecting raw dream–interpretation pairs from three cultural contexts. For the Middle East, we extracted entries from five books. Two sources for the West are from Kaggle, one from Huggingface, and one from webpages which required extensive cleaning, preprocessing, and manual verification. Chinese entries were gathered from both webpages and a SQL database. The raw data distribution across sources is shown in Table 6. Then, all entries were cleaned, deduplicated, consolidated, and manually validated, resulting in 5,568 Islamic, 16,720 Western, and 9,589 Chinese unique entries (Table 1). Further details are provided in Appendix B.1.

To prevent data leakage, we split the dataset into training and test at the level of dream symbols rather than dream-interpretation pairs, ensuring that all interpretations of a given symbol remain within the same split. This procedure is applied independently to the Western, Chinese, and Islamic subsets in order to preserve cultural balance.

Culture	Pairs	QA	MCQ	GB	TF	Train	Test
Islamic	5,568	4,466	5,568	5,568	5,568	4,466	1,102
Western	16,720	16,720	1,675	–	–	15,045	1,675
Chinese	9,589	9,589	959	–	–	8,630	959

Table 1: Dataset statistics of unique pairs and their four task formats: QA, MCQ, Good/Bad (GB), and True/False (TF) across three cultures.

Four Task Formulation From the cleaned dream–interpretation pairs, we categorized dream symbols into 17 groups (Figure fig:category-distribution). For free-form QA and MCQ tasks, we used LLMs to generate culturally specific questions that mimic user queries (Appendix B.3). The gold answer in both tasks was the original interpretation, while four distractors for MCQ were sampled from interpretations of symbols within the same category as the target symbol. This makes four distractors topically plausible while incorrect, increasing the difficulty of the task and preventing trivial elimination strategies.

In addition to QA and MCQ, users often pose queries such as *I dreamed of a snake yesterday, is it a good sign?* or *I dreamed of a snake, and someone told me it means earning more money, do you think this is true?* To better reflect these real-world inquiries, we introduce two additional tasks: *(i)* determining whether a dream is a *good or bad* sign, and *(ii)* verifying whether a dream's meaning matches the user's assumption (*true or false*).

For the Good/Bad task, we used an LLM to label each dream interpretation entry as positive or negative. The True/False task was designed using contrastive reasoning: for each dream symbol, the correct interpretation was paired with a distractor from the same category, yielding QA-style items

with one true and one false option. We extended these two tasks only to the Islamic test splits, as their higher difficulty compared to the Western and Chinese subsets makes them better suited for distinguishing model capabilities and assessing performance consistency across inquiry styles.

In each task extension, we emphasized culture-specific interpretive perspectives when designing questions. For example, classical Chinese traditions draw on the five elements (metal, wood, water, fire, and earth: 金木水火土), Yin–Yang (阴阳), and fate (命格). Western questions reflect astrological, zodiac, semiotic, and psychological views. This ensures that the generated questions are aligned with the sourced culture while maintaining diversity through distinct questions per entry.

We then used Gemini to translate examples: Arabic to English, Chinese to English and Arabic, and vice versa, ensuring that all cases are available in two languages with identical content. Table 1 summarizes the dataset statistics for each task along with the final training and test splits. For both task formulation and translation, we sampled 50-100 cases for each task per language and asked native speakers to validate the quality.

3 Dream Interpretations Across Cultures

Each cultural source differs in the number of dream symbols and interpretations it contains. The Islamic subset has 3,285 unique symbols and a total of 5,568 entries, meaning many symbols link to more than one interpretation. In comparison, the western source has 11,470 unique symbols out of 16,601 total entries and the Chinese source has 8,552 unique symbols out of 9,589 total entries.

That is, symbol-entry ratio is 59% for Islamic source, 68.6% in the Western source and 89.2% in the Chinese. The lower unique-symbol ratio in the Islamic source indicates that individual symbols with multiple interpretations are more frequent than other traditions. For example, the symbol *Wife* appears with several distinct interpretations in the Islamic data. In contrast, the Western and Chinese corpora mostly assign a single interpretation per symbol, emphasizing broader symbol coverage rather than interpretive depth.

For cross-cultural analysis, we used GPT-4o to categorize dream symbols into 17 semantic types with the prompt in Figure 8. Figure 2 presents the percentage distribution of these categories across cultures. We compare their relative frequencies to identify culturally dominant patterns below.

3.1 Culturally Dominant Symbols

Across all corpora, *Tools & Physical Objects* are the most frequent category (Islamic: 17.7%, Western: 17.2%, Chinese: 15.0%), indicating a shared tendency that most dreams present concrete everyday imagery. In the Islamic source, many object-based symbols carry spiritual or practical resonance (e.g., *milk*, *ring*, *sword*, *gold*, *hair*). The Western source similarly includes a wide range of objects, from *tree* and *car* to *computer* and *cake*. The Chinese source also contains common objects (e.g., *rice*, *orange*, *fire*) as well as culturally grounded items such as *jade bracelet* and *Mahjong tiles*.

Islamic Religious Profile Islamic dream interpretation, exemplified by classical works such as those attributed to Ibn Sīrīn, is shaped by Islamic religious concepts and the social realities of pre-modern life. Many symbols are closely tied to religious practice and moral themes, including paradise, fasting, and veiling, as well as objects with historical or ritual relevance. The frequent presence of multiple interpretations for a single symbol reflects contextual nuance and scholarly variation within the tradition.

Consistently, the Islamic corpus contains a comparatively larger share of explicitly religious categories (11.9%) than the Western (5.5%) and Chinese (2.3%) corpora. Illustrative examples include entries referencing *Qur'anic verses*, the *Ka'aba*, *ritual purity*, and revered religious figures, indicating that religious concepts frequently serve as a central interpretive frame.

Chinese Social and Emotional Themes Chinese dream symbols are characterized by detailed, scenario-based entries and culturally grounded figures. The corpus covers everyday experiences (e.g., shopping, exams, relationship milestones) alongside widely recognized spiritual icons such as *Guanyin* (观音) and historical heroes like *Guan Gong* (关公). Traditional frameworks, including *Yin–Yang* and the *Five Elements*, also appear as interpretive resources. Family relations recur as a major theme, with many dreams involving spouses, in-laws, pregnancy, and domestic circumstances, highlighting the importance of kinship and social roles.

Numerous entities involve relationship situations (e.g., *a girlfriend* or *conditions affecting a spouse*) and nuanced affective states (e.g., *crying loudly*,

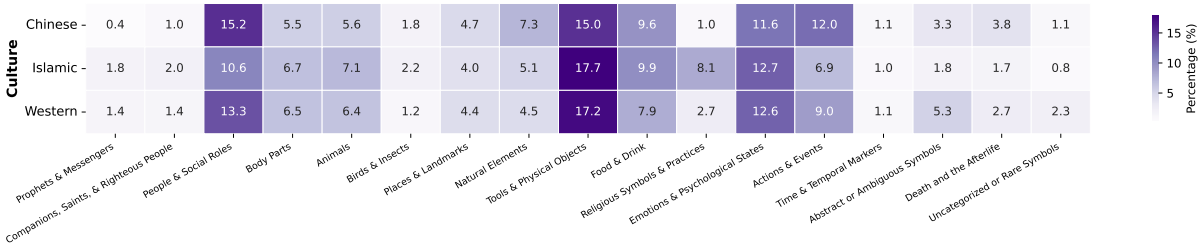

Figure 2: Percentage distribution of dream symbols across 17 categories of three cultures.

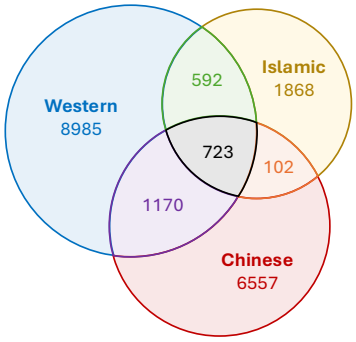

Figure 3: Distribution of unique dream symbols across Islamic, Western, and Chinese cultural traditions.

laughing), suggesting the focus on interpersonal experience and everyday life dynamics. This trend is reflected in the category distribution: the Chinese corpus places greater emphasis on *People and Social Roles* (15.2%), *Actions and Events* (12.0%), and *Emotions and Psychological States* (11.6%).

Western Symbolic Openness The Western corpus shows a relatively higher share of *Abstract or Ambiguous Symbols* (5.3%) than the Chinese (3.3%) and Islamic (1.8%). These entities often convey uncertainty or metaphorical meaning, including *darkness*, *unknown figures*, *mysterious places*, and *unclear voices*. It also includes modern and media-shaped symbols such as *UFO*, which is often interpreted as curiosity, uncertainty, or encounters with something beyond the familiar. In addition, it contains many common dream scenarios (e.g., *falling*, *chasing*, *being trapped*), which are often linked to stress, pressure, or feelings of losing control in everyday life.

3.2 Overlap and Uniqueness of Symbols

A central question in cross-cultural dream studies is the extent to which different traditions share common symbols. Some motifs (e.g., *sun*, *water*, *dog*) appear across cultures, while others are tied to particular cultural contexts.

Figure 3 summarizes how symbols overlap across the Islamic, Western, and Chinese sources. Out of **19,997** total distinct symbols in the combined set, only **723 (3.6%)** appear in all three corpora (22% by IS). These shared symbols mostly reflect broadly common dream imagery, including animals (e.g., *dog*, *cat*), natural elements (e.g., *fire*, *rain*), and basic social roles (e.g., *mother*, *king*). Beyond this shared set, **592** symbols occur only in Islamic and Western (Islamic: 18.0%, Western: 5.2%), **102** only in Islamic and Chinese (Islamic: 3.1%, Chinese: 1.2%), and **1,170** only in Western and Chinese (Western: 10.2%, Chinese: 13.7%).

Each tradition also contains a substantial set of unique symbols: **1,868 (56.9%)** are exclusive to the Islamic corpus, **8,985 (78.3%)** to the Western, and **6,557 (76.7%)** to the Chinese. This high level of uniqueness highlights how dream symbolism reflects culture-specific references and interpretive traditions. For example, the Islamic symbols are grounded in Islamic and Middle Eastern religious imagination (e.g., *Garments of Jannah*). The Western corpus includes modern and media-related symbols (e.g., *television*, *UFO*). The Chinese corpus contains culturally specific figures and practices (e.g., *Guanyin Bodhisattva* (观音), *playing Mahjong*). Together, these examples illustrate how cultural context shapes both which symbols appear and how they are interpreted.

In summary, three traditions share a small common core related to nature, animals, and basic human roles, and each corpus preserves a large culture-specific inventory reflecting local practices, beliefs, and everyday life.

3.3 Summary

Across all three corpora, symbols related to objects and daily life are the most frequent. Shared symbols tend to involve common experiences (e.g., natural elements, animals, and familiar social roles), while unique symbols often reflect culturally grounded references and local scenarios.

The Chinese corpus stands out for its emphasis on social roles and event-based symbols, including many entries that describe relationship situations and everyday activities. In contrast, the Islamic and Western corpora more often focus on a single focal symbol (e.g., an object, person, or place) rather than a full scenario. Emotional states (e.g., *anger*, *joy*) appear in all corpora but remain relatively rare as standalone symbols.

Each culture's dream symbolism is a reflection of its values, environment, and interpretive traditions, from the spiritual undertones of Islamic dreams, to the individualistic and modern scope of Western dreams, to the rich cultural tapestry and pragmatic life scenarios seen in Chinese dreams. Such an analysis highlights how the human experience of dreaming is universal yet colored by cultural lenses.

4 Experiments

In this section, we evaluated 17 models with four tasks, including MCQ as the primary, free-form QA, Good/Bad, and True/False. We aim to answer three questions: *(i)* who wins for dream interpretation, general-purpose LLMs versus large reasoning models (LRMs), can Qwen interpret dreams better from the Chinese perspective? *(ii)* Can continuous supervised fine-tuning (SFT) with data varying from cultures and languages strengthen models' understanding of dreams? Compared with full-parameter SFT, when would LoRA SFT outperform? *(iii)* Who interprets and predicts the future more positively, humans or AIs?

4.1 Experimental Setup

Models We include three commercial LLMs and its reasoning variants: `DeepSeek-v3.1`, `GPT5`, `Claude4-Sonnet`, and 11 open-source models from 1B to 8B: `Qwen3-8B/1.7B` (chat and reasoning), `Llama-3.1-8B`, `Qwen2.5-Math-7B`, `Qwen2.5-1.5B-Instruct`, and their three `DeepSeek-R1` distilled reasoning counterparts, as well as `Llama-3.2-1B-Instruct`. All models used the same prompt templates. Appendix E details models, inference setups and prompts.

Evaluation Tasks and Metrics For the MCQ, Good/Bad and True/False tasks, gold labels are available, and we use accuracy as the evaluation metric. For QA, we employed `GPT5-mini` as a judge (prompts in Figure 18) to assess correctness and sentiment. Correctness was evaluated by comparing model responses against human interpretations on a 1-5 scale. Sentiment was assessed by first identifying responses as positive or negative, then calculating the percentage of positive ones.

LoRA and Full-parameter SFT Setups We used the model-specific chat template to format the question and corresponding interpretations from our QA training set, and then the cross-entropy loss is computed only on the answer. For both full-parameter and LoRA SFT, we train the model for 1 epoch (if not stated otherwise) with a learning rate of 1e-5. A cosine scheduler is applied with the warm-up ratio as 0.2. We tune the gradient accumulation and per-device batch size to ensure larger (7-8B, 64 accumulation steps with 1 sample per device) and smaller (1-2B, 1 accumulation steps with 64 samples per device) models have an equivalent batch size of 64 per step. For LoRA SFT, we add LoRA adapters on {Q,K,V,O} projections with r and α set as 64 and 16 respectively. All models are trained on two A100 GPUs with bf16 data type.

4.2 Who Wins?

Do LRMs Outperform LLMs for Dream Interpretation? For new SOTA LLMs — one model, two modes, a single LLM is trained to support both general-purpose use and deep reasoning within the same parameters, rather than using separate models. Reasoning is controlled by an argument, e.g., `reasoning_effort = medium/minimal` for GPT5 and `thinking = enabled/disabled` for Claude. As shown in Table 2, these new models including GPT5, Claude, DeepSeek-v3.1 and Qwen3 demonstrate that their reasoning mode is consistently superior to or on par with their chat mode. This unified design allows models to flexibly combine general knowledge with reasoning, rather than treating them as disjoint capabilities.

Earlier models (Qwen2.5-1.5B/7B, Llama-3.1-8B) separate chat and reasoning variants. Reasoning models are often inferior or comparable to the chat versions, as the two cannot effectively benefit from one another. This suggest that dream interpretation requires both extensive knowledge and moderate reasoning, making models unified knowledge and reasoning well-suited to the task. Hence, Qwen3-1.7B performs comparably to Qwen2.5-7B, and Qwen3-7B is on par with DeepSeek-v3.1.

Do Chinese-Centric Qwen Perform Better in Zhougong Dream Interpretation? The answer is *No*. Chinese-centric models such as Qwen and

Model	English MCQ Test				Arabic MCQ Test			
	IS	WE	ZH	Avg	IS	WE	ZH	Avg
Closed-source Models								
GPT5	71.7	91.3	**99.8**	87.7	74.4	84.5	99.6	85.4
GPT5-R	**73.9**	**98.0**	**99.8**	**91.3**	**80.7**	96.8	**100.0**	**92.9**
Claude-Sonnet-4	70.9	97.7	99.6	90.3	68.9	98.4	99.7	90.0
Claude-Sonnet-4-R	69.8	97.1	99.6	89.7	68.1	**98.5**	99.9	89.9
DeepSeek-v3.1-Terminus	61.9	97.4	99.4	87.4	65.2	95.0	98.4	87.1
DeepSeek-v3.1-Terminus-R	68.5	96.2	**99.8**	88.9	71.8	93.4	99.5	88.6
Open-source Models								
Qwen3-8B	58.0	96.1	98.8	85.6	54.4	91.6	96.6	81.9
Qwen3-8B-R	59.3	97.1	**99.8**	86.6	61.0	93.7	98.6	85.3
Qwen3-1.7B	51.0	87.0	89.5	77.0	46.2	74.0	71.5	65.2
Qwen3-1.7B-R	50.5	92.0	95.9	80.8	47.1	80.3	81.2	70.7
Qwen2.5-7B-Instruct	47.7	95.0	98.2	81.9	48.6	89.8	94.3	78.8
DeepSeek-R1-Distill-Qwen-7B	43.0	84.9	91.4	74.2	33.6	40.1	37.1	37.4
Llama-3.1-8B-Instruct	40.9	95.8	97.7	80.1	26.0	68.8	78.2	58.6
DeepSeek-R1-Distill-Llama-8B	45.3	93.5	98.2	80.5	34.3	70.4	74.0	60.7
Qwen2.5-1.5B-Instruct	46.2	70.0	54.4	59.0	44.9	54.9	39.3	48.0
DeepSeek-R1-Distill-Qwen-1.5B	33.3	53.7	58.3	48.8	23.3	23.4	22.2	23.1

Table 2: Accuracy (%) of eight LLMs and their reasoning variants on MCQ test sets in two languages. Higher scores between general-purpose and reasoning models are highlighted across three cultures (IS–yellow , WE–blue , ZH–red , Avg–green); best per column is **bolded**.

DeepSeek do not display a significant advantage on the Chinese (ZH) subset. Most models perform well on both Western (WE) and Chinese subsets. Instead, performance on the Islamic (IS) subset is the decisive factor for overall accuracy, with models showing a 10-50% gap compared to Western and Chinese cultures, particularly when Islamic interpretations are presented in Arabic. For instance, Qwen3-8B achieves 98.8% on the ZH subset in English but only 54.4% on the IS in Arabic.

We speculate that while state-of-the-art models have improved in low-resource languages, they still lack cultural nuance optimization. This explains why models, including commercial ones, perform well on Western and Chinese subsets regardless of presentation language, but fail on Islamic cultural assessments, where underrepresented cultural knowledge remains a bottleneck (Chiu et al., 2025).

Language Variation Still Impacts Old Models. On the identical MCQ test set presented in two languages, commercial models and Qwen3-8B demonstrate strong robustness to language variation, performing similarly on English and Arabic. In contrast, smaller and older models show clear disadvantages on the Arabic test set. This suggests that larger SOTA models above 7B have been specifically optimized towards better multilingual communication, handling Arabic better.

Model	English Test			Arabic Test		
	MCQ	GB	TF	MCQ	GB	TF
Qwen3-1.7B	51.0	67.6	52.5	46.2	66.8	48.9
+ LoRA-All	51.2	68.8	52.8	44.1	61.0	48.2
Qwen3-1.7B-R	50.5	66.5	60.0	47.1	67.0	55.9
Qwen3-8B	58.0	67.1	52.7	54.4	68.4	50.3
+ LoRA-All	57.2	67.2	51.4	49.8	68.8	51.4
Qwen3-8B-R	59.3	69.3	58.9	61.0	70.7	55.6
GPT5	71.7	69.6	50.4	74.4	73.9	65.4
GPT5-R	73.9	68.8	55.9	80.7	72.8	69.3

Table 3: Accuracy (%) of Qwen3-1.7B/8B and GPT5 on Islamic culture test sets across three tasks.

Can Different Task Formulations Reveal Consistent Results? *Yes*. Given that most models achieve over 90% accuracy on the WE and ZH cultures, we focus on the Islamic subset to assess consistency across task formulations. As shown in Table 3, the GB and TF tasks largely mirror the MCQ trend: models perform better in English than in Arabic, with the exception of GPT5, which shows robust performance in both languages; overall performance of Qwen3-8B and 1.7B follow GPT5. Table 5 for free-form QA exhibits the same pattern. This confirms that model behavior is consistent

Test Set Language		English MCQ						Arabic MCQ					
SFT Training Data		En		Ar		All		En		Ar		All	
Model	Cul	*Full*	*LoRA*	*Full*	*LoRA*	*Full*	*LoRA*	*Full*	*LoRA*	*Full*	*LoRA*	*Full*	*LoRA*
Qwen3-8B	IS	49.5	57.3	65.4	57.4	60.0	57.2	5.3	53.6	0.5	53.8	34.6	49.8
	WE	97.7	96.2	97.7	96.4	98.2	96.4	92.8	91.5	70.5	91.9	94.8	92.5
	ZH	98.5	98.8	98.7	98.8	99.3	99.1	95.5	96.5	92.7	96.7	96.2	96.8
	Avg	83.7	85.4	**88.4**	85.5	87.2	85.5	67.6	81.6	55.5	**81.9**	77.4	81.0
Qwen3-1.7B	IS	46.2	51.2	50.6	51.4	39.9	51.2	45.8	45.9	48.2	45.2	41.7	44.1
	WE	94.7	87.7	92.2	88.2	93.1	89.8	84.0	74.5	83.0	76.8	80.9	78.2
	ZH	95.6	89.5	93.7	90.2	95.6	90.7	83.5	70.9	85.2	72.5	82.5	73.5
	Avg	**80.6**	77.4	80.3	77.9	78.0	78.7	72.6	65.2	**73.3**	66.4	69.7	66.9
Qwen2.5-7B-Instruct	IS	38.7	48.4	50.8	48.0	42.7	49.9	50.1	48.0	53.6	47.4	51.3	46.4
	WE	93.1	95.7	96.0	95.5	90.3	96.7	93.9	90.9	91.6	91.5	88.4	93.3
	ZH	90.9	98.5	96.1	98.4	93.7	98.5	82.3	94.7	92.6	95.7	90.8	96.7
	Avg	76.5	82.5	82.7	82.3	77.1	**83.4**	78.0	79.2	**80.6**	79.6	78.1	80.3
Llama-3.1-8B-Instruct	IS	40.6	30.9	23.3	34.9	37.3	17.6	26.0	15.4	4.3	12.5	41.6	7.5
	WE	90.6	95.6	89.1	96.4	79.8	94.2	46.3	58.3	59.0	66.0	54.9	51.1
	ZH	55.6	97.5	88.4	98.0	53.8	97.8	25.6	70.7	46.9	77.1	32.4	47.8
	Avg	66.9	77.0	69.5	**78.7**	60.6	72.5	35.1	48.8	39.8	**53.1**	45.2	37.4
Qwen2.5-1.5B-Instruct	IS	55.1	46.2	46.8	45.6	55.6	45.7	45.6	43.8	44.0	43.5	46.6	42.6
	WE	74.8	70.8	72.6	71.4	73.9	71.7	58.4	55.6	59.7	58.3	59.4	58.4
	ZH	50.4	53.7	49.2	54.9	48.2	52.6	34.1	39.3	36.3	41.3	32.2	38.6
	Avg	**62.8**	59.2	59.0	59.6	62.0	59.2	48.5	48.0	49.1	**49.6**	48.7	48.7
Llama-3.2-1B-Instruct	IS	4.9	4.9	4.9	4.9	4.8	4.9	5.7	4.9	4.1	4.9	3.4	4.9
	WE	19.0	19.2	19.1	19.2	19.1	18.9	19.0	18.9	18.9	18.9	19.1	18.9
	ZH	18.9	18.8	18.9	18.9	18.9	19.1	19.1	18.9	18.9	18.9	19.1	18.9
	Avg	14.8	14.9	14.9	14.9	14.8	14.8	**15.1**	14.8	14.6	14.8	14.4	14.8

Table 4: Accuracy (%) of five LLMs fine-tuned by full-parameter SFT vs. LoRA SFT in three training dataset settings: English QA, Arabic QA and their mixture. Evaluation on MCQ test sets.

across tasks. However, even for the "easier" binary classification tasks (GB and TF), accuracies remain below 74%, highlighting limited knowledge of Islamic dream interpretation logic across models.

4.3 Full-parameter vs. LoRA SFT

We fine-tuned six open-source models for one epoch under two settings: LoRA and full-parameter using training data in English, Arabic, and a combination of them, each covering all three cultures.

Does SFT Improve Accuracy? Figure 5 shows that, under appropriate setups, SFT generally improves performance in both languages. As detailed in Table 4, full-parameter SFT is more effective for smaller models (1/1.5B), while LoRA is preferable for larger models (7/8B).

SFT has a limited impact on Western and Chinese subsets due to their already high baselines. The Islamic subset is the key driver of overall accuracy. Improvements in Islamic culture (yellow cells) largely determine performance gains, where most yellow cells appear when training data is presented in either Arabic language or both (All), indicating that Islamic cultural nuances can be more effectively learned in Arabic language than through English language. Similarly, for Western culture (blue cells), gains are mainly observed with English or bilingual training data.

However, SFT can also degrade accuracy, particularly when applying full-parameter SFT to larger models (e.g., `Llama-3.1-8B` and Qwen3-8B), which requires more training epochs to recover disrupted knowledge as below.

Impact of Training Epochs Based on Qwen3-

	Correct		GT-Sent		AI-Sent	
Model	En	Ar	En	Ar	En	Ar
Qwen3-1.7B	3.3	2.6	50.5	52.2	61.3	56.4
+ LoRA-All	2.7	1.9	50.5	53.0	57.3	49.6
Qwen3-1.7B-R	3.4	2.8	51.6	51.4	62.7	56.8
Qwen3-8B	3.6	3.5	50.2	52.2	60.0	56.8
+ LoRA-All	3.2	2.7	51.1	51.1	59.4	53.7
Qwen3-8B-R	3.6	3.5	51.2	52.0	59.3	55.2
GPT5	4.1	4.1	50.5	52.7	56.7	56.5
GPT5-R	4.1	4.2	51.7	53.2	55.4	55.7
Avg	3.5	3.2	50.9	52.2	59.0	55.1

Table 5: Free-form QA responses' correctness and sentiment under non-reasoning, LoRA-SFT-All and reasoning. QA test set includes 2,634 questions reflecting Western and Chinese cultures, presented in two languages. Model responses are more positive than humans' roughly half-half distribution.

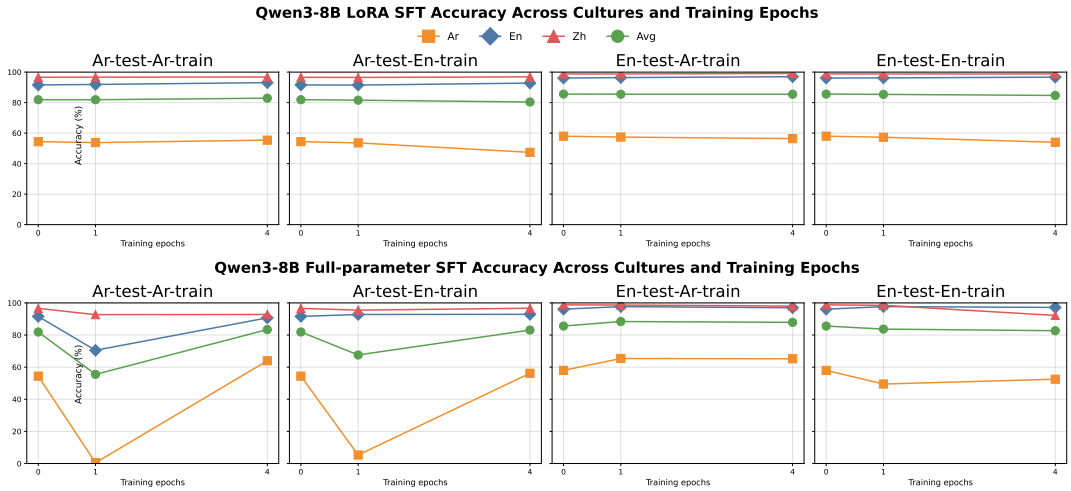

Figure 4: Qwen3-8B Full vs. LoRA SFT accuracy across training epochs and cultures.

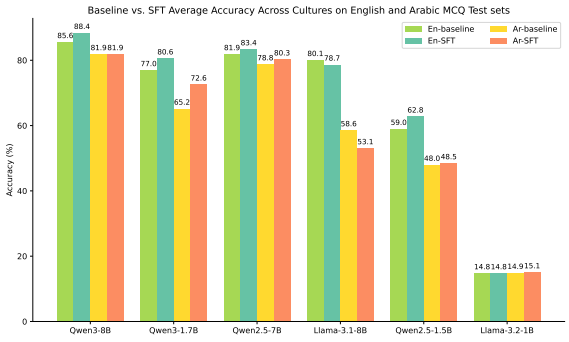

Figure 5: Baseline vs. SFT average accuracy across cultures on English and Arabic MCQ test sets.

8B, we compare the baseline without SFT against full-parameter and LoRA SFT with 1 and 4 training epochs, analyzing the impact of training epochs on dream interpretation across cultures (Figure 4). We find two main results. First, with LoRA SFT, both the number of epochs and the training language have a negligible impact on accuracy when evaluated on English and Arabic MCQ test sets. Second, full-parameter SFT with either English or Arabic data disrupts the model's internal knowledge of Islamic dream interpretation, causing a sharp drop after one epoch, while leaving Western and Chinese subsets largely unaffected. Additional epochs can recover and enhance performance on the Islamic culture subset.

These results suggest that Western and Chinese interpretive traditions are already deeply internalized into parameters during model pretraining, whereas Islamic nuances might be acquired during post-training, which were encoded more shallowly, yielding lower baselines and greater sensitivity to SFT. This aligns with previous findings that knowledge learned during fine-tuning is not well-grounded and tends to induce hallucinations (Gekhman et al., 2024).

4.4 Are AIs More Positive than Humans?

We evaluated three model pairs: Qwen3-7B/1.7B and GPT5 (chat and reasoning) on 2,634 free-form QA examples spanning Western and Chinese cultures in two languages. As shown in Table 5, human interpretations exhibit no sentiment preference, with positive and negative each holding half. In contrast, model outputs show a slight positive tendency, with 59% positive responses in English and 55% in Arabic.

5 Conclusion

We introduce a bilingual dream interpretation dataset spanning Western, Chinese, and Islamic cultures, with 31,877 unique entries. We find that less than 22% of dream symbols overlap across cultures, highlighting the cultural divergence. Using four task formulations, we evaluated 17 LLMs and fine-tuned six models, demonstrating the advantage of the *one model, two modes* paradigm and the capacity of LLMs to handle low-resource languages. However, while models can process the language, they fail to capture cultural nuances, performing markedly worse on Islamic culture interpretations than on Western and Chinese ones. We further identify optimal SFT configurations for models 1-8B, and observe that humans interpret dreams with a roughly neutral balance of positive and negative meanings, whereas models exhibit a slight positive bias. In future work, we plan to focus on enhancing models' sensitivity to cultural nuances and fine-tuning them with reasoning trajectories alongside direct interpretations.

Limitations and Future Work

Niche Scope Dream interpretation is a unique but niche topic. Experimental results and findings may not generalize to broader cultural reasoning tasks.

Insufficient Human Validation We used LLMs to transform dream interpretation entries into four task formulations. Although gold pairs were provided and the model was instructed to preserve content, LLM-induced errors may still occur. We manually inspected 50-100 samples per task for each culture and language to verify translation and task formulation quality. However, this limited validation covers only a small fraction of the dataset and lacks sufficient human auditing and inter-annotator agreement, leaving potential template and judge biases.

For evaluating free-form QA responses, we employed an LLM-based judge (GPT-5-mini) to assess correctness and sentiment. While efficient, this approach may introduce systematic bias. A human-evaluated subset with inter-annotator agreement would strengthen the credibility of the evaluation. Furthermore, the engagement with human experts, such as psychoanalysts or cultural anthropologists, for judging the correctness of free-form answers would be ideal for a subjective and culturally deep task.

Potential Pre-training Data Contamination Our data are mostly derived from dream dictionaries, publicly scraped web sources and books that were likely included in the pre-training corpora of the evaluated models. This raises concerns about the validity of zero-shot performance, as the models may simply be recalling information seen during pre-training rather than demonstrating genuine interpretation or reasoning capabilities.

To mitigate this risk, we design evaluation tasks with diverse formulations including QA. MCQ. Good or Bad, True or False. While models may have memorized free-form text, correctly interpreting dreams when they appear in different ways can still reflect models' generalization and alignment with users' expectations.

Source Reliability and Cultural Scope Some raw data sources lack scholarly validation and may not be fully reliable or culturally representative. Our analysis based on the collected dream symbols may not accurately reflect the true distribution of cultural overlap and divergence. Future work could improve representativeness by collecting first-hand sharing from real users on social media platforms, rather than relying primarily on dream dictionary or websites. Reliability could be further enhanced by using more authoritative sources, such as published books, scholarly translations, religious texts, and academic literature.

In addition to addressing the limitations above, we plan to select challenging and cross-cultural examples and host a public leaderboard or benchmark portal to support evaluations of future models. We also plan to explore research questions, such as whether differences between LoRA and full-parameter fine-tuning reveal novel behaviors in cross-cultural symbolic tasks or low-resource settings, and whether LLM performance can be further improved through prompt or context optimization strategies.

Ethical Statement

Data License and Copyright The Arabic datasets were derived from five books, four of which are ancient books whose authors died between 729-1826 and are therefore in the public domain. The fifth was from the book "Islamic Dream Interpretation" by Dr. Khaled Al-Anbari downloaded from https://www.alanbary.com/, which states "all rights reserved." However, we only extracted 733 dream symbols and their interpretations for research purposes with task reformulations and heavy data transformations. This limited use does not substitute for or replicate the original work and is permissible under established copyright exceptions for scholarly research, including U.S. Copyright Act ("fair use"), the EU Copyright Directive (EU) 2019/790 (Articles on text and data mining for scientific research), and corresponding "fair dealing" provisions in the UK, Canada, and other jurisdictions. Similarly for 77 Chinese examples extracted from the website https://www.yourchineseastrology.com/.

English data were either collected from Kaggle under the license of Apache 2.0 or Huggingface[1] under the license of GPLv3. Both of them allow research usage of the data. We will obey the policy of both. For the public webpage myislamicdream.com, there is no specific license.

Though two Chinese datasets were collected from websites, including https://www.

[1] https://huggingface.co/datasets/JosephusCheung/GuanacoDataset

yourchineseastrology.com/ and https://www.zgjmorg.com/, the content shown in these webpages is from an ancient Chinese book — zhougongjiemeng (《周公解梦》).

We did not collect any dream-interpretation pairs with personal information, Also, we did not use the raw data. We extracted dream symbols rather than using the original dream descriptions. This approach complies with website privacy policies and protects user anonymity. See more about data license in Table 6. If any data is found to require additional permissions, we will contact the content owner for authorization or exclude the relevant portions from the final release.

References

Edgar Altszyler, Sidarta Ribeiro, Mariano Sigman, and Diego Fernández Slezak. 2017. The interpretation of dream meaning: Resolving ambiguity using latent semantic analysis in a small corpus of text. *Consciousness and cognition*, 56:178–187.

Lorenzo Bertolini, Valentina Elce, Adriana Michalak, Giulio Bernardi, and Julie Weeds. 2023. Automatic scoring of dream reports' emotional content with large language models. *arXiv preprint arXiv:2302.14828*.

Lorenzo Bertolini, Adriana Michalak, and Julie Weeds. 2024. Dreamy: a library for the automatic analysis and annotation of dream reports with multilingual large language models. In *Sleep Medicine*, volume 115, pages 406–407. Elsevier RADARWEG 29, 1043 NX AMSTERDAM, NETHERLANDS.

Mark Blagrove, Laura Farmer, and Elvira Williams. 2004. The relationship of nightmare frequency and nightmare distress to well-being. *Journal of sleep research*, 13(2):129–136.

Abigail P Blyler and Martin EP Seligman. 2024. Personal narrative and stream of consciousness: an ai approach. *The Journal of Positive Psychology*, 19(4):592–598.

Rosalind Cartwright. 2011. Dreaming as a mood-regulation system. In *Principles and practice of sleep medicine*, pages 620–627. Elsevier.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Francis Crick and Graeme Mitchison. 1983. The function of dream sleep. *Nature*, 304(5922):111–114.

Susanne Diekelmann and Jan Born. 2010. The memory function of sleep. *Nature reviews neuroscience*, 11(2):114–126.

G William Domhoff. 2017. *The emergence of dreaming: Mind-wandering, embodied simulation, and the default network*. Oxford University Press.

G William Domhoff and Adam Schneider. 2008. Studying dream content using the archive and search engine on dreambank. net. *Consciousness and Cognition*, 17(4):1238–1247.

Valentina Elce, Giacomo Handjaras, and Giulio Bernardi. 2021. The language of dreams: Application of linguistics-based approaches for the automated analysis of dream experiences. *Clocks & Sleep*, 3(3):495–514.

S Freud. 1900. The interpretation of dreams, vol. 4, trans. *J. Strachey.(London, The Hogarth Press*, pages 26–28.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7765–7784, Miami, Florida, USA. Association for Computational Linguistics.

Calvin Hall and Robert Van de Castle. 1966. The content analysis of dreams.

Daniel E Harris-McCoy. 2012. Artemidorus' oneirocritica: Text, translation, and commentary.

P Hauri. 1975. Categorization of sleep mental activity for psychophysiological studies. *The experimental study of sleep: Methodological problems*, pages 271–281.

Sheldon Juncker. 2023. Dreaming with ai. *Poligrafi: revija za religiologijo, mitologijo in filozofijo*, 28(109/110).

Tracey L Kahan and Stephen P LaBerge. 2011. Dreaming and waking: Similarities and differences revisited. *Consciousness and Cognition*, 20(3):494–514.

Mayte H Laureano and Hiram Calvo. 2024. Computational study of dream interpretations: Psychoanalytic human vs artificial analyses. In *2024 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–9. IEEE.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.

Patrick McNamara, Kelly Duffy-Deno, Tom Marsh, and Thomas Jr Marsh. 2019. Dream content analysis using artificial intelligence. *International Journal of Dream Research*, pages 42–52.

David Nadeau, Catherine Sabourin, Joseph De Koninck, Stan Matwin, Peter D Turney, et al. 2006. Automatic dream sentiment analysis. In *Proceedings of the Workshop on Computational Aesthetics at the Twenty-First National Conference on Artificial Intelligence*.

Kate Niederhoffer, Jonathan Schler, Patrick Crutchley, Kate Loveys, and Glen Coppersmith. 2017. In your wildest dreams: the language and psychological features of dreams. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, pages 13–25.

Yuval Nir and Giulio Tononi. 2010. Dreaming and the brain: from phenomenology to neurophysiology. *Trends in cognitive sciences*, 14(2):88–100.

Amir H Razavi, Stan Matwin, Joseph De Koninck, and Ray Reza Amini. 2014. Dream sentiment analysis using second order soft co-occurrences (sosco) and time course representations. *Journal of Intelligent Information Systems*, 42(3):393–413.

Camila Sanz, Federico Zamberlan, Earth Erowid, Fire Erowid, and Enzo Tagliazucchi. 2018. The experience elicited by hallucinogens presents the highest similarity to dreaming within a large database of psychoactive substance reports. *Frontiers in neuroscience*, 12:7.

Michael Schredl. 2010. Dream content analysis: Basic principles. *International Journal of Dream Research*, pages 65–73.

Francesca Siclari, Benjamin Baird, Lampros Perogamvros, Giulio Bernardi, Joshua J LaRocque, Brady Riedner, Melanie Boly, Bradley R Postle, and Giulio Tononi. 2017. The neural correlates of dreaming. *Nature neuroscience*, 20(6):872–878.

Serge Thill and Henrik Svensson. 2011. The inception of simulation: a hypothesis for the role of dreams in young children. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Matthew P Walker and Els van Der Helm. 2009. Overnight therapy? the role of sleep in emotional brain processing. *Psychological bulletin*, 135(5):731.

Erin J Wamsley. 2014. Dreaming and offline memory consolidation. *Current neurology and neuroscience reports*, 14(3):433.

Erin J Wamsley and Robert Stickgold. 2011. Memory, sleep and dreaming: experiencing consolidation. *Sleep medicine clinics*, 6(1):97.

Antonio Zadra and Robert Stickgold. 2021. *When brains dream: Understanding the science and mystery of our dreaming minds*. WW Norton & Company.

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.

Xiaofang Zheng and Richard Schweickert. 2023. Differentiating dreaming and waking reports with automatic text analysis and support vector machines. *Consciousness and Cognition*, 107:103439.

Contents

A Related Work

A.1 Tracing Dream Study

Dreams have been a topic of human curiosity for centuries, dating back to ancient times. Historical approaches to dream interpretation began with Artemidorus in the 2nd century AD, who systematically studied dream content and proposed interpretive techniques in his work *Oneirocritica* (Harris-McCoy, 2012). A major shift occurred in the 19th century with Sigmund Freud's *The Interpretation of Dreams* (Freud, 1900). He studied that dreams express repressed desires and serve to relieve internal tension, thereby supporting sleep and overall well-being. Later theories proposed that dreams contribute to emotional regulation and conflict resolution (Cartwright, 2011; Walker and van Der Helm, 2009), memory consolidation (Diekelmann and Born, 2010), and forgetting irrelevant information to enhance learning (Crick and Mitchison, 1983). Some perspectives have also compared dreams to simulations that help individuals prepare for future challenges and threats (Thill and Svensson, 2011).

In recent decades, scientific interest in dreams has increased, particularly in understanding their psychological and neurological relevance. Research has highlighted connections between dreaming and psychophysical health, and pointed to potential roles in sleep-dependent memory processing (Wamsley and Stickgold, 2011; Wamsley, 2014; Zadra and Stickgold, 2021). Furthermore, as internally generated conscious experiences, dreams provide a valuable model for studying the nature of consciousness itself (Nir and Tononi, 2010; Siclari et al., 2017). Despite the growing interest, the processes underlying dream generation and the exact functions of dreams remain partially understood. A major challenge in dream research is the difficulty in quantitatively assessing dream content in a reproducible manner (Elce et al., 2021).

Dream reports, which document the content of dreams recalled by individuals, are essential to the study of dreaming. These reports offer insights into the connection between dreams and waking life (Blagrove et al., 2004) and have long been used as a medium to examine conscious experience during sleep (Nir and Tononi, 2010; Siclari et al., 2017). Thus, many efforts have focused on collecting and analyzing dream narratives or reports (Hall and Van de Castle, 1966; Hauri, 1975; Schredl, 2010).

A.2 Dream Interpretation Datasets

DreamBank is among the most widely used resources, comprising thousands of annotated dream narratives used in psychological and linguistic research (Domhoff and Schneider, 2008). In addition to public resources, Laureano and Calvo (2024), compiled a proprietary dataset of dreams from 20

patients, with each analyzed by multiple human psychoanalysts and GPT-4. However, manual annotation is time-intensive and typically requires trained human experts, which can hinder the reproducibility and scalability of dream research (Elce et al., 2021).

A.3 Automatic Dream Analysis

To address these limitations, researchers have increasingly explored the use of natural language processing (NLP) tools to automate the analysis of dream reports. They applied linguistic and computational techniques to analyze dream content. Niederhoffer et al. (2017) examined dream narratives using the Hall and Van de Castle (HVDC) framework, highlighting a higher prevalence of negative emotions, especially sadness in dreams. McNamara et al. (2019) introduced the Dream Content Analysis System (DCAS), which used AI to identify gender-related patterns in dream themes and their relation to mood. Elce et al. (2021) demonstrated that methods such as graph analysis, distributional semantics, and dictionary-based techniques can capture both semantic and structural properties of dream narratives.

Overall, these approaches often rely on models trained on general-purpose corpora such as Wikipedia (Nadeau et al., 2006; Razavi et al., 2014; Altszyler et al., 2017; Sanz et al., 2018; McNamara et al., 2019; Bertolini et al., 2023). However, there are debates regarding how closely dream reports resemble other general types of textual data (Kahan and LaBerge, 2011; Domhoff, 2017; Zheng and Schweickert, 2023). Some evidence suggests that the semantic characteristics of dream reports may diverge significantly from those found in waking narratives (Altszyler et al., 2017). If dream reports are indeed unique in structure and content, the effectiveness of using general-domain NLP models, especially in unsupervised settings, could be substantially limited (Bertolini et al., 2023).

A.4 LLM Dream Analysis

With the advancement of LLMs, there is growing interest in their use for automated dream analysis. Building on DreamBank, the DReAMy toolkit (Bertolini et al., 2024) offers an open-source framework that leverages multilingual LLMs to automatically annotate dream reports for emotions and characters based on HVDC framework. Blyler and Seligman (2024) used GPT-4 to generate personal narratives and streams of consciousness, while Juncker (2023) employed the model for psychological reflection and dream interpretation. Laureano and Calvo (2024) compared GPT-4's interpretations with those of human analysts, finding that while both were coherent, the AI displayed distinctive linguistic patterns. GPT-4 tended to use more semantic categories such as vision and health-related terms and fewer grammatical elements like impersonal pronouns. Using LIWC features and Naïve Bayes classification, they achieved 99% accuracy in differentiating between the two sources, indicating the significant differences between human-written and LLM-generated dream analysis. These approaches reduce reliance on manual dream annotation while also expose the gap between human and LLM in dream interpretation.

Summary While these efforts have advanced dream understanding, little attention has been devoted to *dream interpretation*, which seeks to derive symbolic, cultural, and contextual meaning from dream content.

Most publicly available datasets and studies on dream interpretation are centered on English and adopt linguistic, emotional, psychological, or biological perspectives to analyze dream narratives. Such approaches primarily depend on an individual's current condition and recent experiences. By contrast, in cultures such as China, dream interpretation is not only dependent on individuals, but also grounded in centuries of accumulated wisdom and collective observation. Ancient dream-interpretaion records like ZhougongJieMeng provide population-level generalizations, functioning as a form of prior knowledge or statistical distribution of interpretations across thousands of years. These traditions draw on systematic principles such as the Five Elements (metal, wood, water, fire, earth) and Yin–Yang to derive meaning, offering a perspective that extends beyond the individual to collective cultural experience.

To mitigate this gap, we collect a dream interpretation datasets from three culture sources including Arabic, Chinese and Western context, presented in two langauges (Arabic and English). Meanwhile, we examine current state-of-the-art LLMs in dream interpretations across cultures and improved it by continuous training.

B Dataset Curation

In this section, we elaborate how we curate the dataset from collecting raw dream-interpretation

pairs from three cultural sources, to converting them into four task formulations, to splitting them into training and evaluation benchmark.

B.1 Raw Data Sources

To construct our dataset, we first collected raw dream–interpretation pairs from a diverse set of cultural traditions and online repositories. The sources span three major cultural contexts: *Islamic*, *Western*, and *Chinese*. Within each culture, we curated data from multiple publicly available websites and pre-compiled datasets, ensuring broad coverage of interpretations across different schools and perspectives. Table 6 summarizes the raw data sources and their respective sizes. These raw collections serve as the foundation for subsequent task-specific transformations and benchmark construction.

B.1.1 Islamic Culture Corpora

Our dataset combines classical and modern sources of Islamic dream interpretation. Most of the content comes from traditional texts written by well-known scholars in books including **Al–Nabulsi**, **Al–Ihsaei**, **Al–Anbari**, **Ibn Sirin**, and **Ibn Shahin**. We cleaned the data and extracted dream-interpretation entries. Each entry typically consists of a dream symbol paired with an interpretation. When a dream symbol appeared in multiple sources with different interpretations, we retained all available explanations to capture diverse scholarly perspectives, reflecting the natural evolution of dream interpretation from traditional to modern views. After removing duplicates, the dataset includes a total of **5,568** entities.

B.1.2 Western Culture Corpora

Western culture corpora combine modern, classical, and community-based interpretations. It includes structured resources and digitized texts that cover a range of symbolic meanings and interpretation styles. The raw collection draws from four sources: *(i) Dictionary of Dreams*, 1,041 symbol-interpretation pairs; *(ii) Dream Dictionary*, 2,080 entries of a similar format; *(iii) DreamBook (Guanaco Format)*, 9,497 user written dreams paired with interpretations generated by a language model; and *(iv)* `myIslamicDream`[7], 96,404 symbolic entries presented in English. The combined raw count is 109,022 entries.

Content from `myislamicdream.com` was collected by a custom Python scraper. The script iterated over paginated tag pages at the pattern `/tags/{page}`, extracted internal links that ended with `.html`, deduplicated links, and then fetched each page. For each page, the parser retained only pages with at least one non-empty paragraph and a minimum visible text length, and it discarded short or empty pages. The crawler used standard request headers and added a one second delay between requests.

Cleaning and Consolidation After collection, we cleaned and consolidated to obtain unique pairs. We first removed empty English interpretations, and then normalized the interpretation text by deleting boilerplate lead-in phrases, such as "Refer to ...", removing parenthetical source notes starting with "(Provided ...)" and deleting footer style material appended at the end of some pages, for instance trailing text "People Who Read This Article Also Read" and embedded domain mentions like "(... www.xyz.org ...)". Afterwards, we excluded entries that were cross references (e.g., "See ...") rather than dream interpretations, and standardized whitespace by removing line breaks and collapsing multiple spaces. Finally, we dropped any residual empty interpretations that arose after text normalization and performed exact match de-duplication on the normalized interpretations to consolidate repeated content across sources.

When a symbol appeared in more than one source, we retained distinct interpretations that provided substantive content and removed non-informative duplicates. We collected **16,720** unique dream-interpretation pairs.

B.1.3 Chinese Cultural Corpora

The Chinese dataset includes traditional ZhouGong interpretations, structured categorical records, and contemporary astrological perspectives. The collection is drawn from three sources: *(i) Zhougong Dream Dictionary*, parsed from the Zhou Gong website[2], yielding 9,508 entries in Chinese; *(ii) Zhougongjiemeng Database*, reconstructed from a publicly available SQL export[3], containing 9,543 entries in Chinese; and *(iii) Your Chinese Astrology*, an English-language site presenting dream interpretations in a Chinese astrological framework[4], contributing 77 entries.

[2] `https://www.zgjm.org`

[3] `https://blog.csdn.net/jianghulangzhongshen/article/details/107043786`

[4] `https://www.yourchineseastrology.com`

Dataset	Data License	Lang	Raw Size	Size
tafsiralahlam.net (Al-Ahsa'i)[2]	Book by Al Ahsa'i (1826)	AR	441	424
tafsiralahlam.net (Ibn Shahin)[2]	Book by Ibn Shahin (1468)	AR	1,043	1,034
tafsiralahlam.net (Ibn Sirin)[2]	Book by Ibn Sireen (729)	AR	1,096	1,088
tafsiralahlam.net (Nabulsi)[2]	Book by Nabulsi (1731)	AR	2,291	2,291
alanbar_bary.com (Dr. Khaled Al-Anbari)[3]	Book by AlNbari (allow research use)	AR	733	731
Total Islamic Entries				**5,568**
Dictionary of Dreams (kaggle/manswad)[4]	Apache 2.0	EN	1,040	898
Dream Dictionary (kaggle/yuvrajsanghai)[5]	Apache 2.0	EN	2,080	2,041
Dreambook Guanaco Format (hf/n3rd0)[6]	GPLv3 (GNU General Public License v3)	EN	9,497	7,662
myislamicdream.com[7]	No License	EN	96,404	6,119
Total Western Entries				**16,728**
yourchineseastrology.com[8]	–	ZH	77	77
zgjmorg.com (Zhougong)[9]	Ancient Chinese book《周公解梦》	ZH	9,508	7,325
Zhougongjiemeng database[10]	CC 4.0 BY-SA	ZH	9,543	2,187
Total Chinese Entries				**9,589**
Total				**31,877**

Table 6: Raw source and size across three cultures. Size refers to numbers after preprocessing.

Cleaning and Consolidation Both the *Zhougong Dream Dictionary* and the *Zhougongjiemeng Database* consist of collections of HTML pages, each processed with a custom parser. During preprocessing, empty or redundant entries were removed, as many of these corresponded to boilerplate pages that merely linked to other content rather than providing substantive interpretations. Because both sources originate from the traditional Zhou Gong corpus, a substantial amount of content was duplicated across them. To consolidate overlapping material, we applied a semantic similarity model (`paraphrase-multilingual-MiniLM-L12-v2`) to compare entries. Pairs with a similarity score above 0.8 were treated as duplicates, and in such cases the longer, more detailed interpretation was retained. Both datasets also underwent additional text cleaning. Redundant lead-in phrases, such as "梦见……意味着什么？ and 阅读本文的人还阅读了", were removed using regular expressions to retain only the substantive interpretation content. After this deduplication and cleaning process, the resulting Chinese dataset contained 9,512 unique entries.

The *Your Chinese Astrology* source was processed with a custom parser, which cleaned dream titles by removing boilerplate prefixes (e.g., "Dream Meaning and Interpretation about/of") and extracted interpretation content. The resulting text was normalized and converted into a plain, consistently formatted representation suitable for further processing. This source contributed **77** entries in English. Therefore, there are **9,589** entries in total from Chinese culture.

B.2 Dream Symbol Category

We applied GPT-4o using prompt shown in Figure 8 to categorize dream symbols for all three cultures into 17 types including prophets and messengers; companions and righteous figures; people and social roles; body parts; animals; birds and insects; places and landmarks; natural elements and phenomena; tools and objects; food and drink; religious symbols and practices; emotional and psychological states; actions and events; time-related concepts; abstract or ambiguous symbols; and symbols related to death and the afterlife. Figure 2 shows the distribution of dream categories.

B.3 Four Task Formulation

Based on clean dream-interpretation pairs, we formulated four tasks including free-form question

[2] https://tafsiralahlam.net
[3] https://www.alanbary.com
[4] https://www.kaggle.com/datasets/manswad/dictionary-of-dreams
[5] https://www.kaggle.com/datasets/yuvrajsanghai/dream-dictionary
[6] https://huggingface.co/datasets/JosephusCheung/GuanacoDataset
[7] https://www.myislamicdream.com
[8] https://www.yourchineseastrology.com
[9] https://www.zgjmorg.com
[10] https://blog.csdn.net/jianghulangzhongshen/article/details/107043786

answering (QA), multi-choice question answering (MCQ), is it a *good or bad* dream (GB), and is this interpretation of the dream *true or false* (TF).

B.3.1 QA Conversion

Islamic Culture To extend the Islamic subset into a question–answer format, we generated culturally grounded questions based on the available symbol–interpretation pairs. For each dream symbol, multiple questions were formulated in five major variants of Arabic languages: Modern Standard Arabic, Gulf, Egyptian, Levantine, and Maghrebi. The prompts were designed to simulate realistic user queries, ranging from conversational expressions to more formal requests, thereby reflecting how speakers across regions might naturally inquire about dream meanings.

Crucially, the answers paired with these questions were always the original interpretations taken directly from the classical Islamic sources. No modifications or paraphrases were introduced: the interpretive content remains identical to that provided by the authoritative texts. The only variation lies in the phrasing of the questions, which differs between dialects.

This procedure produced a dialect-sensitive Arabic QA dataset in which linguistic diversity on the question side is matched with consistent, unaltered interpretations on the answer side. Such a design ensures that the dataset both preserves the symbolic integrity of the source material and provides a realistic testing ground for evaluating models under conditions of dialectal variation in user queries.

Western/Chinese Culture To enhance question diversity and cultural relevance, we used LLMs (e.g., `Gemini-1.5-pro`) to generate questions. Given a dream symbol and its associated interpretation, the models were prompted with culturally tailored templates, guiding them to produce natural questions aligned with the interpretive traditions of the source data. See prompts in Figures 6 and 7 in Appendix C.

For the Chinese culture subset, the prompt design emphasizes classical Chinese dream interpretation perspectives, such as 金木水火土, 阴阳 and 命格. Western culture templates highlight Western interpretation conventions, interpreting from astrological, zodiac. semiotic and psychological perspectives. Thus, the generated questions are aligned with the interpretive context of the source culture while ensuring diversity (multiple distinct questions per entry). This results in 16,720 QA samples for Western cultural resources and 9,589 QA samples from Chinese culture.

B.3.2 MCQ Generation

Islamic Culture We extend the Islamic dream interpretation pairs into MCQ format by combining each symbol with its original interpretation and a set of four distractors. The correct answer is always taken directly from the original sources, ensuring that the interpretive content remains unchanged. distractors are sampled from other symbols within the same semantic category, which makes them topically plausible while still incorrect. This category-based sampling increases the difficulty of the task and prevents trivial elimination strategies.

To generate the question text and options, we provide the symbol, its correct interpretation, and the four distractors to an LLM `Gemini-2.0-flash`, together with a structured prompt shown in Figure 9. The model is instructed to produce a single well-formed question in Modern Standard Arabic, accompanied by exactly five answer options (A–E) with the correct answer randomly positioned.

Western/Chinese Culture Similar to the procedure of Islamic culture, while we only generated MCQs for the test split of Western/Chinese cultures. 10% entries from the Western and Chinese subsets were independently sampled to preserve cultural balance. This results in 1,675 MCQ samples for western culture and 959 for Chinese.

Each sample's correct interpretation is paired with four distractors randomly drawn from other symbols within the same cultural subset. The five options (A–E) are shuffled, with the correct answer randomly positioned to avoid bias, and a post-processing step ensures uniform answer distribution. Once options are finalized, each entry is passed to `Gemini-1.5-pro` which generates a natural-language question contextualized to the respective cultural framework. The model is supplied with a system prompt that includes both the correct interpretation and its distractors, together with a curated set of culturally appropriate MCQ templates distinct for Western and Chinese dream theories. These templates guide the model to produce diverse and well-phrased questions while maintaining interpretive fidelity. Robustness of the pipeline is ensured via timeout protection, automatic retries, and multithreaded execution.

After expansion through distractor pairing and

question generation, the final MCQ dataset contains 2,634 Western samples. All MCQ entries remain traceable to the original dataset through the consistent `id` key. The final dataset includes the dream symbol, interpretation, multiple-choice options, the correct answer label, and the generated question. Detailed prompt structures for Western and Chinese MCQ generation are presented in Figures 12 and 13 respectively in Appendix D.

B.3.3 Good/Bad and True/False

In addition to free-form questions and straightforward MCQs, users often pose queries such as *I dreamed of a snake yesterday, is it a good sign?* or *I dreamed of a snake, and someone told me it means earning more money, do you think this is true?* To better reflect these real-world inquiry scenarios, we introduce two additional tasks: *(i)* determining whether a dream is a *good or bad* sign, and *(ii)* verifying whether a dream's meaning matches the user's assumption (*true or false*).

For Good/Bad task, we applied Gemini-1.5-Pro to classify each dream-interpretation entry as either good or bad. The input prompt includes the (dream, interpretation) pair, and the model is asked to return the label and brief explanation. This setup encourages the model to reflect on overall sentiment, cultural associations, and emotional tone. For example, a symbol associated with blessings or success would typically be labeled as good, while one linked to fear or misfortune might be labeled as bad.

The True/False task relies on contrastive reasoning. For each dream symbol, we pair its correct interpretation with a distractor interpretation drawn from the same category. Gemini-1.5-Pro is prompted to generate two short QA-style items: one based on the correct interpretation (true), and one based on the distractor (false), written to appear plausible despite being incorrect.

Note that we only extended these two tasks for Islamic culture test splits to evaluate model performance consistency across different inquiry styles. The same approach can be applied to produce more data.

B.4 Training and Evaluation Splits

We divide the dataset into training and test splits to enable reliable model development and evaluation. Within each cultural subset, the final processed symbol–interpretation pairs serve as the foundation for multiple task formats.

To prevent data leakage, the splits are created at the level of dream symbols rather than dream-interpretation pairs, ensuring that all interpretations of a given symbol remain within the same split. This procedure is applied independently to the Western, Chinese, and Islamic subsets in order to preserve cultural balance. Table 1 summarizes the dataset statistics for each task and the final splits of the training and test sets.

C Prompts for QA Generation

Figure 6 and Figure 7 present the detailed prompt used for curating the dataset for the QA task in Western and Chinese cultures. Gemini-1.5-pro is used to synthesize contextually appropriate questions from the given symbol-interpretation pair and predefined question templates.

D Prompts for MCQ Generation

In Figure 8, we present the prompt used for categorizing Arabic dream entities into symbolic categories. Figure 9, Figure 12 and Figure 13 are prompts used for generating the MCQ benchmark dataset in three cultures.

E Evaluation

E.1 Models

We include three commercial LLMs and their reasoning variants:

- `DeepSeek-v3.1-Terminus`
- `GPT5-2025-08-07`
- `Claude-Sonnet-4-20250514`

We also included 11 open-source models from `1B` to `8B`:

- `Qwen3-8B/1.7B` (chat and reasoning)
- `Llama-3.1-8B`
- `Deepseek-R1-Distill-Llama-8B`
- `Qwen2.5-Math-7B`
- `Deeepseek-R1-Distill-Qwen-7B`
- `Qwen2.5-1.5B-Instruct`
- `DeepSeek-R1-Distill-Qwen-1.5B`
- `Llama-3.2-1B`

All models used the same prompt templates.

You are a helpful assistant to interpret dreams from Western culture and theory. This is a symbol and interpretation pair in English from Western dream interpretation.

symbol: `{symbol}`
interpretation: `{interpretation}`

Use the following template of questions to generate an appropriate question for each symbol–interpretation pair. You may also generate original questions based on the given content. Start the question directly, without adding introductory text. Both the question and the interpretation should explicitly reference Western culture. Each symbol can result in between one and five questions depending on the interpretation. The output must be returned as a list of questions in JSON format without any additional text or explanation.

- I had a strange dream involving {symbol}. Could it be reflecting something deeper going on in my life?
- When {symbol} appears in a dream, could it point to something unresolved within me?
- I've been thinking about a dream I had with {symbol} in it, what might that say about my current state?
- Could dreaming about {symbol} be linked to recent emotions or stress I've been experiencing?
- What kind of message or warning might a dream about {symbol} be trying to send me?
- If I keep dreaming about {symbol}, could it be a sign that my mind is trying to process something?
- I dreamed about {symbol} last night, could this relate to something I haven't been paying attention to?
- In Western dream interpretation, what might the presence of {symbol} suggest about my inner world?
- What personal insight could I gain from a dream that prominently features {symbol}?
- Could {symbol} in my dream represent something I'm avoiding or afraid to confront?
- What does it typically mean if {symbol} shows up in a dream during a time of change?
- If I see {symbol} in a dream, might it reflect my mindset or relationships lately?
- I'm curious whether dreaming of {symbol} might connect to a personal challenge or transition.
- Is it possible that {symbol} showing up in my dream symbolizes a decision or dilemma I'm facing?
- What could my subconscious be working through if I keep seeing {symbol} in my dreams?
- What does seeing {symbol} in a dream mean?
- What does seeing {symbol} in a dream symbolize in Western culture?

Figure 6: Prompt used with Gemini 1.5 Pro to generate culturally grounded questions based on Western dream interpretations.

E.2 Inference Configuration

For all pretrained and fine-tuned open-source models (including both reasoning and non-reasoning models), we used identical inference settings: temperature = 0.0, greedy decoding only, and a maximum token limit of 4096. For all commercial models, we used the default parameters provided by the API.

E.3 Prompts for Evaluation

Figure 14, Figure 15, Figure 16 and Figure 17 present the detailed prompt template on the MCQ, Good/Bad classification, True/False classification and free-form QA tasks, respectively. Figure 18 shows the LLM-as-a-Judge prompt used to evaluate the correctness and sentiment of the model response for the QA task.

You are a helpful assistant to interpret dreams from Chinese culture and traditions. This is a symbol and interpretation pair in English from Chinese dream interpretation.
symbol: `{symbol}`
interpretation: `{interpretation}`
Use the following template of questions to generate an appropriate question for each symbol–interpretation pair. You may also generate original questions based on the given content. Each symbol can result in between one and five questions depending on the interpretation. The output must be returned as a list of questions in JSON format without any additional text or explanation.

- What does it mean to dream about {symbol}?
- Can you interpret the meaning of {symbol} appearing in my dream?
- I dreamed about {symbol}, what might this symbolize?
- What is the traditional Chinese interpretation of seeing {symbol} in a dream?
- What could be the deeper meaning of {symbol} in a dream?
- Could dreaming of {symbol} be a sign or omen? What does it represent?
- How should I understand the appearance of {symbol} in my dream?
- What are the possible meanings of dreaming of {symbol} repeatedly?
- I saw {symbol} in my dream and it felt significant, what could it mean?
- What does Chinese dream culture say about dreaming of {symbol}?
- My grandmother used to say dreams carry meanings, what could {symbol} mean if seen in a dream?
- In Chinese folk tradition, how is {symbol} interpreted in dreams?
- How would a traditional dream interpreter explain seeing {symbol}?
- Could the dream of {symbol} indicate something about my future or fortune?
- Is there a symbolic or spiritual meaning to dreaming about {symbol}?
- I'm facing stress lately and dreamed of {symbol}. Could it reflect something in my life?
- I had a peaceful dream with {symbol}, does it reflect emotional or spiritual harmony?
- After dreaming of {symbol}, I've felt uneasy. Could it signal a warning or imbalance?
- What might it mean if I see {symbol} in recurring dreams related to family or work?
- Could the appearance of {symbol} in my dream suggest anything about relationships or health?

Figure 7: Prompt used with Gemini 1.5 Pro to generate culturally grounded questions based on Chinese dream interpretations.

You are an expert in dream interpretation. You will be given a single dream entity, such as a word or phrase, and your task is to assign it to one of the following high-level categories:
1. Prophets and Messengers
2. Companions, Saints, and Righteous People
3. People and Social Roles
4. Body Parts
5. Animals
6. Birds and Insects
7. Places and Landmarks
8. Natural Elements and Phenomena
9. Tools and Physical Objects
10. Food and Drink
11. Religious Symbols and Practices
12. Emotions and Psychological States
13. Actions and Events
14. Time and Temporal Markers
15. Abstract or Ambiguous Symbols
16. Symbols Related to Death and the Afterlife
17. Uncategorized or Rare Symbols
Dream Entity: `{entity}`
Output Format (respond with only the category number):
Category: X

Figure 8: Prompt used with GPT-4o to categorize dream symbols into one of 17 symbolic categories.

You are an expert of dream interpretation. Below is the dream symbol between double ticks "`symbol`" with its correct interpretation:
"`{symbol}`": {interpretation}
And the following are a list of four wrong interpretations of the previous dream symbol:
`{wrong_interp}`
Using this information, write *one* multiple-choice question about the symbol "`{symbol}`" using the following rules: • All output should be in Modern Standard Arabic. • Only write one question. • Provide 5 options (A–E), with only one correct answer. • Randomize the position of the correct answer among A–E. • Format your output as valid JSON in the following structure:

```
{
  "question": "....",
  "options": [
    "A) ...",
    "B) ...",
    "C) ...",
    "D) ...",
    "E) ..."
  ],
  "correct_answer": "<correct_answer>"
  // either "A", "B", "C", "D", or "E"
}
```

Figure 9: Generation prompt for Arabic MCQs. The prompt specifies one correct interpretation and four distractors, and enforces a JSON output format in Modern Standard Arabic.

Review the following question according to five criteria: 1. Proper formatting 2. Correct language 3. Relevance to the symbol 4. Exactly one correct answer 5. Absence of theologically inappropriate content
If any issue is found, *fix it* and rewrite the question in the exact format below. If no issue exists, return the question unchanged.
Context: `{ctx}`
Question: `{s}`
• Reply **only** with the question, no comments or explanations. • Use this layout *verbatim*:

```
Question:
...
A) ...
B) ...
C) ...
D) ...
E) ...
Correct answer: X
```

Figure 10: Self-critique prompt (English translation). This iterative self-refinement step is inspired by the Self-Refine framework (Madaan et al., 2023).

Respond *only* with JSON in the form: `{"pass": true/false, "reasons": ["..."]}`
Evaluate the question below based on the same five criteria: 1. Formatting 2. Language 3. Symbol relevance 4. Valid and unambiguous correct answer 5. Theological safety
Context: `{ctx}`
Question: `{s}`

Figure 11: Judging prompt given to GPT-4.1-mini. This design follows the LLM-as-Judge methodology validated by Zheng *et al.* (Zheng et al., 2023).

You are a helpful assistant to interpret dreams from Western culture and theory. This is a set of symbol, interpretation, and options in English from Western dream interpretation.
symbol: `{symbol}`
interpretation: `{interpretation}`
possible options: `{options}`
The interpretation is the correct answer, while the rest of the options are distractors.
Use the following template of questions to generate an appropriate multiple-choice question. You can also create an original question based on the symbol, interpretation, and options:

- Which of the following best describes the meaning of dreaming about {symbol} in Western dream interpretation?
- According to Western dream theory, what does dreaming of {symbol} symbolize?
- In Western dream interpretation, what might it signify if you dream about {symbol}?
- Dreaming about {symbol} in Western culture typically indicates which of the following?
- Based on Western dream interpretation, what is the likely meaning of {symbol} in your dream?
- If I dream of {symbol}, what would it typically symbolize in Western theory?
- What could be the deeper meaning of {symbol} in a dream according to Western culture?
- According to Western dream interpretation, what is the meaning of {symbol} in a dream?

Figure 12: Prompt used with Gemini 1.5 Pro to generate MCQs based on Western dream interpretations.

You are a helpful assistant to interpret dreams from Chinese culture and traditions. This is a set of symbol, interpretation, and options in English from Chinese dream interpretation.
symbol: `{symbol}`
interpretation: `{interpretation}`
possible options: `{options}`
The interpretation is the correct answer, while the rest of the options are distractors.
Use the following template of questions to generate an appropriate multiple-choice question. You can also create an original question based on the symbol, interpretation, and options:

- Which of the following best describes the meaning of dreaming about {symbol} in Chinese dream interpretation?
- According to Chinese dream culture, what does dreaming of {symbol} symbolize?
- In Chinese dream interpretation, what might it signify if you dream about {symbol}?
- Dreaming about {symbol} in Chinese culture typically indicates which of the following?
- Based on Chinese dream interpretation, what is the likely meaning of {symbol} in your dream?

Figure 13: Prompt used with Gemini 1.5 Pro to generate MCQs based on Chinese dream interpretations.

System Prompt
You are a helpful assistant.

User Prompt
Question
`{question}`

Options
`{choices}`

==========
Answer the above question by only returning the option (A, B, C, D or E) without any further explanation.

Figure 14: Prompt template for evaluating models on the MCQ task.

System Prompt
You are a helpful assistant.

User Prompt
If I dream of symbol, is it a good or bad sign? Return "Good" or "Bad" without any further explanation.

Figure 15: Prompt template for evaluating models on the Good/Bad classification task.

System Prompt
You are a helpful assistant.

User Prompt
Question
`{question}`

Interpretation
`{interpretation}`

==========
Given the above question and its interpretation, determine the correctness of the interpretation. Return "True" or "False" without any further explanation.

Figure 16: Prompt template for evaluating models on the True/False classification task.

System Prompt
You are a helpful assistant.

User Prompt
`{question}`

Figure 17: Prompt template for evaluating models on the free form QA task.

System Prompt
You are a helpful assistant.

User Prompt
Question
`{question}`

Ground Truth Interpretation
`{gt_interpretation}`

Model Interpretation
`{model_interpretation}`

==========
Given the above question, ground truth interpretation and model interpretation.
You have three tasks:
1. Determine if the model interpretation is correct (an integer score between 1 to 5, 1 means entirely incorrect, 5 means entirely correct).
2. Determine the sentiment of ground truth interpretation (positive, negative).
3. Determine the sentiment of model interpretation (positive, negative).

Return the results in the following JSON format without any further explanation.

```
{
"correctness": an integer score between 1 to 5,
"gt_sentiment": "positive" or "negative",
"model_sentiment": "positive" or "negative"
}
```

Figure 18: Prompt template for evaluating QA task performance with LLM-as-a-Judge approach where GPT-5-mini serves as the judge.