Evaluating Cultural Knowledge and Reasoning in LLMs Through Persian Allusions

Melika Nobakhtian[1] **Yadollah Yaghoobzadeh**[1,2] **Mohammad Taher Pilehvar**[3]
[1]Tehran Institute for Advanced Studies, Khatam University, Iran
[2]School of Electrical and Computer Engineering,
College of Engineering, University of Tehran, Tehran, Iran
[3]School of Computer Science and Informatics, Cardiff University
m.noubakhtian.5740@khatam.ac.ir y.yaghoobzadeh@ut.ac.ir
pilehvarmt@cardiff.ac.uk

Abstract

Allusion recognition—a task demanding contextual activation of cultural knowledge—serves as a critical test of 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 incorporating allusions in novel contexts. By combining knowledge assessments, multiple-choice tasks, and open-ended recognition, we analyze whether failures stem from knowledge gaps or activation challenges. Evaluations across eleven LLMs highlight a notable observation: models exhibit strong foundational knowledge and high multiple-choice accuracy, yet performance drops substantially in open-ended tasks, especially 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 from difficulties in spontaneously activating cultural references without explicit cues. We propose allusion recognition as a benchmark for contextual knowledge deployment, highlighting the need for training paradigms that bridge factual recall and culturally grounded reasoning. Our code, datasets and results are available at https://github.com/MelikaNobakhtian/Allusion.

1 Introduction

Allusion—the indirect reference to a culturally or historically significant entity—presents 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 novel contexts. This makes allusion recognition an ideal benchmark for evaluating model recall: the ability to access and deploy knowledge already stored, 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 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 the conversion of open-ended questions into an MC format for efficiency (Zhang et al., 2024). However, these investigations primarily target numerical or logical tasks, leaving figurative language, particularly allusions, largely 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 rigorous test of cultural reasoning. Despite the potential, limited work has focused on allusions as a primary task for evaluating LLMs (Han et al., 2025).

We introduce an evaluation framework for allusion recognition in Persian literature, a tradition rich in symbolic and indirect references. We construct two datasets: (1) a dataset of 200 annotated samples of classical Persian poetry (PersPoems), and (2) a dataset of 75 LLM-generated allusive

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 25725–25737
November 4-9, 2025 ©2025 Association for Computational Linguistics

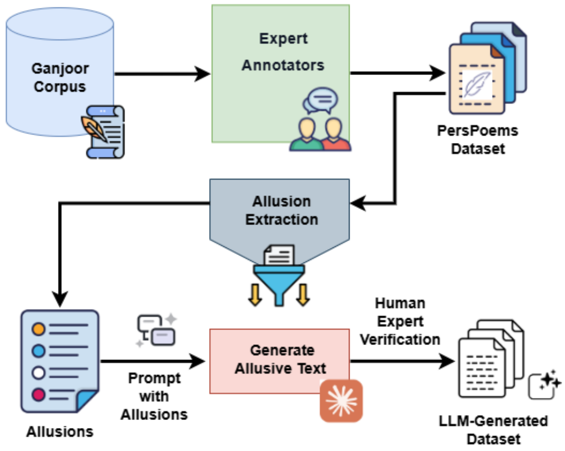

Figure 1: Data Construction Pipeline for Persian Poems (PersPoems) and LLM-Generated Datasets.

texts incorporating the same allusions in novel out-of-distribution contexts. These datasets disentangle knowledge from memorization, probing LLMs' ability to recognize allusions in unfamiliar settings. Our dual framework integrates MC tasks, which isolate discriminative skills, and open-ended recognition tasks, which test spontaneous knowledge activation, complemented by a knowledge assessment covering 127 core allusions.

Our results reveal a critical discrepancy: while LLMs demonstrate strong factual knowledge of allusions (e.g., correctly identifying referenced entities), they struggle to activate this knowledge in open-ended tasks, with performance dropping sharply compared to MC settings. Reasoning-optimized models show more consistent performance across datasets, indicating better integration of cultural reasoning. This gap between knowledge and recognition suggests that recall failure, and not knowledge deficiency, limits LLMs' interpretive capabilities, highlighting the challenges of applying contextual knowledge in figurative and culturally embedded settings.

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 construction and annotation procedures are provided in Appendix A.

2.1 Persian Poems (PersPoems)

To assess LLMs' ability to detect allusions, we build a dataset of 200 Persian poetry couplets (PersPoems), annotated with *poet*, *theme*, *entities*, and *description* (See Table 1 for an example from the PersPoems dataset). Sourced from Ganjoor[1], a free and open-source web collection of Persian poets' works, the dataset is validated by domain experts and spans six themes: *mythical-historical*, *religious*, *mystical*, *quranic*, *romantic*, and *other*, following the categorization used by Shamisa (1996) and Shamisa (2008). The distribution of allusions is shown in Table 2.

Two expert annotators, each with a college degree and teaching experience in Persian literature, independently identified allusions and provided explanations based on their domain knowledge. For quality control, these annotations were cross-checked against online resources, with community-validated online data used as the final adjudicator in cases of disagreement. This process achieved an inter-annotator agreement of 84.5%, ensuring a 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, selected for its strong grasp of Persian culture. Domain experts manually review all generated texts to ensure allusion accuracy. Initially, a larger pool of texts was generated, but only 75 were ultimately retained. Texts that were ambiguous, easily confused with unrelated allusions, or did not clearly reflect the intended allusive target were excluded. This filtering ensures that the final set of generated texts is both reliable and of high quality.

3 Evaluation

We evaluate LLMs on the allusion recognition task in three different settings: knowledge assessment, MC recognition, and open-ended recognition.

3.1 Knowledge Assessment

To establish LLMs' baseline knowledge of allusions, we compiled 127 distinct allusions from the Persian Poems dataset, including variant forms. The task is open-ended: given an allusion prompt, the model is asked to explain an allusion. We designed a protocol to evaluate LLMs across: (i)

[1] https://ganjoor.net/

Allusion Category	religious	quranic	mythical-historical	romantic	mystical	other
Count	112	58	31	19	12	2

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

Source identification: origin text, historical, or cultural context; (ii) Semantic explanation: 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; and (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. We only considered responses in the first category as demonstrating true knowledge. Partial and 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 understanding of allusions, we can more precisely analyze whether failures in subsequent tasks result from missing knowledge or from an inability to deploy existing knowledge effectively in novel contexts.

3.2 Multiple-Choice Recognition

We assessed LLMs' ability to recognize allusions using a multiple-choice format, bridging knowledge possession and open-ended identification. For each sample in both datasets, LLMs were presented with the text and five allusion options, selecting the correct one or a sixth option if no allusion was deemed to be present, thereby testing confident negative recognition. Allusion options were designed to distract the LLM from the correct choice.

Distractors are selected systematically: (i) For *religious* or *Quranic* allusions, they are drawn from the same category to leverage their diversity. (ii) For *mythical-historical*, *mystical*, *romantic*, or *other* allusions, they are pooled from related categories sharing conceptual or narrative similarities.

This setup evaluates fine-grained discrimination between similar allusions. By isolating recognition from generation, we can determine whether LLMs struggle with distinguishing closely related allusions or with retrieving them without explicit cues, providing insight into the mechanisms underlying allusion recognition.

3.3 Open-ended Recognition

We assessed LLMs' ability to recognize allusions with no candidate option available, evaluating cultural knowledge retrieval and textual interpretation in a naturalistic setting. Using both PersPoems and the LLM-Generated datasets, we implemented a multi-stage protocol: (i) Allusion detection: models determine whether a text contains an allusion, leveraging subtle linguistic and contextual cues; (ii) Allusion identification: for texts containing allusions, models specify the exact reference, requiring active knowledge retrieval; and (iii) Thematic integration: models explain how the allusion enriches or transforms the text's meaning, assessing interpretive depth. All outputs were manually validated to ensure quality. Detailed prompts for each evaluation stage are provided in the Appendix.

4 Results

We evaluated 11 open- and closed-source LLMs for knowledge assessment and allusion recognition on both the PersPoems dataset and the LLM-Generated 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.

Knowledge assessment. We first assess LLMs' knowledge of Persian allusions according to Section 3.1. Most models demonstrate strong proficiency , with seven exceeding 90% accuracy. Open-source models such as DeepSeek-V3-chat and Llama-3.3-70b-instruct perform comparably to closed-source models. In contrast, o1-mini

			PersPoems Dataset		LLM-Generated Dataset	
	Model Name	**Knowledge**	**Open-ended**	**Multi-choice**	**Open-ended**	**Multi-choice**
Open-source	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
Closed-source	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. Since Claude was used in the LLM-Generated dataset, we do not include it.

(63.8%), R1-distill-Qwen-32b (52.7%), and QwQ-32b (44.9%) exhibit notable performance drops.

Performance on PersPoems. On PersPoems, LLMs perform strongly in multiple-choice recognition, with accuracies ranging 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, yields lower performance. Claude-3.5-Sonnet achieves the highest at 80.5%, followed by GPT-4.1 and Gemini-2.0-Flash at 74.0%, and DeepSeek-R1 at 72.5%. The gap between multiple-choice and open-ended performance is notable: Claude-3.5-Sonnet drops 13.0%, while R1-distill-Qwen-32b experiences a decline exceeding 50.0%, highlighting challenges in spontaneous allusion recognition.

Performance on LLM-generated dataset. 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%), following closely. In open-ended recognition, GPT-4.1 scores highest at 74.6%, followed by DeepSeek-R1 and Gemini-2.0-Flash (both 72.0%). Despite these strong results, performance gaps between multiple-choice and open-ended formats remain substantial, with even the smallest gaps (21.4% and 22.6%) highlighting challenges in the spontaneous recognition of allusions in novel contexts.

Cross-dataset performance analysis. Our analysis reveals patterns of model performance stability across datasets. RL-trained models like DeepSeek-R1, o1-mini, and QwQ-32b exhibit remarkable consistency in open-ended settings, with minimal declines (<1%) from PersPoems to LLM-Generated text. In contrast, non-RL models like DeepSeek-v3-chat, GPT-4o-mini, and Gemma-3-27b-it show substantial drops (17.4%, 14.5%, and 12.5%, respectively), indicating RL training enhances generalization to novel contexts. Within the DeepSeek family, DeepSeek-R1's stability contrasts sharply with the decline of DeepSeek-v3-chat, further highlighting the effect of RL on reasoning. Notably, R1-distill-Qwen-32b, a distilled variant of DeepSeek-R1, experiences a 50% performance drop, indicating that distillation may fail to preserve cultural knowledge and related reasoning cabilities.

Qualitative analysis of performance gaps. To investigate the performance gap between multiple-choice and open-ended recognition, we analyzed instances where models demonstrated knowledge and succeeded in multiple-choice tasks but failed in open-ended ones. Representative examples from both closed- and open-source models reveal common patterns. For instance, DeepSeek-R1 possessed knowledge of the story of Yusuf and Zulaika (in which women cut their hands instead of bergamot upon seeing Yusuf) and correctly identified this allusion in multiple-choice tasks across different examples. However, in open-ended settings, the model only successfully recognized the allusion when explicit narrative elements were present. When presented with a poem that referenced cutting hands and bergamot without explicitly mentioning Yusuf and Zulaika, the model failed to make

Model	EM	Cohen's κ
Llama3.3	0.89	0.79
Gemma3 27B	0.91	0.82
DeepSeek R1	0.93	0.81
DeepSeek V3	0.93	0.84
QwQ-32B	0.86	0.70
R1-distill Qwen-32B	0.87	0.64
Gemini-2.0 Flash	0.87	0.61
GPT-4o Mini	0.89	0.78
GPT-4.1	0.94	0.83
Claude 3.5-Sonnet	0.91	0.66
o1-mini	0.91	0.81

Table 4: Alignment (Cohen's κ) and Exact Match (EM) between human evaluation and LLM-as-Judge (GPT-4o) for allusion recognition on the *PersPoem* dataset with access to ground truth (GT) allusions.

the connection. This pattern suggests that, in the absence of explicit narrative cues or the prompting effect of multiple-choice options, models struggle to activate relevant knowledge frameworks. Similarly, Claude-3.5-Sonnet exhibits a common pattern across many LLMs in allusion recognition when handling allusions derived from quotations, Quranic verses, or Hadiths. While the model readily identifies such allusions when options are provided, it frequently fails in open-ended scenarios, particularly when the allusive text lacks explicit markers or conventional framing devices that signal a quotation or reference.

An intriguing complementary pattern emerges in the opposite scenario: models with limited knowledge nonetheless achieve relatively high performance on multiple-choice tasks through surface-level lexical matching strategies rather than genuine comprehension. We provide a detailed analysis with representative examples in Appendix B.

LLM-as-Judge evaluation. To assess the viability of automated evaluation for allusion recognition in open-ended settings, we conducted experiments using GPT-4o as a judge model. We explored two scenarios: evaluation without ground truth and evaluation with ground truth access.

In the first scenario, where the judge model evaluates allusion recognition without access to correct answers, we observe significant discrepancies between evaluation metrics. While Exact Match (EM) scores appear high (0.81 for QwQ-32B and 0.80 for Gemini Flash), Cohen's κ values are considerably lower, with Gemini Flash achieving only 0.31 and QwQ-32B reaching 0.62. Given the poor performance of these two models in this setting, we did not evaluate additional models without reference answers. This disparity occurs because the judge frequently labels outputs with incorrect allusions as correct, artificially inflating EM values while the low κ scores reveal poor alignment with human evaluations when accounting for chance agreement.

When provided with ground truth allusions and their descriptions, the LLM-as-Judge shows marked improvement across all models, as shown in Table 4. Both EM and Cohen's κ values increase substantially, with models like Gemma 3B and DeepSeek V3 achieving κ values of 0.82 and 0.84, respectively, indicating very high inter-rater agreement. However, variability remains significant across different models, with κ values ranging from 0.61 for Gemini-2.0 Flash to 0.84 for DeepSeek V3.

To contextualize this variability, we conducted an additional experiment comparing two human annotators on Claude 3.5-Sonnet outputs, which had achieved a relatively lower κ value (0.66) in the LLM-as-Judge evaluation. The human-to-human comparison yielded an EM of 91.0% and a κ of 0.71, suggesting that some variability in LLM-as-Judge results may reflect inherent task subjectivity rather than tool unreliability. These findings demonstrate that while LLM-as-Judge without ground truth proves unreliable, providing ground truth transforms it into an effective evaluation tool that can serve as a scalable alternative to manual human evaluation for allusion recognition tasks.

5 Conclusions

We introduced two Persian allusion datasets, PersPoems and LLM-Generated datasets, designed to probe LLMs' cultural reasoning beyond memorization. Using multiple-choice and open-ended formats, we assessed both discriminative ability and spontaneous knowledge activation. We did our evaluation on six open-source and five closed-source LLMs. Most LLMs 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 (and 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 high school teachers and women. 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 information from online resources, which included community discussions on Ganjoor and other educational websites. This cross-checking process was necessary because we could not solely rely on these online resources due to their varying reliability and lack of availability for all poems.

Consensus was established through a clear adjudication process. 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 aforementioned online resources. These resources were used as the final point of reference because they typically represent conclusions reached by either educational authorities or through a collaborative community consensus over time. This method ensured that our final annotations were not only expert-validated but also aligned with broader, community-accepted interpretations.

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 difficulty, 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 5.

B High multiple-choice performance but limited knowledge

In the section 4, we focus on cases where LLMs possess knowledge of an allusion and succeeded in multiple-choice (MC) settings but fail in open-ended tasks. An intriguing complementary pattern emerges in the opposite scenario: models with limited foundational knowledge nonetheless achieve relatively high performance on MC tasks. What drives this apparent discrepancy?

Consider Qwen/QWQ-32B, which shows relatively low allusion knowledge (44.9 overall), yet achieves solid MC performance (74.0% on PersPoem and 69.3% on LLM-Generated). We analyzed several examples where the model did not exhibit knowledge of a particular allusion but still selected the correct answer in MC settings—while consistently failing in open-ended formats.

What enables the model to succeed in MC format despite lacking deeper understanding? A closer examination suggests that surface-level lexical similarities between the poem and the correct allusion option can guide selection. For instance, in the poem گفت آن یار کز او گشت سرِ دار بلند، جُرمش این بوَد که اسرار هویدا می‌کرد ("He said, that friend whose head was raised on the gallows, his only crime was that he revealed the secrets"), which alludes to the execution of Hallaj. Although QWQ lacked explicit knowledge of this allusion in the knowledge assessment, it correctly selected the answer in the MC task. The presence of the keyword "gallows" (دار) in both the poem and the allusion option likely served as a cue, enabling the model to make the correct connection.

This lexical-matching heuristic can extend beyond exact word overlap. For instance, considering the poem نماز در خم آن ابروان محرابی، کسی کند که به خون جگر طهارت دارد ("Only one who has purified with the blood of their heart can pray in the arch of those altar-like eyebrows"), the model correctly selects the allusion to the quote رکعتان فی عشق لایصح الوضوهما الا بدم ("Two rak'ahs of love, and their ablution is not accepted but with blood") in the MC setting, despite not having knowledge of the quote. Here, associations between semantically related terms—like "prayer" and "rak'ah", "ablution" and "purification", likely guided the model to the correct answer.

However, this strategy can also result in errors. When a model relies excessively on superficial similarity, it may select an allusion that merely resembles the poem, rather than the one intended. For example, in the poem یا چو درختم که به امر رسول، بیخ کشان آمدم اندر فلا ("Or like a tree that, by the Prophet's command, was uprooted and came into the wilderness"), the model incorrectly selects "The Prophet's Ascension" as the allusion, misled by surface-level keywords. The true allusion, lacking obvious lexical cues, goes unrecognized.

These examples indicate that in the absence of deep understanding, LLMs frequently default to pattern-matching strategies—leveraging surface or semantic overlap in multiple-choice tasks. While this can inflate apparent performance, it does not reflect genuine comprehension.

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 5: 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 6: 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 7: 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 8: Structured translated allusion analysis prompt for open-ended evaluation.

Example Instruction and Options

Task: 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 task:

1. Carefully read the text/poem and determine if an allusion exists.
2. Select the most accurate option from the 5 provided choices.
3. Respond with only the correct option number (1–5) in the JSON format below:

```
{
  "selected_option": 1
}
```

If no allusion is present, respond with:

```
{
  "selected_option": 0
}
```

Poem for analysis: *In that drunkenness, I cut an orange / Now the orange is real, and my hand is weary*

Available options:

1. Bringing out the camel from the mountain by Prophet Saleh
2. The story of Bal'am Ba'ura
3. The story of cutting hands when seeing Prophet Joseph
4. The Day of Judgment
5. The Holy Spirit and the Messiah

Table 9: English translated example of the allusion recognition task with poem and multiple-choice options.