Evaluating Cultural Knowledge and Reasoning in LLMs Through Persian Allusions

Anonymous ACL submission

Abstract

Allusion recognition—a task demanding contextual activation of cultural knowledge—serves as a critical test of large language models' (LLMs) ability to deploy stored information in open-ended, figurative settings. We introduce a framework for evaluating Persian literary allusions through (1) classical poetry annotations and (2) LLM-generated texts embedding allusions in novel contexts. By combining knowledge assessments, multiple-choice tasks, and open-ended recognition, we isolate whether failures stem from knowledge gaps or activation challenges. Evaluations across 11 LLMs reveal a critical disconnect: while models exhibit strong foundational knowledge and high multiple-choice accuracy, their performance drops significantly in open-ended settings, particularly for indirect references. Reasoning-optimized models generalize better to novel contexts, whereas distilled models show marked degradation in cultural reasoning. The gap underscores that LLMs' limitations arise not from missing knowledge but contextual recall failure—an inability to spontaneously activate cultural references without explicit cues. Our work positions allusion recognition as a benchmark for evaluating contextual knowledge deployment, urging training paradigms that bridge factual recall and culturally grounded reasoning.

1 Introduction

Allusion—the indirect reference to a culturally or historically significant entity—poses a unique challenge for both human readers and language models. Recognizing an allusion requires more than surface-level comprehension: it demands retrieving culturally situated background knowledge and applying it in a new context. This makes allusion recognition an ideal setting for evaluating model recall—the ability of a model to retrieve and deploy knowledge it already possesses, rather than merely generating plausible continuations.

Poet	حافظ (Hafez)
Theme	مذهبی (Religious)
Entities	خلیل (Khalil), سرد (Cold)", جان (Soul), آتش (Fire), رب (Lord)
Content	یا رب این آتش که در جان من است سرد کن آن‌سان که کردی بر خلیل (O Lord, cool this fire that is in my soul, as you did for Khalil.)
Allusion	حضرت ابراهیم (Prophet Abraham)
Description	گلستان کردن آتش توسط خداوند بر حضرت ابراهیم (The cooling of the fire by God upon Prophet Abraham)

Table 1: An example from the Persian Poems (PersPoems) Dataset.

While large language models (LLMs) excel in factual recall and generalization, their ability to activate knowledge in open-ended, figurative settings remains underexplored. Prior work has critiqued multiple-choice (MC) formats for LLM evaluation and advocated for open-ended tasks to assess reasoning (Myrzakhan et al., 2024), with some proposing MC conversions for efficiency (Zhang et al., 2024). However, these studies focus on numerical or logical tasks, leaving figurative language, particularly allusion, understudied. Unlike metaphors or idioms (Chakrabarty et al., 2022; Khoshtab et al., 2025; Rezaeimanesh et al., 2025), allusions are less formulaic and demand recognition of indirect, culturally embedded references, making them a robust test of cultural reasoning. Limited work, such as Han et al. (2025), explores allusion through a Chinese historical allusion dataset to fine-tune models for improved poetry generation.

We introduce an evaluation framework for allusion recognition in Persian literature, a tradition rich in symbolic and indirect references. We construct two datasets: (1) 200 annotated lines of classical Persian poetry (PersPoems), and (2) 75 LLM-generated allusive texts embedding the same allusions in novel out-of-distribution contexts.

These datasets isolate knowledge from memorization, probing LLMs' ability to recognize allusions in unfamiliar settings. Our dual framework combines multiple-choice tasks (isolating discriminative skills) and open-ended recognition (testing spontaneous knowledge activation), alongside a knowledge assessment of 127 core allusions.

Key findings reveal a critical disconnect: while LLMs exhibit factual knowledge of allusions (e.g., identifying referenced entities), they struggle to activate it in open-ended tasks, with performance dropping sharply compared to multiple-choice settings. Reasoning-optimized models generalize better across datasets, suggesting improved cultural reasoning integration. The disparity between knowledge and recognition reveals that recall failure, not lack of knowledge, hinders LLMs' interpretive capabilities, underscoring challenges in contextual knowledge application.

2 Allusion Datasets

Here we describe the two datasets used in our experiments: a collection of Persian poems containing allusions and a set of allusive LLM-generated texts created to test allusion recognition capabilities beyond potential training data memorization. Detailed information about dataset construction and annotation is provided in Appendix A.

2.1 Persian Poems (PersPoems)

To assess LLMs' ability to detect allusions, we build a dataset of 200 Persian poetry lines (PersPoems), annotated with "poet," "theme," "entities," and "description" (Table 1). Sourced from "Ganjoor," the dataset is validated by domain experts and spans six themes: Mythical-Historical, Religious, Mystical, Quranic, Romantic, and Other, based on works like (Shamisa, 1996, 2008). Allusion distribution is shown in Table 2.

Two expert annotators with Persian literature degrees independently identified allusions, providing explanations compared with online resources. Consensus annotations were finalized; otherwise, community-validated online data was used. This process achieves an 84.5% inter-annotator agreement, ensuring reliable ground truth for LLM evaluation.

2.2 LLM-Generated

To investigate memorization versus true understanding, we develop a dataset of 75 novel allusive texts generated by Claude 3.7 Sonnet, chosen for its strong grasp of Persian cultural elements. Using diverse allusions from our Persian poetry collection, we prompt the model to create passages embedding these allusions indirectly. This enables testing LLMs' ability to identify allusions in new contexts, emphasizing reasoning over training data retrieval.

Allusion Category	Count
Religious	112
Quranic	58
Mythical-Historical	31
Romantic	19
Mystical	12
Other	2

Table 2: Distribution of allusion categories in the PersPoems dataset, containing 200 poems, with some containing multiple types of allusions.

3 Evaluation Methodology

We detail our experimental setup for evaluating LLMs' allusion recognition, covering knowledge assessment, multiple-choice recognition, and open-ended recognition testing.

3.1 Knowledge Assessment

To establish LLMs' baseline knowledge of allusions, we compiled 127 distinct allusions from the Persian Poems dataset, including canonical and variant forms. We designed a protocol to evaluate LLMs across: (i) Source identification: Origin text, historical, or cultural context. (ii) Semantic explication: Literal and figurative meanings. (iii) Narrative components: Story arcs, key characters, and plot elements. (iv)Domain-specific details: (a) Quranic references: Surah, verse, and revelation context. (b) Hadith citations: Narrator and contextual meaning. (c) Mystical concepts: Philosophical frameworks, symbolism, and history.

An expert manually classified responses as (1) complete and accurate, (2) partial or imprecise, or (3) incorrect or absent knowledge. Partial or incorrect responses indicate knowledge gaps affecting recognition. This assessment helps determine if recognition failures arise from knowledge deficits or ineffective application in context. By quantifying each model's foundational knowledge of the allusions themselves, we can more precisely analyze whether recognition failures in later tasks stem from knowledge gaps or from inability to deploy

Model Name	Knowledge	PersPoems Dataset		LLM-Generated Dataset	
		Open-ended	Multi-choice	Open-ended	Multi-choice
Llama3.3 70B	93.7	47.5	90.5	41.3	92.0
Gemma3 27B	86.6	58.5	88.5	46.0	90.6
DeepSeek R1	93.7	**72.5**	91.0	**72.0**	**94.6**
DeepSeek V3	**96.1**	64.0	**92.5**	46.6	92.0
QwQ-32B	44.9	39.0	74.0	40.0	69.3
R1-distill Qwen-32B	52.7	22.5	79.5	20.0	81.3
Gemini-2.0 Flash	97.6	74.0	92.0	**72.0**	**96.0**
GPT-4o Mini	95.2	58.5	88.5	44.0	89.3
GPT-4.1	97.6	74.0	**93.5**	74.6	**96.0**
Claude 3.5-Sonnet	**100.0**	**80.5**	**93.5**	—	—
o1-mini	63.8	40.5	84.0	40.0	86.6

Table 3: Allusion recognition accuracy (%) on PersPoems and LLM-Generated Datasets across evaluation types. The "Knowledge" column reports the knowledge assessment accuracy. (The upper section lists open-source models; the lower shows closed-source models.) (Since Claude was used in the LLM-Generated dataset, we do not include its numbers.)

existing knowledge effectively when encountering allusions in context in different settings.

3.2 Multiple-Choice Recognition

We assess LLMs' ability to recognize allusions using a multiple-choice format, bridging knowledge possession and open-ended identification. For each sample in both datasets, LLMs received the text and five allusion options, selecting the correct one or "0" for no allusion, testing confident negative recognition. Distractors were chosen strategically: (i) For "Religious" or "Quranic" allusions, options were from the same category, leveraging their diversity. (ii) For "Mythical-Historical", "Mystical", "Romantic", or "Other" allusions, distractors were pooled from these related categories, sharing conceptual or narrative similarities.

This setup tests fine-grained discrimination between similar allusions. By isolating recognition from generation, we identify whether LLMs struggle with distinguishing allusions or retrieving them without cues, clarifying recognition mechanisms.

3.3 Open-ended Recognition

We evaluate LLMs' ability to autonomously recognize allusions without options, testing cultural knowledge retrieval and textual interpretation in a naturalistic setting. Using both datasets, we implemented a multi-stage protocol: (i) Allusion Detection: Models identify if a text contains an allusion, using subtle linguistic and contextual cues. (ii) Allusion Identification: For allusive texts, models specify the exact reference, requiring active knowledge retrieval. (iii) Thematic Integration: Models explain how the allusion enriches or transforms the text's meaning, assessing interpretive depth.

Outputs were manually validated for accurate recognition assessment. Designed prompts for each phase of evaluation are available in Appendix.

4 Results

We systematically evaluate 11 open- and closed-source LLMs for knowledge assessment and allusion recognition on both the Persian Allusive Poems dataset and the Generated Allusive Text dataset in open-ended and multiple-choice settings. Our evaluation includes six open-source models: Llama-3.3 (AI@Meta, 2024), Gemma-3 (Team, 2025a), DeepSeek-R1, DeepSeek-v3-chat, R1-distill-Qwen-32b (DeepSeek-AI, 2025) and QwQ-32b (Team, 2025b) and five close-source models: Claude-3.5-Sonnet (Anthropic, 2024), gpt-4o-mini, GPT-4.1 (OpenAI et al., 2024), o1-mini (OpenAI, 2024) and Gemini-2.0 Flash (Sundar Pichai and Kavukcuoglu, 2024). Table 3 presents the knowledge assessment and accuracy percentages for each model in both datasets, revealing notable patterns in allusion recognition capabilities.

Knowledge Assessment We first assess LLMs' foundational knowledge of Persian allusions according to Section 3.1. Most LLMs show strong proficiency, with seven models exceeding 90% accuracy. Open-source models like DeepSeek-V3-chat and Llama-3.3-70b-instruct perform comparably to closed-source models. However, o1-mini (63.8%), R1-distill-Qwen-32b (52.7%), and QwQ-32b (44.9%) exhibit notable performance drops.

Performance on PersPoems In PersPoems, LLMs perform strongly in multiple-choice recognition, with accuracies from 74.0% to 93.5%, most exceeding 88%. Claude-3.5-Sonnet and GPT-4.1 lead at 93.5%. In contrast, open-ended recognition, requiring independent allusion identification, shows lower performance. Claude-3.5-Sonnet tops at 80.5%, followed by GPT-4.1 and Gemini-2.0-Flash (74.0%), and DeepSeek-R1 (72.5%). The performance gap between formats is notable, with Claude-3.5-Sonnet dropping 13.0% and R1-distill-Qwen-32b exceeding 50.0%, highlighting challenges in unprompted allusion recognition.

Performance on LLM-Generated On the LLM-Generated dataset, designed to test novel allusions, models show strong multiple-choice performance, with top accuracies reaching 96.0%. Gemini-2.0-Flash and GPT-4.1 lead at 96.0%, with open-source models DeepSeek-V3-chat (94.6%) and Llama3.3-70B (92.0%) close behind. In open-ended recognition, GPT-4.1 scores highest at 74.6%, followed by DeepSeek-R1 and Gemini-2.0-Flash (both 72.0%). Performance gaps between formats remain significant, with the smallest gaps (21.4% and 22.6%) still indicating challenges in unprompted allusion recognition on novel content.

Cross-Dataset Performance Analysis Our analysis highlights model performance stability across datasets. RL-trained models like DeepSeek-R1, o1-mini, and QwQ-32b show remarkable consistency in open-ended settings, with minimal declines (0.5–1.0 percentage points) from PersPoems to LLM-Generated allusive text. In contrast, non-RL models like DeepSeek-v3-chat, GPT-4o-mini, and Gemma-3-27b-it exhibit significant drops (17.4, 14.5, and 12.5 percentage points, respectively), indicating RL training enhances generalization to novel contexts. Within the DeepSeek family, DeepSeek-R1's stability contrasts with DeepSeek-v3-chat's decline, underscoring RL's impact on reasoning. However, R1-distill-Qwen-32b, distilled from DeepSeek-R1, shows a 50.0% performance drop, suggesting distillation fails to transfer cultural knowledge and reasoning.

Qualitative Analysis of Performance Gaps To explore the multiple-choice versus open-ended performance gap, we analyzed cases where models showed knowledge and succeeded in multiple-choice tasks but failed in open-ended ones. We selected representative examples from both closed-source and open-source models to identify common patterns.

For example, DeepSeek-R1 exhibited the knowledge of the story of Yusuf and Zulaika (where women cut their hands instead of bergamot when seeing Yusuf) and correctly identified this allusion in multiple-choice settings across different examples. However, in the open-ended setting, it successfully identified the allusion only when explicit narrative elements were present. When presented with a poem that merely referenced cutting hands and bergamot without explicitly mentioning Yusuf and Zulaika, the model failed to make the connection. This pattern suggests that without explicit narrative markers or the prompting effect of multiple-choice options, models struggle to activate relevant knowledge frameworks. Similarly, Claude-3.5-Sonnet demonstrated a pattern common across multiple LLMs when encountering allusions derived from quotations, Quranic verses, or hadiths. While the model could readily identify such allusions when presented with options, it frequently failed to recognize these same references in open-ended scenarios, particularly when the allusive text lacked explicit markers or conventional framing devices that would signal quotation or reference.

5 Conclusions

We introduced two Persian allusion datasets—PersPoems and LLM-Generated—designed to probe LLMs' cultural reasoning beyond memorization. Using multiple-choice and open-ended formats, we assess both discriminative ability and spontaneous knowledge activation. We did our evaluation on six open-source and five closed-source LLMs. Most LLMs tested showed strong factual knowledge and high accuracy in multiple-choice questions, but performance drops significantly in the open-ended setting. This gap reveals that recall failure—not lack of knowledge—limits interpretive understanding. We also found that LLMs post-trained for reasoning using RL generalize better to our LLM-generated data, pointing to the need for training and evaluation methods that support contextual cultural inference.

6 Limitations

Our study on LLMs' allusion recognition capabilities has several limitations: it focuses solely on allusion rather than other figurative devices (metaphor, irony, symbolism); examines only Persian cultural and literary allusions, potentially missing cross-cultural patterns; relies on allusions generated by a single LLM which may introduce biases; and would benefit from a more comprehensive taxonomy of failure modes. Future research should investigate whether the observed gap between knowledge possession and application extends to other figurative language forms, conduct cross-linguistic comparative studies, employ diverse generation strategies, and develop detailed error pattern analyses to improve LLM reasoning for figurative language understanding.

References

AI@Meta. 2024. Llama 3 model card.

Anthropic. 2024. Claude 3.5 sonnet model card addendum.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Zhonghe Han, Jintao Liu, Yuanben Zhang, Lili Zhang, Lei Wang, Zequn Zhang, Zhihao Zhao, and Zhenyu Huang. 2025. Copiously quote classics: Improving chinese poetry generation with historical allusion knowledge. *Computer Speech Language*, 90:101708.

Paria Khoshtab, Danial Namazifard, Mostafa Masoudi, Ali Akhgary, Samin Mahdizadeh Sani, and Yadollah Yaghoobzadeh. 2025. Comparative study of multilingual idioms and similes in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8680–8698, Abu Dhabi, UAE. Association for Computational Linguistics.

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. 2024. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*.

OpenAI. 2024. Openai o1-mini: Advancing cost-efficient reasoning.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Sara Rezaeimanesh, Faezeh Hosseini, and Yadollah Yaghoobzadeh. 2025. Large language models for Persian-English idiom translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7974–7985, Albuquerque, New Mexico. Association for Computational Linguistics.

Sirus Shamisa. 1996. *The Dictionary of Allusions: Mythological, Narrative, Historical, and Religious Allusions in Persian Literature (Farhange Talmihat: Isharat-i asatiri, dastani, tarikhi, mazhabi dar adabiyat-i Farsi)*. Ferdowsi.

Sirus Shamisa. 2008. *The Dictionary of Refrences in Persian Literature: Myths, Traditions, Customs, Beliefs, Sciences, ... (farhange esharate adabiyate farsi: asatir, sonan, adab, eeteghadat, oloom and ...)*, volume 2. Mitra. In two volumes.

Demis Hassabis Sundar Pichai and Kuray Kavukcuoglu. 2024. Introducing gemini 2.0: our new ai model for the agentic era.

Gemma Team. 2025a. Gemma 3.

Qwen Team. 2025b. Qwq-32b: Embracing the power of reinforcement learning.

Ziyin Zhang, Zhaokun Jiang, Lizhen Xu, Hongkun Hao, and Rui Wang. 2024. Multiple-choice questions are efficient and robust llm evaluators. *Preprint*, arXiv:2405.11966.

A Extended Dataset Description

1.1 Persian Poems (PersPoems) Dataset Details

Our dataset encompasses poems with allusions distributed across six distinct thematic categories, providing a comprehensive representation of Persian literary tradition. These categories are derived from and expand upon the taxonomies presented in seminal works on Persian allusions, particularly Farhang-e Talmihat (Shamisa, 1996, Dictionary of Allusions). While this work primary classification focuses on mythological, fictional, historical, and religious references, the subsequent publications explore additional dimensions including allusions

to cultural customs, traditional sciences, and astronomical and medical beliefs of pre-modern Persia (Shamisa, 2008). Drawing upon this framework, we develop a more fine-grained classification system to better capture the nuanced cultural dimensions of Persian allusions.

These six categories are:

- Mythical-Historical
- Religious
- Mystical
- Quranic
- Romantic
- Other

The "Mythical-Historical" category comprises allusions rooted in historical events, legendary narratives, and Persian mythology, drawing from texts such as the Shahnameh. This aligns with (Shamisa, 1996) mythological and historical classifications but emphasizes the often inseparable nature of myth and history in Persian literature. "Religious" allusions reference stories of prophets, saints, and notable religious figures whose narratives form an integral part of religious heritage beyond explicit Quranic references.

The "Mystical" category contains references to Sufi concepts, philosophical ideals, and narratives about renowned gnostics or individuals with spiritual accomplishments—a dimension particularly prominent in Persian poetry yet deserving of distinct categorization from general religious content. "Quranic" allusions—separated from the broader "Religious" category due to their specific textual authority and prominence in Persian poetry—directly reference specific verses, expressions, or rhetorical structures from the Quran, as well as notable hadiths and quotations from Islamic figures.

The "Romantic" category encompasses references to canonical love narratives from Persian literature such as Leili and Majnun or Khosrow and Shirin. While these stories have historical or mythical origins, their exceptional prevalence and cultural significance in Persian poetry merits their classification as a distinct category of allusions, serving as archetypal frameworks through which poets explore themes of love and devotion. Finally, the "Other" category accommodates references to Persian cultural practices, societal conventions, and folkloric elements that do not fit neatly into the other categories but represent important aspects of Iranian cultural identity.

Examples in the dataset can belong to more than one thematic category, reflecting the multidimensional nature of many Persian allusions. The allusions in our dataset span a wide cultural spectrum, from famous romantic narratives to religious quotations and Quranic verses. This diversity makes the dataset particularly valuable for assessing LLMs' cultural knowledge and interpretive capabilities, as successful allusion recognition requires familiarity with concepts and figures from religious and mythological texts.

1.2 Annotation Process Details

To establish a robust human performance baseline and ensure annotation reliability, we employed a rigorous validation process. We provided the collected poems to two expert annotators with academic degrees in Persian literature and extensive teaching experience. Both of them are women and they are high school teachers. They did this work voluntarily with no payment to help to develop a Persian dataset for allusions. These annotators independently identified allusions present in each poem, providing brief (1-2 sentence) explanations. We then compared these annotations with the information gathered from online resources. When at least two annotators agreed on an identified allusion, this was established as the final annotation. In cases of disagreement, we defaulted to the allusion mentioned in the online resources, as these typically represent conclusions reached by either educational authorities or through collaborative community consensus.

1.3 LLM-Generated Dataset Creation

When evaluating LLMs on well-known poetic works, there exists a significant methodological concern: these texts may have been included in the models' training data, potentially resulting in performance based on memorization rather than genuine understanding. Authentic allusion recognition requires complex reasoning—identifying allusive markers, connecting contextual elements to external references, and accurately determining the specific allusion being invoked.

To address this limitation and assess LLMs' capability for genuine allusion recognition, we construct a novel dataset comprising 75 artificially generated allusive texts created using Claude 3.7 Sonnet. The generation process began with a careful analysis of our collected Persian poems to extract a diverse set of 75 representative allusions. This curated collection spans a spectrum of dif-

ficulty, from relatively straightforward and commonly recognized allusions to more sophisticated and nuanced references. We deliberately exclude extremely obscure allusions that would pose unreasonable challenges to both human experts and LLMs, ensuring the dataset serves as a fair and informative benchmark for evaluating allusion recognition capabilities.

For the generation protocol, we instruct Claude 3.7 Sonnet to produce creative literary passages that incorporate the selected allusions indirectly. The model is tasked with crafting texts that reference allusive elements through artistic and creative signals without explicitly naming the allusion itself. You can see the prompt for this part in Appendix Table 4.

Translated Prompt: Creative Literary Text Generation with Allusions

You must write a creative non-poetic literary text that artistically alludes to an ancient cultural-literary reference through indirect means.

Objectives:

- Create a literary text with elevated language containing layered allusions to ancient stories/myths/narratives
- Maintain harmony between textual atmosphere and the essence of the original allusion
- Develop new narratives preserving core concepts of the source material

Composition Guidelines:

Creative Process Framework

1 .**Essence Extraction**: Analyze core spirit and message of the allusion
2 .**Symbol Mapping**: Identify key symbols/colors/numbers from source material
3 .**Contextual Translation**: Reinterpret elements through contemporary metaphors
4 .**Narrative Weaving**: Construct emotionally resonant story architecture
5 .**Linguistic Enrichment**: Employ literary devices and evocative imagery

Output Specifications:

Composition Requirements

- 4-6 lines of text
- Indirect symbolic references (no explicit naming)
- Layered literary devices (metaphor/synecdoche/allegory)
- Self-contained narrative with ancient resonance
- Output contains **only** the generated text

Allusion and its Details: `{allusion}`

Table 4: Structured prompt for generating allusion-rich literary texts

Translated Prompt: Allusion Knowledge Test

I intend to present a literary allusion to you. Your task is to demonstrate whether you are truly familiar with the origin and source of this allusion.

Instructions:

- Identify the exact source (Quran, Hadith, historical story, myth, etc.)
- Explain the main meaning and concept
- Describe the full story with important details
- For Quranic references: Mention Surah & verse + context
- For Hadiths: Specify narrator & context
- For prophetic stories: Detail key events
- For mystical concepts: Explain origins & usage

Response Format:

Example Response

```
[
  {
    "title": "Short title",
    "full_explanation": "Detailed explanation..."
  }
]
```

Unfamiliar Response:

Null Response

```
[
  {
    "title": null,
    "full_explanation": null
  }
]
```

Allusion to analyze: `{allusion}`

Table 5: English version of the allusion knowledge assessment prompt with structured response formats

Translated Prompt: Allusion Detection Test

I will present you with a text that may contain indirect allusions to known stories, historical events, religious narratives, or literary works (verses, hadiths, religious tales, prophets, or mythological legends).

Your Tasks:

- Carefully read the text/poem and determine if an allusion exists
- Select the most accurate option from the 5 provided choices
- Respond with only the correct option number (1-5)

Analysis Method:

Step-by-Step Process

1 .**Identify Clues**: Detect special words, phrases, symbols, or imagery suggesting allusion
2 .**Evaluate Options**: Analyze all 5 choices against identified clues
3 .**Select Option**: Choose the most accurate match
4 .**Format Response**: Provide only the option number

Response Format:

Valid Responses

When allusion exists:

```
[{
  "selected_option": 3
}]
```

No allusion found:

```
[{
  "selected_option": 0
}]
```

Text to Analyze: `{text}`

Options:

1 .`{option_1}`
2 .`{option_2}`
3 .`{option_3}`
4 .`{option_4}`
5 .`{option_5}`

Table 6: Structured translated prompt for allusion detection in texts with multiple-choice evaluation system.

Translated Prompt: Allusion Detection Test

I intend to present you with a verse of poetry or text that may contain indirect allusions (talmīḥ) to recognized stories, historical events, religious narratives, or literary works.

Your Tasks:

- Carefully analyze the text to detect potential allusions
- Identify the referenced story/event/work if present
- Explain the allusion's significance within the text

Analysis Protocol:

Step-by-Step Evaluation

1 .**Detection**: Identify potential allusion markers in the text
2 .**Verification**: Confirm reference validity through contextual analysis
3 .**Interpretation**: Determine the allusion's semantic contribution

Response Schema:

JSON Output Specifications

When allusion exists:

```
[{
  "reference": "Identified story/event/work",
  "explanation": "Contextual significance analysis"
}]
```

No allusion detected:

```
[{
  "reference": null,
  "explanation": null
}]
```

Subject Text: `{text}`

Table 7: Structured translated allusion analysis prompt for open-ended evaluation