GhazalBench: Evaluating LLM Understanding and Canonical Surface-Form Access in Persian Ghazals

Anonymous ACL submission

Abstract

Persian poetry plays an active role in Iranian cultural practice, where verses by canonical poets such as Hafez are frequently quoted, paraphrased, or completed from partial cues. Supporting such interactions requires language models to engage not only with poetic meaning but also with culturally canonical surface form. We introduce GhazalBench, a benchmark for evaluating how large language models (LLMs) interact with Persian ghazals under usage-grounded conditions. Unlike prior work that primarily studies memorization as a liability, GhazalBench examines settings where access to exact surface form is functionally important for culturally grounded interaction. The benchmark evaluates two complementary abilities: poem-to-prose understanding and canonical surface-form access under varying semantic and lexical cues. Across several proprietary and open-weight multilingual LLMs, we observe a consistent dissociation: models generally capture poetic meaning but struggle to produce exact verse completions in open-ended settings, while recognition-based settings substantially reduce this gap. Parallel experiments on English sonnets show markedly stronger completion performance, suggesting that these limitations are tied more to differences in training exposure than to inherent architectural constraints. Our findings highlight the need for evaluation frameworks that jointly assess meaning, form, and cue-dependent access to culturally significant texts. GhazalBench is available at https://anonymous.4open.science/r/GhazalBench/.

1 Introduction

Large language models (LLMs) are increasingly deployed in multilingual and culturally diverse settings, where successful interaction requires more than general linguistic fluency (Kim et al., 2025). Beyond grammatical correctness and semantic adequacy, models are often expected to engage with *culturally grounded language practices*—forms of expression embedded in shared history, tradition, and collective memory (Hershcovich et al., 2022; Joshi et al., 2020). While prior work on LLM memorization has predominantly framed verbatim reproduction as a risk—highlighting concerns such as privacy leakage and unintended reproduction of training data (Carlini et al., 2021, 2022)—there exist domains in which access to exact surface form is not a failure mode but a functional requirement. In particular, interaction with culturally canonical texts often relies on the ability to recognize, complete, and reproduce fixed expressions under partial or indirect cues. Evaluating whether LLMs can support such interactions remains an open challenge.

Classical Persian poetry provides a particularly salient example of this setting. Poets such as Hafez are enduring cultural references in Persian-speaking societies, especially in Iran (Foltz, 2016). Verses from Hafez's ghazals are routinely quoted, partially recited, paraphrased, or alluded to in everyday conversation and social rituals (Ordoubadian et al., 2006). These interactions are often triggered by fragmentary or indirect cues and rely on shared familiarity with canonical poetic form. As a result, effective participation requires not only semantic understanding, but also reliable access to the *exact surface form* of culturally canonical text.

Existing evaluations of LLMs on poetry have largely emphasized creative generation, stylistic imitation, or aesthetic quality (Weatherby et al., 2024; Walsh et al., 2024b; Heuser, 2025). While informative, such settings do not capture how poetry functions as a shared cultural resource. Consequently, the complementary question of when access to exact surface form is *functionally required*—and how models access canonical forms under realistic usage conditions—remains largely unexplored.

In this work, we introduce **GhazalBench**, a benchmark for evaluating LLM interaction with Persian ghazals of Hafez through *usage-grounded*

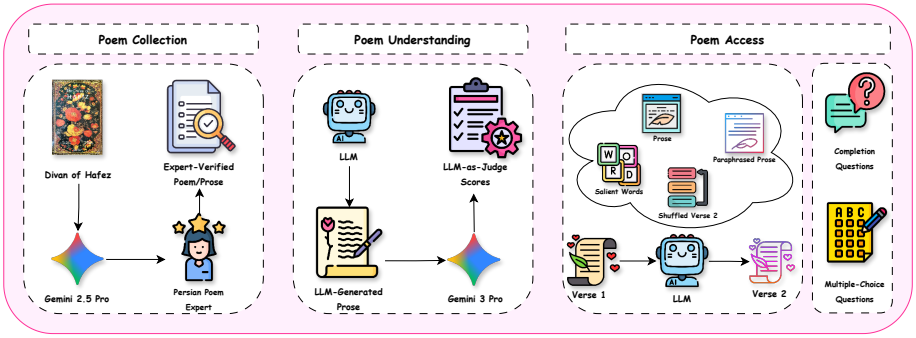

Figure 1: Pipeline illustrating the steps for poem collection, prose generation / evaluation, and assessment of second-verse access (completion or recognition) given the first verse and contextual cues.

diagnostic scenarios. Rather than asking whether models can generate poetry, we ask whether they can engage with poetry in ways that reflect how it is actually used by speakers. Our evaluation spans two complementary dimensions: (i) *poem-to-prose understanding*, where models produce concise, neutral Persian paraphrases of ghazal couplets, and (ii) *canonical surface-form access*, where models generate or identify missing verses under varying semantic and lexical cues. By combining completion-based and recognition-based settings, these scenarios are inspired by established distinctions between recall and recognition in cognitive psychology (Tulving et al., 1972; Tulving and Thomson, 1973), enabling us to disentangle semantic understanding from access to exact form.

Experiments on both proprietary and open-weight multilingual LLMs reveal a consistent pattern. While models demonstrate strong semantic understanding through accurate prose paraphrasing, they struggle with tasks requiring production of the exact surface form of poetry, particularly in completion-based settings. Performance improves substantially in recognition-based settings, revealing a strong gap between recognition and completion. Cross-lingual comparisons with English sonnets further show that this limitation is amplified in Persian, suggesting that access to canonical surface form depends strongly on training signal density rather than reflecting an inherent architectural constraint.

This paper makes the following contributions: (i) We introduce GhazalBench, a benchmark for evaluating LLM interaction with culturally canonical Persian poetry under realistic usage scenarios. (ii) We propose a set of diagnostic evaluation scenarios that distinguish semantic understanding from access to canonical surface form, grounded in recall–recognition theory. (iii) We show a systematic dissociation between prose-level understanding and completion-based access to canonical poetic form, revealing a strong gap between recognition and completion. (iv) We provide cross-lingual evidence suggesting that access to canonical surface form depends in part on training exposure, with substantially stronger completion performance observed for English sonnets than Persian ghazals.

2 Related Work

Memorization and Surface-Form Access in LLMs Memorization in LLMs has been studied primarily through the lens of privacy leakage, unintended reproduction of training data, and overfitting (Carlini et al., 2021, 2022). In parallel, retrieval-augmented generation (RAG) has been proposed as an alternative to parametric memorization by providing access to external documents at inference time (Lewis et al., 2020; Izacard et al., 2022). These lines of work largely focus on factual knowledge and document-oriented retrieval.

More recently, several studies have examined memorization of culturally canonical texts. Zhang et al. (2024) analyze verbatim memorization of English and French literature, while Attia et al. (2026) and Mensa et al. (2025) study proverb completion in Arabic and Italian. In poetry, D'Souza and Mimno (2023) evaluate poem completion under different prompt cues, and Chen et al. (2025); Walsh et al. (2024a) investigate access to canonical poems using partial context or surface overlap.

While these studies quantify memorization or completion performance, they primarily focus on whether models have retained culturally salient texts and how such retention varies across languages and domains. In contrast, our work is motivated by a different question: not whether models

memorize canonical texts, but whether they can access canonical surface forms under realistic interaction scenarios. We therefore evaluate both semantic understanding and surface-form access through usage-grounded tasks inspired by everyday interactions with Persian poetry, while systematically varying the semantic and lexical cues available to the model. This allows us to study the relationship between meaning, cueing conditions, and access to exact poetic form.

Poetry Understanding and Evaluation Prior work has evaluated LLMs' understanding and explanation of poetry using automatic metrics, LLM-as-Judge frameworks, and human assessment (Ma et al., 2025). Benchmarks have examined poetic form understanding in English sonnets (Walsh et al., 2024a) and semantic understanding in multilingual settings, including Arabic poetry (Al Ghallabi et al., 2025). While these studies provide structured evaluations of poetic understanding, they do not investigate access to canonical poetic form under varying semantic and lexical cues.

Poetry Generation A substantial literature has explored poetry generation with LLMs, focusing on style, rhyme, meter, and creativity across languages (Walsh et al., 2024b; Chen et al., 2024; Hernández Martínez, 2025). In contrast, our work is not concerned with generating new poetry, but with whether models can access existing canonical poetry under realistic scenarios.

Existing Persian Benchmarks Several recent benchmarks evaluate LLMs on Persian language understanding, cultural knowledge, and literary content. Examples include literary comprehension in Khayyam Challenge (Ghahroodi et al., 2024), allusion understanding in PersPoems (Nobakhtian et al., 2025), creativity evaluation in Persian literary text generation (Tourajmehr et al., 2025), and proverb understanding and generation (Hosseini et al., 2026; Kalhor and Bahrak, 2026). Our work complements these efforts by focusing on verse-level interaction with culturally canonical poetry and evaluating access to canonical surface form under realistic usage scenarios.

3 Benchmark Construction

To construct **GhazalBench**, we extract the first 100 ghazals of Hafez from the *Divan of Hafez* (Shirazi, 2009). The ordering of ghazals in the *Divan* is not based on popularity, chronology, or literary significance; instead, ghazals are arranged alphabetically according to the final letter of their rhyme. A ghazal is a classical poetic form composed of thematically linked but structurally independent couplets, each consisting of two verses with fixed meter and rhyme (de Bruijn, 2000). Appendix A provides an overview of this poetic form.

Each ghazal is collected couplet by couplet together with a corresponding prose explanation. To obtain this information, we query Gemini 2.5 Pro (Comanici et al., 2025), a model with strong Persian capabilities,[1] to extract the ghazal number, couplet number, couplet text, and prose explanation.

All extracted outputs are manually verified against the original printed edition by authors proficient in Persian poetry, and any transcription errors are corrected. In addition, the poetic meter of each ghazal is obtained from Ganjoor, a widely used repository of Persian poetry.[2]

For the experiments reported in this paper, we use a subset of 50 ghazals comprising 438 couplets. Detailed benchmark statistics, including dataset size, task distribution, and prompt characteristics, are reported in Appendix B. Figure 1 presents an overview of the benchmark construction and evaluation pipeline.

We focus on Hafez because his ghazals constitute one of the most widely shared and culturally canonical poetic corpora in Persian, making them a natural testbed for studying access to culturally entrenched surface forms.

3.1 Poem-to-Prose Understanding

To assess LLMs' semantic understanding of Persian ghazals, we evaluate their ability to paraphrase individual couplets into natural, standard Persian prose. Each model is given a single couplet and asked to produce a concise prose rendering that faithfully conveys its meaning. To guide output style, we provide two example paraphrases and instruct models to avoid lengthy explanations, interpretive commentary, and complex syntactic constructions. The goal is to elicit neutral and accessible paraphrases that align stylistically with the reference prose explanations used in the benchmark. Importantly, because Classical and Modern Persian remain largely mutually intelligible, this task is better viewed as an *intralingual paraphrasing* problem than a translation task.

[1] https://mcinext-mizan-llm-leaderboard.hf.space/
[2] https://ganjoor.net/

To examine the generated prose, we report automatic similarity metrics between generated and reference prose. Specifically, we compute BERTScore F1 using XLM-RoBERTa-large (Conneau et al., 2020), chrF++ (Popović, 2015), and BLEU-4 (Papineni et al., 2002). These metrics provide complementary measures of semantic and lexical similarity. Additionally, we provide a judge-based evaluation using Gemini 3 Pro in Appendix C.

3.2 Canonical Surface-Form Access

To assess models' access to canonical poetic form, we evaluate their ability to generate or identify the second verse of a couplet when provided with its first verse. This setting reflects common quotation-based interactions in Persian, where speakers rely on shared familiarity with canonical poetry. Our evaluation spans a set of diagnostic scenarios that vary the type and strength of semantic and lexical cues available to the model. By comparing completion-based and recognition-based settings, we investigate how access to canonical surface form depends on cue availability and task formulation.

These scenarios are intentionally diagnostic rather than naturalistic. Rather than simulating everyday human recall, they are designed to probe the accessibility of canonical poetic forms under controlled cueing conditions. Differences in performance across scenarios provide insight into the relationship between semantic understanding, lexical cues, and access to exact surface form. The following sections describe each scenario in detail.

3.2.1 Canonical Continuation

This scenario models a common quotation-based interaction in which a speaker recites the first verse of a well-known couplet and expects the listener to complete it or recognize it among several alternatives. Each model is provided with the **poet's name** (Hafez) and the first verse of the couplet. We evaluate performance under two settings: *completion* and *multiple-choice recognition*.

In the completion setting, the model generates the second verse directly, requiring open-ended access to the canonical continuation. In the recognition setting, the model selects the correct continuation from a fixed set of candidates.

For the multiple-choice setting, we include two carefully constructed distractors in addition to the correct answer. One distractor is drawn from a different couplet within the *same ghazal*, while the other is drawn from a couplet in a *different ghazal* in the dataset. To prevent trivial elimination based on form or meaning, all distractors are constrained to share the same poetic meter as the correct verse—a property automatically satisfied for distractors drawn from the same ghazal, since each ghazal uses a single meter. Among verses with matching meter, we select the distractor with the highest semantic similarity to the correct verse, measured using embeddings from the *paraphrase-mpnet-base-v2* model (Reimers and Gurevych, 2019) and cosine similarity.

By jointly controlling for meter and semantic proximity, this setup probes access to canonical surface form rather than reliance on rhythmic patterns or coarse semantic cues. Answer options are randomly shuffled for each instance to mitigate positional bias (Pezeshkpour and Hruschka, 2024).

3.2.2 Prose-Mediated Access

In this scenario, models are provided with the poet's name together with a prose explanation of the target couplet. We evaluate performance using the same *completion* and *multiple-choice recognition* settings as in Section 3.2.1.

This setting examines whether semantic information can facilitate access to canonical poetic form. By comparing completion and recognition performance under meaning-based prompts, we assess the extent to which models can generate or identify the target verse from its semantic content rather than from direct lexical cues.

3.2.3 Prose-Paraphrase Robustness

This scenario examines whether access to canonical poetic form is mediated by semantic content or by surface-level lexical overlap between the prose explanation and the target verse. We create a robustness condition by paraphrasing each original prose explanation using Gemini 3 Flash (Google LLC, 2025), instructing the model to preserve meaning while substantially altering vocabulary, phrasing, and sentence structure.

To assess paraphrase quality, we compute BERTScore F1 to measure semantic equivalence and Self-BLEU (Post, 2018) to quantify lexical divergence from the original explanations. BERTScore is computed at the sentence level and averaged across the dataset, while Self-BLEU is computed at the corpus level. The resulting BERTScore F1 of 0.891 indicates high semantic fidelity, while the Self-BLEU score of 1.574 reflects

minimal surface overlap.

Models are then evaluated using these paraphrased explanations under the same settings as Section 3.2.2. Comparing performance between original and paraphrased explanations allows us to assess whether access to canonical form is driven primarily by underlying semantic representations or by lexical overlap between the cue and the target verse.

3.2.4 Fragment-Conditioned Access

This scenario evaluates whether models can access canonical poetic forms when provided with minimal lexical cues. In addition to the poet's name and the first verse, each model is given two salient words drawn from the second verse and asked to produce the complete verse.

Salient words are selected using Gemini 3 Flash, which is instructed to identify two words that are maximally informative and memorable—terms that a human might plausibly remember even when the remainder of the verse is forgotten. This yields fragmentary prompts that preserve limited lexical identity while omitting syntactic structure and most semantic content.

We evaluate this scenario using completion only, as providing salient words would make recognition-based settings largely trivial. This condition probes whether sparse lexical fragments can facilitate access to canonical poetic form.

3.2.5 Shuffled-Form Access

This scenario probes whether models maintain order-sensitive representations of canonical poetic text by isolating sensitivity to word order independently of lexical content.

In the completion setting, models are provided with the poet's name, the first verse, and a word-shuffled version of the target second verse, and are asked to reconstruct the verse in its canonical order. This evaluates whether models can recover canonical structure when lexical identity is preserved but word order is disrupted.

For the recognition setting, we follow the setup of Section 3.2.1 but restrict the task to a binary choice. The correct canonical verse is presented alongside a distractor consisting of a word-shuffled version of the same verse. By controlling for lexical identity and semantic content, this design isolates sensitivity to canonical ordering.

Success in the shuffled-form completion setting should not be interpreted solely as evidence of access to canonical surface form. Because lexical identity is fully preserved, models may also benefit from general linguistic knowledge and sensitivity to syntactic regularities. The task therefore probes the interaction between canonical form knowledge and broader language modeling capabilities.

4 Experimental Setup and Results

4.1 Setup

Language Models for Evaluation We evaluate seven state-of-the-art multilingual LLMs on GhazalBench: GPT-5.2 (OpenAI, 2025) and Claude Sonnet 4.5 (Anthropic, 2025), two proprietary models, and DeepSeek V3.2 (Liu et al., 2025), DeepSeek R1 (DeepSeek-AI, 2025), Gemma 3 27B Instruct, Gemma 3 12B Instruct (Gemma Team and Google DeepMind, 2025), and Llama 4 Scout (AI, 2025), which are open-weight models. These models are selected to represent both closed- and open-weight paradigms, enabling a comparative analysis of their capabilities in Persian poetic understanding and memorization.

Prompts For all prompts used in both benchmark construction and model evaluation, we employ zero-shot prompting and run each prompt independently, without enabling web search, to ensure that the models rely solely on their internal knowledge. Moreover, as the focus of our study is accuracy rather than creativity, we set the temperature to 0 and the top-p value to 1. Examples of each prompt used in the study are provided in Appendix D.

4.2 Main Results

Poem Understanding Table 1 presents corpus-level automatic evaluation scores for the generated prose. We exclude ROUGE-L, as its scores are consistently zero—a result expected for poem-to-prose paraphrasing, where semantic adequacy rather than surface overlap is the primary objective. Claude Sonnet 4.5 outperforms the open-weight models across all reported metrics.

The relatively small gap in BERTScore F1 indicates that all models preserve high-level semantic content reasonably well, with Claude showing a modest advantage. In contrast, larger differences in chrF++ and BLEU-4 suggest that Claude's outputs exhibit stronger lexical and phrase-level alignment with the reference prose. For reference, the BLEU-4 score between the original poetic couplets and the reference prose is only 4.6, confirming that low

BLEU is inherent to the poem-to-prose transformation and not a sign of poor model quality.

Model	BERTScore F1	chrF++	BLEU-4
GPT-5.2	.91	32.5	6.7
Sonnet 4.5	**.93**	**37.8**	**12.3**
DeepSeek V3.2	.92	33.0	8.0
DeepSeek R1	.91	32.0	7.7
Gemma3 27B	.91	30.5	6.2
Gemma3 12B	.91	28.2	5.0
Llama4	.91	29.5	6.4

Table 1: Corpus-level automatic evaluation metrics for LLM-generated prose of Hafez's ghazals.

Evaluation of Verse Completion To evaluate completion-based access to canonical poetic form, we first apply Persian-specific text normalization, including character normalization (e.g., Arabic 'ي' → 'ی', 'ك' → 'ک'), removal of diacritics, zero-width non-joiners, and kashida characters, and whitespace normalization.

After normalization, a prediction is classified as an *exact match* if its Levenshtein edit distance from the reference verse is at most 1, 2, or 3 characters for verses shorter than 40, between 40 and 79, or 80+ characters, respectively. This criterion allows for minor typographical variation while requiring near-verbatim reproduction of the target verse.

For predictions not classified as exact matches, we compute Character Error Rate (CER), defined as the edit distance normalized by the length of the reference verse. Predictions with CER ≤ 0.2 are labeled *partial matches*, while all remaining outputs are classified as *no match*. This threshold is selected after empirically comparing cutoff values of 0.1, 0.15, 0.2, and 0.25 and choosing the one that best matched human judgments of "near-correct" completion.

Table 2 reports completion performance using exact and partial matches across different cueing conditions. Example outputs and common failure modes are provided in Appendix E.

As shown in Table 2, prose explanations provide modest improvements in completion performance for Claude Sonnet 4.5, GPT-5.2, and DeepSeek V3.2. However, these gains diminish when the prose is paraphrased, suggesting that models benefit more from lexical overlap with the cue than from access to its abstract semantic content alone. Similarly, providing two salient words from the target verse yields little improvement and occasionally degrades performance, indicating that sparse lexical fragments are generally insufficient to support access to canonical poetic form.

In contrast, supplying a word-shuffled version of the target verse consistently produces the largest gains for most models. Because lexical identity is fully preserved while syntactic order is disrupted, this result suggests that access to canonical form is strongly mediated by lexical content. The pattern is particularly pronounced in open-weight models, which benefit substantially from shuffled-form cues despite exhibiting near-zero completion performance under weaker cueing conditions.

Comparing the two DeepSeek models, DeepSeek R1 achieves slightly higher completion rates than V3.2 across most conditions, with the largest difference in the salient□word setting. This pattern may reflect R1's reasoning□oriented architecture: when provided with sparse lexical fragments, a model that can reason explicitly over meter, rhyme, and semantic associations may be better able to reconstruct a plausible continuation than a model that relies on more direct completion.

Across all settings, GPT-5.2 substantially outperforms the other models, achieving completion rates approaching 50%, whereas Llama 4 and the Gemma 3 models remain near zero. More broadly, with the exception of GPT-5.2, completion performance remains low even under strong cueing conditions. Taken together, these findings reveal a substantial gap between semantic understanding and access to exact poetic form. While models can often explain the meaning of a verse, producing its canonical continuation remains challenging.

Model	Name	+Prose	+Para	+Salient	+Shuffle
GPT-5.2	**44.5**	**51.6**	**44.5**	**33.8**	**23.1**
Sonnet 4.5	3.6	7.1	4.6	4.6	11.9
DeepSeek V3.2	7.1	7.3	5.3	3.2	11.7
DeepSeek R1	9.6	10.5	7.3	8.2	13.0
Gemma3 27B	2.0	2.0	2.0	2.0	1.4
Gemma3 12B	0.2	0.2	0.0	0.0	0.2
Llama4	0.7	1.6	1.4	0.7	1.6

Table 2: Successful second verse completion percentages. The second column is when only the name of Poet is given, and the other four correspond to our name + different contextual cues.

Recognition-Based Evaluation Table 3 reports model accuracy on the multiple-choice verse identification task across contextual conditions. For nearly all models, providing a prose explanation substantially improves performance. Paraphrased

prose yields similarly strong gains, although performance is generally lower than with the original prose (except for Llama 4). The strong performance with both original and paraphrased prose may partly reflect keyword overlap between the prose and the target verse. This pattern suggests that models benefit from both lexical overlap and semantic information when identifying canonical verses.

When models fail to select the correct continuation, they disproportionately choose distractors drawn from the same ghazal rather than from different ghazals (Figure 4). Because all options are controlled for poetic meter, this tendency indicates sensitivity to shared thematic or stylistic properties within a ghazal beyond exact surface form.

Compared to completion-based settings, performance in the recognition setting is markedly higher across all contextual conditions. This contrast is particularly pronounced for Llama 4 and the Gemma 3 models, which achieve competitive recognition performance despite near-zero completion performance. The resulting gap between completion and recognition is one of the central findings of this work: models often fail to produce canonical verses directly, yet can reliably identify them when presented among alternatives. This suggests that access to canonical poetic form is substantially more robust under recognition than under open-ended completion.

Table 3 also reports results on the binary-choice shuffled-form task, where the distractor consists of a word-shuffled version of the target verse. All models perform well above chance level, indicating sensitivity to canonical word order beyond lexical identity alone. GPT-5.2 achieves the highest accuracy by a substantial margin, while Claude Sonnet 4.5 and DeepSeek R1 also demonstrate strong performance. These results complement the shuffled-form completion experiments and suggest that models encode meaningful information about canonical verse structure even when they struggle to reproduce the verse itself.

Finally, similar to the completion setting, DeepSeek R1 consistently outperforms DeepSeek V3.2 in recognition settings, while performing slightly worse on poem-to-prose understanding (Table 1). The advantage is particularly pronounced in the shuffled□form condition, where reordering jumbled words to identify the canonical verse likely benefits from the explicit, step□by□step reasoning that DeepSeek R1 employs. This pattern suggests that reasoning-oriented models may be better able to exploit structural cues present in canonical poetry, although further investigation is needed to understand the mechanisms underlying this advantage.

Model	Name	+Prose	+Para	+Shuffle
GPT-5.2	**93.8**	**99.8**	**99.5**	**99.5**
Sonnet 4.5	66.0	98.9	97.7	95.4
DeepSeek V3.2	47.5	92.5	92.5	79.0
DeepSeek R1	55.0	97.0	96.3	92.0
Gemma3 27B	59.1	97.3	93.4	86.8
Gemma3 12B	52.7	92.7	91.1	65.3
Llama4	54.6	86.5	90.2	77.4

Table 3: Accuracy (%) on the multiple-choice verse recognition task under varying contextual cues.

4.3 Cross-Lingual Analysis: Shakespeare's Verse Completion

To assess whether the limitations observed in Persian are specific to a comparatively low-resource language or reflect a broader limitation of current LLMs, we conduct parallel completion experiments on *Shakespeare's Sonnets* (Shakespeare, 2002). Shakespeare's sonnets provide a useful comparison point as a canonical and highly structured poetic form in a high-resource language, serving as a rough English analogue to Persian ghazals in terms of literary status and formal constraint (Gray et al., 1995). From the full set of 154 sonnets, we randomly sample 50 sonnets, yielding 350 line pairs for evaluation.

To align the evaluation across poetic traditions, we treat two consecutive sonnet lines as the functional equivalent of a ghazal couplet. Because reference prose explanations are unavailable for Shakespeare's sonnets, we evaluate three cueing conditions: (i) poet's name only, (ii) poet's name plus salient words from the target line, and (iii) poet's name plus a word-shuffled version of the target line. In all settings, the first line is provided and the model is asked to generate the second. Salient words are extracted using Gemini 3 Flash. To avoid trivial cues, sentence-initial capitalization is removed from both salient-word and shuffled-line prompts.

Before presenting the results, we compare basic corpus statistics for the Persian and English datasets. As shown in Appendix F, Persian verses are shorter in characters but exhibit similar word counts, vocabulary sizes, and lexical diversity. Both corpora also follow strong metrical and rhyming conventions.

While the two traditions are not directly equivalent, these similarities reduce the likelihood that large performance differences are driven solely by superficial corpus characteristics.

Table 4 reports the results. All models achieve substantially higher completion performance in English than in Persian, with particularly large gains for the closed-source models. This finding suggests that the difficulties observed for Persian ghazals are not solely a consequence of the task itself, but are likely influenced by differences in training exposure to canonical literary texts. In contrast, Llama 4 and the Gemma 3 models continue to exhibit low completion performance, indicating limited access to canonical poetic forms in both languages.

Across conditions, the strongest performance is consistently observed when models are provided with a word-shuffled version of the target line, mirroring the pattern found for Persian ghazals. Salient-word prompts, by contrast, provide little benefit and occasionally reduce performance. Together, these results reinforce the importance of lexical identity for completion-based access to canonical poetic form and provide further evidence that the limitations observed in Persian are at least partly distributional rather than architectural.

Model	Name	+Salient	+Shuffle
GPT-5.2	92.6	**97.7**	98.0
Sonnet 4.5	**96.3**	94.0	**98.9**
DeepSeek V3.2	80.3	60.0	81.1
DeepSeek R1	91.7	94.3	97.1
Gemma3 27B	6.0	5.4	20.6
Gemma3 12B	0.6	1.2	6.0
Llama4	19.7	20.3	24.0

Table 4: Successful second line completion percentages in Shakespeare's sonnets.

5 Discussion and Conclusion

We introduced GhazalBench, a benchmark for evaluating how large language models engage with *culturally canonical poetry* under *usage-grounded diagnostic scenarios*. Rather than treating poetry solely as a generation task, our benchmark evaluates whether models can participate in interaction patterns common among Persian speakers, including paraphrasing, recognizing, completing, and reconstructing verses from Hafez's ghazals.

Across models and evaluation settings, we observe a clear dissociation between *semantic understanding* and *access to canonical surface form*. While LLMs perform well on poem-to-prose paraphrasing and recognition-based verse identification, completion performance remains low across a wide range of cueing conditions. In other words, understanding the meaning of a ghazal does not necessarily imply the ability to produce its canonical form.

This completion–recognition gap has important implications for both evaluation and deployment. Models that can accurately explain or recognize a verse may still struggle to participate in quotation-based or allusive interactions, which are central to the cultural use of poetry. As a result, evaluation paradigms that rely primarily on paraphrasing or recognition tasks may substantially overestimate a model's ability to engage with culturally grounded language practices.

Our cross-lingual experiments on Shakespeare's sonnets further suggest that these limitations are at least partly distributional rather than architectural. Several models exhibit substantially stronger completion performance in English than in Persian, indicating that access to canonical surface forms depends heavily on training exposure. However, the substantial variation across models on Persian ghazals shows that this capability is not fixed. While some models exhibit near-zero completion rates across conditions, GPT-5.2 achieves markedly stronger performance, demonstrating that access to culturally canonical poetic forms can be learned to a significant extent. More broadly, the results point to an important challenge for multilingual LLMs: semantic competence may generalize across languages more readily than access to culturally canonical forms.

Future Work. Future work could extend this framework to other literary traditions, genres, and languages, enabling broader cross-cultural analysis of canonical surface-form access. Additional diagnostic tasks, such as evaluating the understanding of named entities, historical references, and literary allusions, could help disentangle cultural knowledge from access to exact form. At the modeling level, our findings motivate approaches that improve access to culturally significant canonical texts while preserving the safeguards required for sensitive or copyrighted content. More broadly, incorporating usage-grounded diagnostics into evaluation suites may reveal important limitations that remain hidden under standard generation and recognition benchmarks.

Limitations

Our study focuses on a single Persian poetic form (the ghazal) and a single poet (Hafez). Although this choice is motivated by the central cultural role of Hafez's ghazals in Persian literary practice, it constrains the generality of our findings. In addition, our cross-lingual analysis is restricted to one comparison language (English) and one poet (Shakespeare), which limits the scope of conclusions that can be drawn about cross-cultural poetic knowledge in LLMs. Moreover, we only assess second-verse access in our settings, while first-verse access, which is likely more challenging for causal language models, is not explored. Finally, we adopt a single zero-shot prompting strategy, which may not capture the full range of model behaviors when engaging with culturally canonical poetic texts.

References

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Meta AI blog (LLaMA 4 announcement).

Wafa Al Ghallabi, Ritesh Thawkar, Sara Ghaboura, Ketan Pravin More, Omkar Thawakar, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. 2025. Fann or flop: A multigenre, multiera benchmark for arabic poetry understanding in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20235–20255.

Anthropic. 2025. Claude sonnet 4.5 system card. Technical system card / model documentation. Accessed: 2025-12-13.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Mena Attia, Aashiq Muhamed, Mai Alkhamissi, Thamar Solorio, and Mona T. Diab. 2026. Beyond understanding: Evaluating the pragmatic gap in LLMs' cultural processing of figurative language. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7238–7265, Rabat, Morocco. Association for Computational Linguistics.

Junyi Bian, Baojun Lin, Ke Zhang, Zhaohui Yan, Hong Tang, and Yonghe Zhang. 2019. Controllable length control neural encoder-decoder via reinforcement learning. *arXiv preprint arXiv:1909.09492*.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Tong Chen, Faeze Brahman, Jiacheng Liu, Niloofar Mireshghallah, Weijia Shi, Pang Wei Koh, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2025. Parapo: Aligning language models to reduce verbatim reproduction of pre-training data. *arXiv preprint arXiv:2504.14452*.

Yanran Chen, Hannes Gröner, Sina Zarrieß, and Steffen Eger. 2024. Evaluating diversity in automatic poetry generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19671–19692.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Johannes Thomas Pieter de Bruijn. 2000. □azal i. history. *Encyclopædia Iranica*. Accessed: 2026-01-04.

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

Lyra D'Souza and David Mimno. 2023. The chatbot and the canon: Poetry memorization in llms. *Proceedings http://ceur-ws. org ISSN*, 1613:0073.

Richard C Foltz. 2016. *Iran in world history*. Oxford University Press.

Gemma Team and Google DeepMind. 2025. Gemma 3. *Technical Report*. Available at https://storage.googleapis.com/deepmind-media/gemma/Gemma3Report.pdf.

Omid Ghahroodi, Marzia Nouri, Mohammad Vali Sanian, Alireza Sahebi, Doratossadat Dastgheib, Ehsaneddin Asgari, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. 2024. Khayyam challenge (persianmmlu): Is your llm truly wise to the persian language? *arXiv preprint arXiv:2404.06644*.

Google DeepMind. 2025. Gemini 3 pro model card. Google DeepMind Tech. Note. Published Nov. 2025. Available at https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

Google LLC. 2025. Introducing gemini 3 flash: Frontier intelligence built for speed. Google Official Blog / Product Announcement. Available at https://blog.google/products/gemini/gemini-3-flash/.

Elizabeth T Gray and 1 others. 1995. *The Green Sea of Heaven: Fifty Ghazals from the Díwán of □áfì□*. White Cloud Press.

Leire Hernández Martínez. 2025. Metric poetry generation using large language models.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam De Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, and 1 others. 2022. Challenges and strategies in cross-cultural nlp. *arXiv preprint arXiv:2203.10020*.

Ryan Heuser. 2025. Generative aesthetics: On formal stuckness in ai verse.

Faezeh Hosseini, Mohammadali Yousefzadeh, and Yadollah Yaghoobzadeh. 2026. Ffe-hallu: Hallucinations in fixed figurative expressions: Benchmark of idioms and proverbs in the persian language. *arXiv preprint arXiv:2601.20105*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 1(2):4.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Ghazal Kalhor and Behnam Bahrak. 2026. Masalbench: A benchmark for contextual and cross-cultural understanding of persian proverbs in llms. *arXiv preprint arXiv:2601.22050*.

Gopichand Kanumolu, Lokesh Madasu, Pavan Baswani, Ananya Mukherjee, and Manish Shrivastava. 2023. Unsupervised approach to evaluate sentence-level fluency: Do we really need reference? *arXiv preprint arXiv:2312.01500*.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Eunhye Kim, Kiroong Choe, Minju Yoo, Sadat Shams Chowdhury, and Jinwook Seo. 2025. Beyond tools: Understanding how heavy users integrate llms into everyday tasks and decision-making. *arXiv preprint arXiv:2502.15395*.

Klaus Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology*, 4th edition. SAGE Publications, Inc., Thousand Oaks, CA.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *arXiv preprint arXiv:2203.05227*.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Bolei Ma, Yina Yao, and Anna-Carolina Haensch. 2025. Capabilities and evaluation biases of large language models in classical chinese poetry generation: A case study on tang poetry. *arXiv preprint arXiv:2510.15313*.

Enrico Mensa, Lorenzo Zane, Calogero Jerik Scozzaro, Matteo Delsanto, Tommaso Milani, and Daniele P. Radicioni. 2025. Easy to complete, hard to choose: Investigating LLM performance on the ProverbIT benchmark. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, volume 4112 of *CEUR Workshop Proceedings*, pages 722–734, Cagliari, Italy. CEUR-WS.org.

Michail Mersinias and Kyle Mahowald. 2023. For generated text, is nli-neutral text the best text? *arXiv preprint arXiv:2302.08577*.

Melika Nobakhtian, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2025. Evaluating cultural knowledge and reasoning in llms through persian allusions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25725–25737.

OpenAI. 2025. Introducing GPT-5.2. Blog post. Accessed: 2026-05-14.

Reza Ordoubadian and 1 others. 2006. *The Poems of Hafez*, volume 6. Ibex Publishers, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

William Shakespeare. 2002. *The complete sonnets and poems*. Oxford University Press.

Hafez Shirazi. 2009. *Divan-e Hafez with Complete Explanation of the Couplets*, 8th edition. Ganjineh.

Charles Spearman. 2010. The proof and measurement of association between two things. *International Journal of Epidemiology*, 39(5):1137–1150.

Armin Tourajmehr, Mohammad Reza Modarres, and Yadollah Yaghoobzadeh. 2025. Evaluating the creativity of llms in persian literary text generation. *arXiv preprint arXiv:2509.18401*.

Endel Tulving and Donald M Thomson. 1973. Encoding specificity and retrieval processes in episodic memory. *Psychological review*, 80(5):352.

Endel Tulving and 1 others. 1972. Episodic and semantic memory. *Organization of memory*, 1(381-403):1.

Melanie Walsh, Maria Antoniak, and Anna Preus. 2024a. Sonnet or not, bot? poetry evaluation for large models and datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15568–15603.

Melanie Walsh, Anna Preus, and Elizabeth Gronski. 2024b. Does chatgpt have a poetic style? *arXiv preprint arXiv:2410.15299*.

Scarlet Weatherby, Noel Ashbourne, and Jacob Palmerston. 2024. Exploring poetic creativity in large language models: A dynamic multi-agent framework for poem generation. *Authorea Preprints*.

Xinhao Zhang, Olga Seminck, and Pascal Amsili. 2024. Remember to forget: A study on verbatim memorization of literature in large language models. In *Proceedings of the Computational Humanities Research Conference 2024*, volume 3834 of *CEUR Workshop Proceedings*, pages 961–981, Aarhus, Denmark. CEUR-WS.org.

A Ghazal Introduction

The ghazal is a poetic form typically centered on romantic themes, consisting of five to twelve couplets. The first verse of the opening couplet rhymes with all subsequent even□numbered verses. Etymologically, ghazal means "talking with the beloved, declaring one's love, recounting youth, and the conversation of companionship and love." Table 5 provides some examples of Hafez's ghazal couplets along with their English translations.

B Benchmark Statistics

Dataset Overview GhazalBench is constructed at the couplet level and consists of 438 Persian couplets and 350 English couplets. Each couplet is evaluated under multiple task settings, resulting in a total of 5,430 prompts per model across all tasks. This evaluation design enables systematic analysis of model behavior across different understanding and access conditions.

Task Distribution Table 6 summarizes the distribution of tasks across languages and evaluation settings.

Couplet Length We report the length of couplets measured in characters.

- **Persian:** mean = 65.94, std = 6.06
- **English:** mean = 85.72, std = 6.18

Prompt Length We further report the length of input prompts (in characters).

- **Persian (Paraphrasing):** mean = 1110.94, std = 6.06
- **Persian (Recognition):** mean = 723.87, std = 112.06
- **Persian (Completion):** mean = 935.90, std = 290.10
- **English (Completion):** mean = 1250.15, std = 416.76

Discussion While the dataset is moderate in size, the benchmark is designed to systematically vary access conditions at the couplet level. This results in a large number of evaluation instances per model and enables controlled comparisons across multiple task settings, particularly for analyzing completion, recognition, and semantic mediation in poetic access.

Couplet Text	English Translation
هرگز نمیرد آن که دلش زنده شد به عشق / ثبت است بر جریده عالم دوام ما	Never dies the one whose heart is quickened by love; Our everlasting life is inscribed in the world's record.
تا مرا عشق تو تعلیم سخن گفتن کرد / خلق را ورد زبان مدحت و تحسین من است	Ever since your love taught me the art of speech, My praise and admiration have become the constant refrain on people's tongues.
ندای عشق تو دیشب در اندرون دادند / فضایِ سینهٔ حافظ هنوز پر ز صداست	The call of your love was raised within my soul last night; The space of Hafez's chest is still filled with its echo.
راهیست راه عشق که هیچش کناره نیست / آنجا جز آن که جان بسپارند چاره نیست	The path of love is a road with no end in sight; There, there is no choice but to surrender one's soul.
لطیفه‌ایست نهانی که عشق از او خیزد / که نام آن نه لب لعل و خط زنگاریست	There is a hidden mystery from which love arises, A mystery not called ruby lips or the faint mustache of youth.

Table 5: Examples of Hafez's ghazal couplets along with their English translations.

Language	Task Type	# Instances	Proportion
Persian	Paraphrasing	438	10%
Persian	Recognition	1,752	40%
Persian	Completion	2,190	50%
English	Completion	1,050	100%

Table 6: Task distribution across languages and evaluation settings.

C LLM-as-Judge Evaluation

C.1 Evaluation Dimensions

We evaluate the generated prose using Gemini 3 Pro as an LLM-as-Judge (Google DeepMind, 2025). For each instance, the judge receives the full ghazal, the reference prose explanation, and the model-generated paraphrase. It evaluates the paraphrase along six dimensions: **Semantic Fidelity**, **Naturalness & Fluency**, **Clarity & Simplicity**, **Interpretation Avoidance**, **Neutral Tone Preservation**, and **Length & Brevity Control**. These dimensions reflect commonly used criteria in natural language generation evaluation (Li et al., 2022; Kanumolu et al., 2023; Reiter and Belz, 2009; Mersinias and Mahowald, 2023; Bian et al., 2019). Each dimension is scored on a five-point scale, and the judge additionally provides an overall score.

C.2 Results

Table 7 reports the LLM-as-Judge evaluation of model-generated prose, broken down by evaluation dimension. All models perform well on the poem-to-prose task, indicating generally strong semantic understanding of Hafez's ghazals. Across all dimensions, the closed-source models consistently outperform the open-weight models, suggesting higher proficiency in Persian literary prose.

Among the evaluation dimensions, *Neutral Tone Preservation* receives the highest scores for all models. The lowest-scoring dimension for the closed-source models is *Interpretation Avoidance*, while for the open-weight models it is *Semantic Fidelity*. This pattern aligns with proprietary models' stronger ability to produce faithful and stylistically appropriate paraphrases. These trends are consistent with the corpus-level automatic evaluation, particularly along the dimensions of *Semantic Fidelity* and *Clarity*. At the same time, automatic metrics remain limited in capturing discourse-level qualities such as interpretation avoidance and tone preservation, underscoring the value of complementary, human-aligned evaluation for literary paraphrasing tasks.

C.3 Human Validation

To evaluate the reliability of the LLM-as-Judge scores, we randomly sample 50 couplets and their corresponding prose generated by DeepSeek V3.2. Two authors independently rate each prose along the six dimensions. We compute weighted κ, Krippendorff's α, and percent agreement to measure inter-annotator reliability (Cohen, 1968; Krippendorff, 2019; Artstein and Poesio, 2008). The two annotators' scores are then averaged to form gold labels. Finally, we calculate Spearman's ρ, Kendall's τ, and mean absolute error (MAE) between the LLM scores and the gold labels to assess human–LLM alignment (Spearman, 2010; Kendall, 1938).

Table 8 reports these metrics. As shown, inter-annotator agreement is high across most dimensions, with Naturalness & Fluency achieving the

Dimension	GPT-5.2	Sonnet 4.5	DeepSeek V3.2	DeepSeek R1	Gemma3 27B	Gemma3 12B	Llama4
Semantic Fidelity	4.4	**4.4**	3.8	3.3	3.3	2.9	2.7
Naturalness & Fluency	**4.8**	4.7	4.5	4.3	4.5	4.1	3.7
Clarity & Simplicity	**4.8**	4.8	4.5	4.4	4.4	4.0	3.7
Interpretation Avoidance	4.2	**4.4**	4.0	3.4	3.5	3.5	3.3
Neutral Tone Preservation	**5.0**	5.0	4.9	4.6	4.9	4.9	4.8
Length & Brevity Control	4.6	**4.7**	4.6	4.3	4.6	4.6	4.3
Overall Score	**4.5**	4.4	4.0	3.5	3.6	3.1	2.9

Table 7: LLM-as-Judge dimension-wise scores for LLM-generated prose of Hafez's ghazals.

highest reliability ($\kappa = \mathbf{0.873}$, $\alpha = \mathbf{0.874}$, 90% exact agreement). The LLM scores exhibit strong rank correlation with the averaged human judgments on Semantic Fidelity ($\rho = \mathbf{0.743}$, $\tau = \mathbf{0.635}$), confirming that the LLM captures a fine-grained quality signal for this dimension. In contrast, Neutral Tone Preservation shows a high percentage of exact agreement (80%) yet the weakest human–LLM alignment ($\rho = 0.186$) and the lowest MAE (0.190), suggesting a restricted rating range where most scores cluster at the top of the scale; the few detectable human distinctions are not reflected in the LLM's scores.

D Example Prompts

D.1 Poem Collection Prompt

An example prompt used to collect ghazals and their respective prose, couplet by couplet, is provided in Table 9.

D.2 Understanding Prompts

D.2.1 Prose Generation

An example of the prompt used to generate prose from Hafez's ghazal at the couplet level is provided in Table 10.

D.2.2 Prose Evaluation

An example of the prompt used to assess LLM-generated prose for a Hafez ghazal, with Gemini 3 Pro serving as the LLM-as-Judge, is shown in Figure 2 (original Persian) and its English translation in Figure 3.

D.3 Auxiliary Prompts

The following auxiliary prompts are used to construct canonical surface-form access questions and are not themselves used for model evaluation.

D.3.1 Prose Paraphrasing

D.3.2 Salient Words Selection

D.4 Completion / Recognition Prompts

D.4.1 Completion Questions

Example prompts for completing Hafez's ghazals at the verse level, given different input information, are provided in Tables 14–17.

D.4.2 Multiple-Choice Questions

Example prompts for verse-level recognition of Hafez's ghazals under different input conditions are provided in Tables 18–20.

D.4.3 English Prompts

Example prompts for line completion of Shakespeare's sonnets under different input conditions are provided in Tables 21–23.

E Failure Modes

Verse Completion Table 24 provides examples of failure categories observed in our LLM-generated data for completion experiments. Hallucination is when LLMs access a plausible but incorrect verse that does not exist. Prosaic Paraphrase is when models access an explanation of the verse's meaning in prose format. Incomplete Completion is when models access only a part of the correct verse. Contextual Confusion is when models access a verse from another ghazal of Hafez.

Verse Recognition Figure 4 presents the distribution of distractor options for the recognition-based task across different scenarios.

F Corpus Statistics: Persian vs. English

Table 25 compares the Persian ghazal couplets and English sonnet couplets used in our cross-lingual analysis. All Persian text was normalized using the pipeline described in Section 4.2; English text was lowercased and stripped of punctuation before tokenisation. Character lengths and word counts

Dimension	Weighted κ	Krippendorff α	% Agree	Spearman ρ	Kendall τ	MAE
Semantic Fidelity	0.808	0.806	0.580	**0.743**	**0.635**	0.710
Naturalness & Fluency	**0.873**	**0.874**	**0.900**	0.448	0.404	0.410
Clarity & Simplicity	0.654	0.656	0.680	0.628	0.564	0.450
Interpretation Avoidance	0.694	0.685	0.600	0.472	0.414	0.800
Neutral Tone Preservation	0.561	0.562	0.800	0.186	0.182	**0.190**
Length & Brevity Control	0.647	0.651	0.580	0.309	0.267	0.700

Table 8: Inter-annotator reliability and human–LLM alignment for each evaluation dimension. Human scores are averaged to compute alignment metrics.

شما یک مدل زبانی هستید و وظیفه دارید صرفاً متن کتاب «دیوان حافظ با شرح کامل ابیات» که در این فایل آپلود شده است را بخوانید و بدون هیچ‌گونه تغییر، حذف، بازنویسی یا افزودن، محتوای آن را استخراج نمایید.

لطفاً دستورالعمل‌های زیر را دقیقاً رعایت کنید:
1. هر غزل را به صورت بیت به بیت بررسی کنید.
2. برای هر بیت، شرح یا تفسیر مربوط به همان بیت را عیناً و بدون تغییر از متن کتاب استخراج و درج نمایید.
3. خروجی را در قالب یک جدول با چهار ستون زیر ارائه دهید:
- «شماره غزل»
- «شماره بیت»
- «متن بیت»
- «شرح یا تفسیر»
4. در ستون «متن بیت»، دو مصراع هر بیت را با علامت « / » از یکدیگر جدا کنید.
5. در ستون «شرح یا تفسیر» فقط و فقط متن کتاب را عیناً بیاورید و هیچ توضیح یا تحلیل شخصی خود را اضافه نکنید.
6. اگر برای بیتی شرحی در کتاب وجود ندارد، ستون مربوط به شرح را خالی بگذارید.
7. ساختار جدول را به گونه‌ای ارائه دهید که بتوان آن را به راحتی در فایل lecxE یا VSC ذخیره و پردازش نمود.

توجه: رعایت امانت‌داری کامل در نقل متن کتاب ضروری است.

English Translation:
You are a language model. Your task is to read only the text of the book *"Divan of Hafez with Complete Commentary on the Couplets"* provided in this uploaded file, and extract its content without any modification, omission, paraphrasing, or addition.

Please strictly follow these instructions:
1. Process each ghazal couplet by couplet.
2. For each couplet, extract and reproduce the corresponding prose exactly as it appears in the book, without any changes.
3. Present the output as a table with the following four columns:
- "Ghazal Number"
- "Couplet Number"
- "Couplet Text"
- "Prose"
4. In the "Couplet Text" column, separate the two verses of each couplet using the symbol " / ".
5. In the "Prose" column, include only the exact text from the book and do not add any personal explanation or interpretation.
6. If no prose is provided for a couplet in the book, leave the corresponding cell empty.
7. Structure the table so that it can be easily exported to and processed in Excel or CSV format.

Note: Strict fidelity to the original text is essential.

Table 9: Example prompt for collecting ghazals and their prose from the Divan of Hafez.

are computed per verse/line, then averaged across the corpus.

شما یک مدل زبانی هستید که برای بازنویسی ابیات غزل‌های حافظ به نثر طبیعی طراحی شده‌اید.

به شما یک بیت از یکی از غزل‌های حافظ داده می‌شود. وظیفه‌ی شما این است که این بیت را به نثر طبیعی و استاندارد فارسی بازنویسی کنید.

لطفاً دستورالعمل‌های زیر را دقیقاً رعایت کنید:
1- از توضیح‌های طولانی، تحلیل‌های عمیق یا ساختارهای پیچیده خودداری کنید.
2- نثر بیت را در قالب یک پاراگراف کوتاه فارسی ارائه دهید.
3- از تکرار خود بیت در نثر آن خودداری کنید.
4- خروجی را فقط در قالب زیر بازگردانید و از ارائه هرگونه توضیح اضافی خودداری کنید:
<answer>نثر بیت</answer>

نمونه 1:
ورودی:
<couplet>الا یا ایها الساقی ادر کاساً و ناولها / که عشق آسان نمود اول ولی افتاد مشکل‌ها</couplet>
خروجی:
<answer>هان ای ساقی، جام باده را به چرخش درآور و آن را به دست من بده زیرا که گرچه عشق در ابتدا، ساده به نظر می‌رسید اما سخت مشکل افتاده است.</answer>

نمونه 2:
ورودی:
<couplet>اگر به زلف دراز تو دست ما نرسد / گناه بخت پریشان و دست کوته ماست</couplet>
خروجی:
<answer>اگر دست ما به سر زلف بلند تو نرسد گناه از بخت نامساعد و دست کوتاه ماست.</answer>

بیت:
<couplet>گویی از صحبت ما نیک به تنگ آمده بود / بار بربست و به گردش نرسیدیم و برفت</couplet>

English Translation:
You are a language model designed to rewrite couplets from the ghazals of Hafez into natural prose.

You will be given a single couplet from a Hafez ghazal. Your task is to rewrite this couplet into fluent, standard Persian prose.

Please strictly follow these instructions:
1- Avoid lengthy explanations, deep analysis, or complex structures.
2- Present the prose in the form of a short paragraph in Persian.
3- Do not repeat the original couplet in your prose.
4- Return the output strictly in the following format and do not include any additional text:
<answer>Prose</answer>

Example 1:
Input:
<couplet>O cupbearer, pass the cup and offer it around / Love seemed easy at first, but difficulties arose</couplet>
Output:
<answer>O cupbearer, pass the wine and give it to me, for although love first appeared easy, it has now become difficult.</answer>

Example 2:
Input:
<couplet>If our hand does not reach your long tress / It is the fault of our ill fortune and short reach</couplet>
Output:
<answer>If we cannot reach your long hair, it is due to our misfortune and our limited reach.</answer>

Couplet:
<couplet>It seemed she had grown weary of our company / She packed her belongings and left before we could catch up</couplet>

Table 10: Example prompt for generating prose from a Hafez ghazal at the couplet level.

شما یک مدل زبانی هستید که برای بازنویسی نثر فارسی طراحی شده‌اید.

به شما یک متن فارسی داده می‌شود. وظیفه‌ی شما این است که آن متن را با حفظ کامل معنا بازنویسی کنید، به‌گونه‌ای که تا حد امکان واژگان، عبارات و ساختار جمله‌ها تغییر کنند، بدون آن‌که محتوا یا پیام متن دچار تغییر شود.

لطفاً دستورالعمل‌های زیر را دقیقاً رعایت کنید:
1- نثر بازنویسی‌شده باید از نظر معنایی کاملاً معادل نثر ورودی باشد.
2- نثر بازنویسی‌شده باید کاملاً طبیعی، روان و مطابق با فارسی معیار باشد.
3- هیچ‌گونه اطلاعات، توضیح یا محتوای جدیدی نباید به متن اضافه شود و هیچ بخشی از معنا نیز نباید حذف شود.
4- جایگزینی واژگان و عبارات باید به‌گونه‌ای انجام شود که شباهت واژگانی با متن ورودی حداقل شود، مشروط بر آن‌که معنا کاملاً حفظ شود.
5- ساختار جمله‌ها باید تا حد امکان تغییر کند، بدون تغییر در معنا.
6- نثر بازنویسی‌شده باید غیرشاعرانه و صرفاً توضیحی باقی بماند.
7- خروجی را فقط در قالب زیر بازگردانید و از ارائه هرگونه توضیح اضافی خودداری کنید:
<answer>نثر بازنویسی‌شده</answer>

نثر ورودی:
<prose>ای دل با این شتاب کجا می‌روی، فقط به سیب زنخدان یار نگاه مکن و فریفته‌اش مشو بلکه متوجه چاه ذقن او نیز باش که در راهت است.</prose>

English Translation:
You are a language model designed to paraphrase Persian prose.

You will be given a Persian text. Your task is to rewrite the text while fully preserving its meaning, such that the vocabulary, expressions, and sentence structure are changed as much as possible without altering the content or message.

Please strictly follow these instructions:
1- The rewritten prose must be semantically equivalent to the input.
2- The rewritten prose must be natural, fluent, and conform to standard Persian.
3- Do not add any new information, explanation, or content, and do not omit any part of the meaning.
4- Replace words and expressions so as to minimize lexical overlap with the input, while fully preserving the meaning.
5- Modify sentence structure as much as possible without changing the meaning.
6- The rewritten prose must remain non-poetic and purely explanatory.
7- Return the output strictly in the following format and do not include any additional text:
<answer>Paraphrased prose</answer>

Input prose:
<prose>O heart, where are you going in such haste? Do not fix your gaze only on the beloved's dimpled chin and be enchanted by it; also be mindful of the pit of the chin that lies in your path.</prose>

Table 11: Example prompt for paraphrasing prose from a Hafez ghazal at the couplet level.

شما یک مدل زبانی هستید که برای شناسایی واژه‌های برجسته از ابیات غزل‌های حافظ طراحی شده‌اید.

به شما مصراع دوم از یکی از ابیات غزل‌های حافظ داده می‌شود. وظیفه‌ی شما این است که دقیقاً دو واژه از همان مصراع را انتخاب کنید که از نظر معنایی و حافظه‌ای برجسته‌ترین باشند؛ به‌گونه‌ای که حتی در صورت فراموش شدن کل مصراع، بیشترین احتمال به خاطر سپرده شدن را داشته باشند.

لطفاً دستورالعمل‌های زیر را دقیقاً رعایت کنید:
1- واژه‌ها باید دقیقاً به همان صورت سطحی در مصراع ورودی وجود داشته باشند.
2- هیچ واژه‌ای را تغییر ندهید، خلاصه نکنید، یا بازنویسی نکنید.
3- فقط واژه‌های منفرد (نه عبارت یا ترکیب چندواژه‌ای) انتخاب کنید.
4- دو واژه‌ی انتخاب‌شده باید متمایز و غیرتکراری باشند، حتی اگر یک واژه بیش از یک‌بار در مصراع ظاهر شده باشد.
5- خروجی را فقط در قالب زیر بازگردانید و از ارائه هرگونه توضیح اضافی خودداری کنید:
["واژه اول", "واژه دوم"]

مصراع ورودی:
<second_verse>که عشق از پرده عصمت برون آرد زلیخا را</second_verse>

English Translation:
You are a language model designed to identify salient words from couplets of Hafez's ghazals.

You will be given the second verse of a couplet from a Hafez ghazal. Your task is to select exactly two words from that verse that are the most salient in terms of meaning and memorability, such that they are most likely to be remembered even if the entire verse is forgotten.

Please strictly follow these instructions:
1- The selected words must appear exactly in their surface form in the input verse.
2- Do not modify, shorten, or paraphrase any word.
3- Select only single words (not phrases or multi-word expressions).
4- The two selected words must be distinct and non-repeating, even if a word appears multiple times in the verse.
5- Return the output strictly in the following format and do not include any additional text:
["word1", "word2"]

Input verse:
<second_verse>Love would draw Zuleikha out from behind the veil of chastity</second_verse>

Table 12: Example prompt for choosing salient words from a Hafez ghazal at the verse level.

You are a language model designed to identify salient words from lines of Shakespeare's Sonnets.

You are given the second line of a couplet from one of Shakespeare's sonnets. Your task is to select exactly two words from that line that are the most semantically and memorably salient, such that even if the entire line were forgotten, these two words would be the most likely to be remembered.

Please strictly follow the instructions below:
1- The selected words must appear exactly in the input line in their original surface form.
2- Do not modify, paraphrase, lemmatize, or rewrite any word.
3- Select only single words (not phrases or multi-word expressions).
4- The two selected words must be distinct and non-redundant, even if a word appears more than once in the line.
5- Return the output strictly in the following format and do not provide any additional explanation:
["first_word", "second_word"]

Input line:
<second_line>When beauty lived and died as flowers do now,</second_line>

Table 13: Example prompt for choosing salient words from a Shakespeare sonnet at the line level.

شما یک مدل زبانی هستید که برای بازیابی مصراع دوم دقیق و اصیل حافظ از دیوان او طراحی شده‌اید.

به شما مصراع اول از یکی از ابیات غزل‌های حافظ داده می‌شود. وظیفه‌ی شما این است که مصراع دوم واقعی و موجود در دیوان حافظ را که مربوط به همین مصراع اول است، شناسایی کرده و بدون هیچ‌گونه تغییری بازگردانید.

لطفاً دستورالعمل‌های زیر را دقیقاً رعایت کنید:
1- هیچ مصراعی نسازید؛ فقط از متن معتبر دیوان حافظ بازیابی کنید.
2- مصراع باید کاملاً مطابق متن اصلی حافظ باشد؛ هیچ واژه یا نشانه‌ای را تغییر ندهید.
<answer>مصراع دوم واقعی حافظ</answer>

مصراع اول:
<first_verse>خوبان پارسی گو بخشندگان عمرند</first_verse>

English Translation:
You are a language model designed to retrieve the exact and authentic second verse of Hafez from his Divan.

You will be given the first verse of a couplet from a Hafez ghazal. Your task is to identify the true second verse from the Divan of Hafez that corresponds to this first verse, and return it without any modification.

Please strictly follow these instructions:
1- Do not generate any verse; only retrieve it from the authentic text of the Divan of Hafez.
2- The verse must exactly match the original text of Hafez; do not alter any word or punctuation.
<answer>Exact second verse</answer>

Input first verse:
<first_verse>The Persian-speaking beloveds are givers of life</first_verse>

Table 14: Example prompt for completing the second verse of a ghazal's couplet, given the poet's name (Hafez).

شما یک مدل زبانی هستید که برای بازیابی مصراع دوم دقیق و اصیل حافظ از دیوان او طراحی شده‌اید.

به شما مصراع اول یکی از ابیات غزل‌های حافظ داده می‌شود و سپس نثر روان فارسی همان بیت ارائه می‌گردد. وظیفه‌ی شما این است که مصراع دوم واقعی و موجود در دیوان حافظ را که مربوط به همین بیت است، شناسایی کرده و بدون هیچ‌گونه تغییری بازگردانید.

لطفاً دستورالعمل‌های زیر را دقیقاً رعایت کنید:
1- هیچ مصراعی نسازید؛ فقط از متن معتبر دیوان حافظ بازیابی کنید.
2- مصراع باید کاملاً مطابق متن اصلی حافظ باشد؛ هیچ واژه یا نشانه‌ای را تغییر ندهید.
<answer>مصراع دوم واقعی حافظ</answer>

مصراع اول:
<first_verse>غرور حسنت اجازت مگر نداد ای گل</first_verse>

نثر بیت:
<couplet_prose>ای گل، حتماً غرور زیبایی‌ات اجازه نداد که از حال بلبل عاشق خود احوالی بپرسی.</couplet_prose>

English Translation:
You are a language model designed to retrieve the exact and authentic second verse of Hafez from his Divan.

You will be given the first verse of a couplet from a Hafez ghazal, followed by a fluent Persian prose rendering of the same couplet. Your task is to identify the true second verse from the Divan of Hafez that corresponds to this couplet, and return it without any modification.

Please strictly follow these instructions:
1- Do not generate any verse; only retrieve it from the authentic text of the Divan of Hafez.
2- The verse must exactly match the original text of Hafez; do not alter any word or punctuation.
<answer>Exact second verse</answer>

Input first verse:
<first_verse>O flower, perhaps the pride of your beauty did not permit it</first_verse>

Couplet prose:
<couplet_prose>O flower, surely the pride in your beauty did not allow you to ask about the state of your lover nightingale.</couplet_prose>

Table 15: Example prompt for completing the second verse of a ghazal's couplet, given the poet's name (Hafez) and couplet prose.

شما یک مدل زبانی هستید که برای بازیابی مصراع دوم دقیق و اصیل حافظ از دیوان او طراحی شده‌اید.

به شما مصراع اول یکی از ابیات غزل‌های حافظ داده می‌شود و سپس دو واژه از مصراع دوم همان بیت ارائه می‌گردد. وظیفه‌ی شما این است که مصراع دوم واقعی و موجود در دیوان حافظ را که مربوط به همین بیت است، شناسایی کرده و بدون هیچ‌گونه تغییری بازگردانید.

لطفاً دستورالعمل‌های زیر را دقیقاً رعایت کنید:
1- هیچ مصراعی نسازید؛ فقط از متن معتبر دیوان حافظ بازیابی کنید.
2- مصراع باید کاملاً مطابق متن اصلی حافظ باشد؛ هیچ واژه یا نشانه‌ای را تغییر ندهید.
3- خروجی را فقط در قالب زیر بازگردانید و از ارائه هرگونه توضیح اضافی خودداری کنید:
<answer>مصراع دوم واقعی حافظ</answer>

مصراع اول:
<first_verse>غزل گفتی و در سفتی بیا و خوش بخوان حافظ</first_verse>

واژه‌ها:
["نظم", "ثریا"]

English Translation:
You are a language model designed to retrieve the exact and authentic second verse of Hafez from his Divan.

You will be given the first verse of a couplet from a Hafez ghazal, followed by two words from the second verse of the same couplet. Your task is to identify the true second verse from the Divan of Hafez that corresponds to this couplet, and return it without any modification.

Please strictly follow these instructions:
1- Do not generate any verse; only retrieve it from the authentic text of the Divan of Hafez.
2- The verse must exactly match the original text of Hafez; do not alter any word or punctuation.
3- Return the output strictly in the following format and do not include any additional text:
<answer>Exact second verse</answer>

Input first verse:
<first_verse>You composed a ghazal and strung its pearls—come, recite it beautifully, O Hafez</first_verse>

Given words:
["order", "Pleiades"]

Table 16: Example prompt for completing the second verse of a ghazal's couplet, given the poet's name (Hafez) and salient words of the second verse.

شما یک مدل زبانی هستید که برای بازیابی مصراع دوم دقیق و اصیل حافظ از دیوان او طراحی شده‌اید.

به شما مصراع اول یکی از ابیات غزل‌های حافظ داده می‌شود و سپس نسخه‌ای از مصراع دوم همان بیت که ترتیب واژگان آن جابه‌جا شده است ارائه می‌گردد. وظیفه‌ی شما این است که مصراع دوم واقعی و موجود در دیوان حافظ را که مربوط به همین بیت است، شناسایی کرده و بدون هیچ‌گونه تغییری بازگردانید.

لطفاً دستورالعمل‌های زیر را دقیقاً رعایت کنید:
1- هیچ مصراعی نسازید؛ فقط از متن معتبر دیوان حافظ بازیابی کنید.
2- مصراع باید کاملاً مطابق متن اصلی حافظ باشد؛ هیچ واژه یا نشانه‌ای را تغییر ندهید.
3- تنها با جابه‌جایی ترتیب واژگان موجود در نسخه‌ی ارائه‌شده از مصراع دوم، آن را به مصراع صحیح برسانید؛ افزودن یا حذف هرگونه واژه یا نشانه به‌هیچ‌وجه مجاز نیست.
4- خروجی را فقط در قالب زیر بازگردانید و از ارائه هرگونه توضیح اضافی خودداری کنید:
<answer>مصراع دوم واقعی حافظ</answer>

نمونه 1:
ورودی:
<first_verse>و گر رسم فنا خواهی که از عالم براندازی</first_verse>
<shuffled_second_verse>فروریزد هر جان برافشان مویت تا ز هزاران</shuffled_second_verse>
خروجی:
<answer>برافشان تا فروریزد هزاران جان ز هر مویت</answer>

نمونه 2:
ورودی:
<first_verse>دو چشم شوخ تو برهم زده خطا و حبش</first_verse>
<shuffled_second_verse>زلف چین هند به و ماچین داده خراج تو</shuffled_second_verse>
خروجی:
<answer>به چین زلف تو ماچین و هند داده خراج</answer>

مصراع اول:
<first_verse>هنگام وداع تو ز بس گریه که کردم</first_verse>

مصراع دوم با ترتیب واژگانی جابه‌جاشده:
<shuffled_second_verse>رخ نور چشم دور نماندست از مرا تو</shuffled_second_verse>

English Translation:
You are a language model designed to retrieve the exact and authentic second verse of Hafez from his Divan.

You will be given the first verse of a couplet from a Hafez ghazal, followed by a version of the second verse in which the word order has been shuffled. Your task is to identify the true second verse from the Divan of Hafez that corresponds to this couplet, and return it without any modification.

Please strictly follow these instructions:
1- Do not generate any verse; only retrieve it from the authentic text of the Divan of Hafez.
2- The verse must exactly match the original text of Hafez; do not alter any word or punctuation.
3- Restore the correct verse only by rearranging the order of the words provided in the shuffled version of the second verse; adding or removing any word or punctuation is strictly prohibited.
4- Return the output strictly in the following format and do not include any additional text:
<answer>Exact second verse</answer>

Example 1:
Input:
<first_verse>And if you seek the way of annihilation to cast the world aside</first_verse>
<shuffled_second_verse>collapse every soul scatter your hair from thousands of</shuffled_second_verse>
Output:
<answer>Scatter your hair so that thousands of souls collapse from each strand</answer>

Example 2:
Input:
<first_verse>Your mischievous eyes have thrown China and Ethiopia into disorder</first_verse>
<shuffled_second_verse>hair China India to and Machin given tribute your</shuffled_second_verse>
Output:
<answer>To the China of your hair, Machin and India have paid tribute</answer>

Input first verse:
<first_verse>At the moment of parting from you, I wept so intensely</first_verse>

Shuffled second verse:
<shuffled_second_verse>face light eye far remained from me your</shuffled_second_verse>

Table 17: Example prompt for completing the second verse of a ghazal, given the poet's name (Hafez) and a word-shuffled version of the second verse.

شما یک مدل زبانی هستید که برای بازیابی مصراع دوم دقیق و اصیل حافظ از دیوان او طراحی شده‌اید.

به شما مصراع اول از یکی از ابیات غزل‌های حافظ داده می‌شود. سپس یک پرسش سه‌گزینه‌ای مطرح می‌شود که در آن، مصراع دوم واقعی و موجود در دیوان حافظ، که مربوط به همین مصراع اول است، پرسیده می‌شود.

وظیفه‌ی شما این است که تنها یک گزینه‌ی صحیح را انتخاب کنید. پاسخ شما باید فقط شامل حرف گزینه‌ی درست (الف، ب یا ج) باشد و هیچ متن یا توضیح اضافی ارائه نکنید.

مصراع اول:
<first_verse>خیال زلف تو پختن نه کار هر خامیست</first_verse>

گزینه‌ها:
الف) که زیر سلسله رفتن طریق عیاریست
ب) که این لطیفه عشقم ز رهروی یادست
ج) که رستگاری جاوید در کم آزاریست

English Translation:
You are a language model designed to retrieve the exact and authentic second verse of Hafez from his Divan.

You will be given the first verse of a couplet from a Hafez ghazal. You will then be presented with a multiple-choice question asking for the true second verse from the Divan of Hafez that corresponds to this first verse.

Your task is to select only one correct option. Your response must contain only the letter corresponding to the correct option (A, B, or C), and no additional text or explanation.

Input first verse:
<first_verse>To dwell upon your tress is not the work of every novice</first_verse>

Options:
A) For to go beneath the chain is the way of the daring
B) For this subtlety of my love comes from spiritual wayfaring
C) For eternal salvation lies in causing little harm

Table 18: Example prompt for selecting the authentic second verse of a ghazal's couplet, given the poet's name (Hafez)

شما یک مدل زبانی هستید که برای بازیابی مصراع دوم دقیق و اصیل حافظ از دیوان او طراحی شده‌اید.

به شما مصراع اول یکی از ابیات غزل‌های حافظ داده می‌شود و نثر روان فارسی همان بیت نیز ارائه می‌گردد. سپس یک پرسش سه‌گزینه‌ای مطرح می‌شود که در آن، مصراع دوم واقعی و موجود در دیوان حافظ، که مربوط به همین بیت است، پرسیده می‌شود.

وظیفه‌ی شما این است که تنها یک گزینه‌ی صحیح را انتخاب کنید. پاسخ شما باید فقط شامل حرف گزینه‌ی درست (الف، ب یا ج) باشد و هیچ متن یا توضیح اضافی ارائه نکنید.

مصراع اول:
<first_verse>سخن عشق نه آنست که آید به زبان</first_verse>

نثر بیت:
<couplet_prose>سخن عشق نباید که بر زبان آید، پس ای ساقی، می بده و این گفتگو را کوتاه کن.</couplet_prose>

گزینه‌ها:
الف) کاین کرامت سبب حشمت و تمکین من است
ب) ساقیا می ده و کوتاه کن این گفت و شنفت
ج) ناز کم کن که در این باغ بسی چون تو شکفت

English Translation:
You are a language model designed to retrieve the exact and authentic second verse of Hafez from his Divan.

You will be given the first verse of a verse from a Hafez ghazal, along with a fluent Persian prose rendering of the same couplet. You will then be presented with a multiple-choice question asking for the true second verse from the Divan of Hafez that corresponds to this couplet.

Your task is to select only one correct option. Your response must contain only the letter corresponding to the correct option (A, B, or C), and no additional text or explanation.

Input first verse:
<first_verse>The discourse of love is not something that can be spoken aloud</first_verse>

Couplet prose:
<couplet_prose>The matter of love should not be spoken openly; therefore, O cupbearer, pour the wine and bring this conversation to an end.</couplet_prose>

Options:
A) For this grace is the source of my dignity and stature
B) O cupbearer, pour the wine and shorten this conversation
C) Be less coquettish, for many like you have blossomed in this garden

Table 19: Example prompt for selecting the authentic second verse of a ghazal's couplet, given the poet's name (Hafez) and couplet prose.

شما یک مدل زبانی هستید که برای بازیابی مصراع دوم دقیق و اصیل حافظ از دیوان او طراحی شده‌اید.

به شما مصراع اول از یکی از ابیات غزل‌های حافظ داده می‌شود. سپس یک پرسش دوگزینه‌ای مطرح می‌شود که در آن، مصراع دوم واقعی و موجود در دیوان حافظ، که مربوط به همین مصراع اول است، پرسیده می‌شود.

وظیفه‌ی شما این است که تنها یک گزینه‌ی صحیح را انتخاب کنید. پاسخ شما باید فقط شامل حرف گزینه‌ی درست (الف یا ب) باشد و هیچ متن یا توضیح اضافی ارائه نکنید.

مصراع اول:
<first_verse>در گلستان ارم دوش چو از لطف هوا</first_verse>

گزینه‌ها:
الف) زلف سنبل به نسیم سحری می‌آشفت
ب) به زلف نسیم سحری سنبل می‌آشفت

English Translation:
You are a language model designed to retrieve the exact and authentic second verse of Hafez from his Divan.

You will be given the first verse of a couplet from a Hafez ghazal. You will then be presented with a two-choice question asking for the true second verse from the Divan of Hafez that corresponds to this first verse.

Your task is to select only one correct option. Your response must contain only the letter corresponding to the correct option (A or B), and no additional text or explanation.

Input first verse:
<first_verse>Last night in the garden of Eram, through the gentleness of the breeze</first_verse>

Options:
A) The hyacinth's tress was tousled by the dawn breeze
B) By the tress of the dawn breeze, the hyacinth was tousled

Table 20: Example prompt for selecting the authentic second verse of a ghazal's couplet, given the poet's name (Hafez) and a word-shuffled second verse as a distractor.

You are a language model designed to retrieve the exact and authentic second line from Shakespeare's Sonnets.

You are given the first line of a couplet from one of Shakespeare's sonnets. Your task is to identify and return the real second line that corresponds to this first line, exactly as it appears in the authoritative text of Shakespeare's Sonnets, without any modification.

Please strictly follow the instructions below:
1- Do not generate or invent any line; retrieve only from the authentic text of Shakespeare's Sonnets.
2- The returned line must exactly match the original wording and punctuation; do not alter, omit, or normalize any word or symbol.
3- Return the output strictly in the following format and do not provide any additional explanation:
<answer>exact second line from Shakespeare's Sonnets</answer>

First line:
<first_line>After my death, dear love, forget me quite,</first_line>

Table 21: Example prompt for completing the second line of a sonnet's couplet, given the poet's name (Shakespeare).

You are a language model designed to retrieve the exact and authentic second line from Shakespeare's Sonnets.

You are given the first line of a couplet from one of Shakespeare's sonnets, followed by two words taken from the second line of the same couplet. Your task is to identify and return the real second line that corresponds to this couplet, exactly as it appears in the authoritative text of Shakespeare's Sonnets, without any modification.

Please strictly follow the instructions below:
1- Do not generate or invent any line; retrieve only from the authentic text of Shakespeare's Sonnets.
2- The returned line must exactly match the original wording and punctuation; do not alter, omit, or normalize any word or symbol.
3- Return the output strictly in the following format and do not provide any additional explanation:
<answer>verbatim second line from Shakespeare's Sonnets</answer>

First line:
<first_line>Your monument shall be my gentle verse,</first_line>

Words:
["eyes", "o'er-read"]

Table 22: Example prompt for completing the second line of a sonnet's couplet, given the poet's name (Shakespeare) and salient words of the second line.

You are a language model designed to retrieve the exact and authentic second line from Shakespeare's Sonnets.

You are given the first line of a couplet from one of Shakespeare's sonnets, followed by a version of the second line from the same couplet in which the word order has been shuffled. Your task is to identify and return the real second line that corresponds to this couplet, exactly as it appears in the authoritative text of Shakespeare's Sonnets, without any modification.

Please strictly follow the instructions below:
1- Do not generate or invent any line; retrieve only from the authentic text of Shakespeare's Sonnets.
2- The returned line must be completely identical to the original line in wording and punctuation; do not change, normalize, add, or remove any word or symbol.
3- Restore the correct second line solely by reordering the words present in the provided shuffled version; adding or deleting any word or symbol is strictly prohibited.
4- Return the output strictly in the following format and do not provide any additional explanation:
<answer>exact second line from Shakespeare's Sonnets</answer>

Example 1:
Input:
<first_line>Rough winds do shake the darling buds of May,</first_line>
<shuffled_second_line>lease summer's short and too all a date: hath</shuffled_second_line>
Output:
<answer>And summer's lease hath all too short a date:</answer>

Example 2:
Input:
<first_line>Love alters not with his brief hours and weeks,</first_line>
<shuffled_second_line>it to of bears but edge even the doom. out</shuffled_second_line>
Output:
<answer>But bears it out even to the edge of doom.</answer>

First line:
<first_line>As victors of my silence cannot boast;</first_line>

Second line with shuffled word order:
<shuffled_second_line>not was fear I any of from thence: sick</shuffled_second_line>

Table 23: Example prompt for completing the second line of a sonnet's couplet, given the poet's name (Shakespeare) and a word-shuffled version of the second line.

Failure Category	Model	Scenario	Expected Output	Model Output
Hallucination	GPT-5.2	Name	سهی قدان سیه چشم ماه سیما را	چرا ز دوست به دشمن چنین جدایی نیست
Hallucination	Gemma 3 12B	+Prose	تفقـدی نکند طوطی شکرخا را	چون تو نیایی دلبری چه سود مرا
Prosaic Paraphrase	Claude Sonnet 4.5	+Prose	به یاد دار محبان باده پیما را	ز عاشقان بی‌بهره خود یادی کن
Prosaic Paraphrase	DeepSeek V3.2	+Para	نمی‌کنی به ترحم، نِطاق سلسله سست	تو همچنان بی‌شفقی و از جور نمی‌کاهی
Incomplete Completion	DeepSeek R1	+Salient	زلف سنبل به نسیم سحری می‌آشفت	زلف سنبل به شکن بر شاخ گل می‌زد خوشه
Incomplete Completion	Gemma 3 27B	+Shuffle	که کس نگشود و نگشاید به حکمت این معما را	و کس نگشاید حکمت این معما را
Contextual Confusion	Llama 4 Scout	+Name	تبارک الله از این فتنه‌ها که در سر ماست	که در طریق عشقبازی امن و آسایش بلاست
Contextual Confusion	DeepSeek R1	+Name	چو حافظ بنده و هندوی فرخ	ز هر چه رنگ تعلق پذیرد آزاد است

Table 24: Illustrative failure modes observed across models and completion scenarios.

Statistic	Persian Ghazals	English Sonnets
Number of Couplets	438	350
Length (characters)		
First Verse/Line (mean ± std)	31.5 ± 3.4	42.4 ± 3.8
Second Verse/Line (mean ± std)	31.9 ± 3.7	42.3 ± 4.2
Word Count		
First Verse/Line (mean ± std)	7.8 ± 1.3	8.3 ± 1.1
Second Verse/Line (mean ± std)	7.7 ± 1.4	8.1 ± 1.1
Lexical Diversity		
Vocabulary Size (unique words)	2,093	1,696
Type–Token Ratio (TTR)	0.31	0.30
Hapax Legomena Ratio	0.19	0.19

Table 25: Structural statistics for the Persian ghazals and English sonnets.

شما یک مدل زبانی هستید که برای ارزیابی بازنویسی غزل‌های حافظ به نثر طبیعی طراحی شده‌اید.
به شما یک غزل از دیوان حافظ داده می‌شود. سپس «نثر مرجع» آن غزل بیت‌به‌بیت و یک «نثر ارائه‌شده» بیت‌به‌بیت ارائه می‌شود. وظیفه‌ی شما این است که نثر ارائه‌شده را در ابعاد مختلف زیر ارزیابی کنید.

ابعاد ارزیابی (برای هر بیت)

1- دقت معنایی (Semantic Fidelity):

آیا نثر ارائه‌شده به‌درستی و به‌طور کامل معنای بیت اصلی را بدون حذف یا تغییر منتقل کرده است؟

- امتیاز 5: معنای بیت به‌صورت دقیق و کامل و بدون تحریف منتقل شده است.
- امتیاز 3: کلیت معنای بیت منتقل شده، ولی برخی جزئیات یا نکات ریز از بین رفته‌اند.
- امتیاز 1: بخش‌هایی از معنای بیت حذف یا تحریف شده‌اند.

2- طبیعی بودن و روانی (Naturalness & Fluency):

آیا نثر ارائه‌شده روان، طبیعی و شبیه زبان فارسی معیار امروزی است؟

- امتیاز 5: نثر کاملاً روان، طبیعی و شبیه گفتار یا نوشتار روزمره است.
- امتیاز 3: نثر نسبتاً روان است اما گاه نشانه‌هایی از زبان غیرطبیعی دارد.
- امتیاز 1: نثر مصنوعی، ناهماهنگ یا دشوارفهم است.

3- وضوح و سادگی (Clarity & Simplicity):

آیا نثر به‌روشنی و سادگی مفهوم بیت را منتقل کرده و از ابهام یا پیچیدگی خودداری کرده است؟

- امتیاز 5: نثر کاملاً روشن، شفاف و بدون ابهام است.
- امتیاز 3: نثر نسبتاً واضح است اما می‌توانست ساده‌تر و شفاف‌تر باشد.
- امتیاز 1: نثر مبهم یا گیج‌کننده است.

4- پرهیز از تفسیر اضافی (Interpretation Avoidance):

آیا نثر فقط معنای بیت را بازگو کرده و از افزودن تفسیر، توضیح یا تحلیل فراتر از متن اصلی خودداری کرده است؟

- امتیاز 5: نثر دقیقاً در محدوده‌ی معنای بیت باقی مانده و هیچ تفسیر یا توضیح اضافی ندارد.
- امتیاز 3: نثر گاهی اندکی از معنا فراتر رفته، اما کلیت معنا را حفظ کرده است.
- امتیاز 1: نثر شامل تفسیر یا توضیحات فراتر از معنی بیت است.

5- حفظ لحن بی‌طرف (Neutral Tone Preservation):

آیا نثر از هرگونه احساس، قضاوت یا ارزش‌داوری اضافی دوری کرده و لحنی خنثی دارد؟

- امتیاز 5: نثر کاملاً بی‌طرف و خنثی است.
- امتیاز 3: نثر غالبا بی‌طرف است اما برخی واژه‌ها لحن را اندکی تغییر داده‌اند.
- امتیاز 1: نثر دارای لحن احساسی یا جانبدارانه است.

6- کنترل طول و اختصار (Length & Brevity Control):

آیا طول نثر متناسب با بیت است و از توضیحات اضافی یا جزئیات غیرضروری خودداری کرده است؟

- امتیاز 5: نثر دقیق، مختصر و متناسب با بیت است.
- امتیاز 3: نثر طول مناسبی دارد ولی می‌توانست اندکی فشرده‌تر باشد.
- امتیاز 1: نثر طولانی و پر از توضیحات غیرضروری است.

برای هر بیت:

- برای هر بعد، یک امتیاز بین 1 تا 5 اختصاص دهید (1 = بسیار ضعیف، 5 = عالی).
- یک امتیاز نهایی (بین 1 تا 5) ارائه دهید که نشان‌دهنده‌ی ارزیابی کلی شما از آن بیت باشد.

خروجی را دقیقاً در قالب زیر ارائه کنید:

```
[
 {
  "couplet": "بیت اول",
  "scores": {
   "semantic_fidelity_score": <1-5>,
   "naturalness_&_fluency_score": <1-5>,
   "clarity_&_simplicity_score": <1-5>,
   "interpretation_avoidance_score": <1-5>,
   "neutral_tone_preservation_score": <1-5>,
   "length_&_brevity_control_score": <1-5>,
   "overall_score": <1-5>
  }
 },
 {
  "couplet": "بیت دوم",
  "scores": {
   "semantic_fidelity_score": <1-5>,
   "naturalness_&_fluency_score": <1-5>,
   "clarity_&_simplicity_score": <1-5>,
   "interpretation_avoidance_score": <1-5>,
   "neutral_tone_preservation_score": <1-5>,
   "length_&_brevity_control_score": <1-5>,
   "overall_score": <1-5>
  }
 },
 ...
]
```

ورودی:

```json
[
 {
  "couplet": "شربتی از لب لعلش نچشیدیم و برفت / روی مه پیکر او سیر ندیدیم و برفت",
  "reference_prose": "از لب لعل یار، شربتی نچشیدیم، و چهره زیبای چون ماه او را سیر ندیدیم که رفت.",
  "submitted_prose": "هنوز از لبان سرخ و شیرینش شربتی ننوشیده بودیم که رفت و هنوز از دیدن چهره‌ی ماه‌مانندش سیر نشده بودیم که از ما جدا شد."
 },
 {
  "couplet": "گویی از صحبت ما نیک به تنگ آمده بود / بار بربست و به گردش نرسیدیم و برفت",
  "reference_prose": "گویا او از همصحبتی با ما سخت به تنگ آمده بود که بار سفر را بست و ما به گرد او نرسیدیم و رفت.",
  "submitted_prose": "به نظر می‌رسید که از همنشینی با ما به شدت ناراحت و خسته شده بود، چنان عجله داشت که وسایلش را جمع کرد و رفت و ما حتی فرصت نیافتیم که دور و برش بگردیم و با او باشیم."
 },
 {
  "couplet": "پس که ما فاتحه و حرز یمانی خواندیم / وز پی‌اش سوره اخلاص دمیدیم و برفت",
  "reference_prose": "ما برای نصر او سوره فاتحه و دعای حرز یمانی خواندیم و به دنبالش سوره اخلاص را دمیدیم اما باز هم او رفت.",
  "submitted_prose": "ما برای جلب توجه و حفظ او بسیار دعا و ورد خواندیم و سوره‌های مختلفی برای حفاظتش دمیدیم، اما با وجود همه این تلاش‌ها، او باز هم رفت و ما را ترک کرد."
 },
 {
  "couplet": "عشوه دادند که بر ما گذری خواهی کرد / دیدی آخر که چنین عشوه خریدیم و برفت",
  "reference_prose": "ما را فریب دادند که از کنار ما می‌گذری، دیدی که ما چگونه این فریب را پذیرفتیم اما او رفت؟",
  "submitted_prose": "عشوه‌بازی کردند و وعده دادند که به دیدار ما خواهی آمد، اما ببین سرانجام چگونه این وعده‌های فریبنده را باور کردیم و او بدون وفا به قول رفت."
 },
 {
  "couplet": "شد چمان در چمن حسن و لطافت لیکن / در گلستان وصالش نچمیدیم و برفت",
  "reference_prose": "یار خرامان به چمن زیبایی و لطافت پا نهاد و ما در گلستان وصال او نرمک نرمک گام ننهادیم و او رفت.",
  "submitted_prose": "چشمان ما در باغ زیبایی و ظرافت او گشت و گذار کرد، اما متأسفانه در بوستان پیوند و وصال او قدم نگذاشتیم و او رفت."
 },
 {
  "couplet": "همچو حافظ همه شب ناله و زاری کردیم / کای دریغا به وداعش نرسیدیم و برفت",
  "reference_prose": "تمام شب را چون حافظ ناله و افغان کردیم که ای افسوس که ما فرصت وداع نیافتیم و او رفت.",
  "submitted_prose": "مانند حافظ تمام شب گریه و ناله کردیم، اما افسوس که حتی فرصت خداحافظی با او را هم نیافتیم و رفت."
 }
]
```

Figure 2: Example prompt for evaluating LLM-generated prose using Gemini 3 Pro as the LLM-as-Judge.

You are a language model designed to evaluate the rewriting of Hafez's ghazals into natural prose.
You will be given a ghazal from the Divan of Hafez. A couplet-by-couplet "reference prose" version of the ghazal and a couplet-by-couplet "submitted prose" version will then be provided. Your task is to evaluate the submitted prose across the following dimensions.

Evaluation Dimensions (for each couplet)

1- Semantic Fidelity:
Does the submitted prose correctly and completely convey the meaning of the original couplet without omission or distortion?
- Score 5: The meaning of the couplet is conveyed accurately, completely, and without distortion.
- Score 3: The overall meaning is conveyed, but some details or nuances are lost.
- Score 1: Parts of the meaning are omitted or distorted.

2- Naturalness & Fluency:
Is the submitted prose fluent, natural, and similar to contemporary standard Persian?
- Score 5: The prose is completely fluent, natural, and resembles everyday spoken or written language.
- Score 3: The prose is relatively fluent, but occasionally contains unnatural phrasing.
- Score 1: The prose is artificial, inconsistent, or difficult to understand.

3- Clarity & Simplicity:
Does the prose communicate the meaning of the couplet clearly and simply while avoiding ambiguity or unnecessary complexity?
- Score 5: The prose is fully clear, transparent, and unambiguous.
- Score 3: The prose is relatively clear, but could have been simpler or more transparent.
- Score 1: The prose is ambiguous or confusing.

4- Interpretation Avoidance:
Does the prose only restate the meaning of the couplet without adding interpretation, explanation, or analysis beyond the original text?
- Score 5: The prose remains strictly within the meaning of the couplet and contains no additional interpretation or explanation.
- Score 3: The prose occasionally goes slightly beyond the meaning, while still preserving the overall sense.
- Score 1: The prose contains interpretation or explanations beyond the meaning of the couplet.

5- Neutral Tone Preservation:
Does the prose avoid introducing additional emotion, judgment, or subjective evaluation and maintain a neutral tone?
- Score 5: The prose is completely neutral and objective.
- Score 3: The prose is mostly neutral, but some word choices slightly alter the tone.
- Score 1: The prose contains emotional or biased language.

6- Length & Brevity Control:
Is the length of the prose appropriate for the couplet, avoiding unnecessary explanations or details?
- Score 5: The prose is precise, concise, and appropriately proportional to the couplet.
- Score 3: The prose has a reasonable length, but could have been somewhat more concise.
- Score 1: The prose is overly long and filled with unnecessary explanations.

For each couplet:
- Assign a score between 1 and 5 for each evaluation dimension (1 = very poor, 5 = excellent).
- Provide one final score (between 1 and 5) representing your overall evaluation of the couplet.

Return the output exactly in the following format:

```
[
 {
   "couplet": "First couplet",
   "scores": {
     "semantic_fidelity_score": <1-5>,
     "naturalness_&_fluency_score": <1-5>,
     "clarity_&_simplicity_score": <1-5>,
     "interpretation_avoidance_score": <1-5>,
     "neutral_tone_preservation_score": <1-5>,
     "length_&_brevity_control_score": <1-5>,
     "overall_score": <1-5>
   }
 },
 {
   "couplet": "Second couplet",
   "scores": {
     "semantic_fidelity_score": <1-5>,
     "naturalness_&_fluency_score": <1-5>,
     "clarity_&_simplicity_score": <1-5>,
     "interpretation_avoidance_score": <1-5>,
     "neutral_tone_preservation_score": <1-5>,
     "length_&_brevity_control_score": <1-5>,
     "overall_score": <1-5>
   }
 },
 ...
]
```

Input:

```
```json
[
 {
   "couplet": "We did not taste a sip from her ruby lips before she left / We did not gaze enough upon her moon-like face before she left",
   "reference_prose": "We did not taste a sip from the beloved's ruby lips, nor did we fully behold her moon-like face before she departed.",
   "submitted_prose": "Before we could drink from her sweet red lips, she departed, and before we had enough of seeing her moon-like face, she left us."
 },
 {
   "couplet": "It seemed she had grown weary of our company / She packed her belongings and left before we could reach her",
   "reference_prose": "It seemed that she had become greatly tired of our companionship, so she prepared for departure and left before we could catch up to her.",
   "submitted_prose": "She appeared deeply exhausted and upset by our companionship; in great haste she gathered her belongings and left before we even had the chance to remain by her side."
 }
]
```

Figure 3: English translation of the prompt for evaluating LLM-generated prose (original in Figure 2).
```

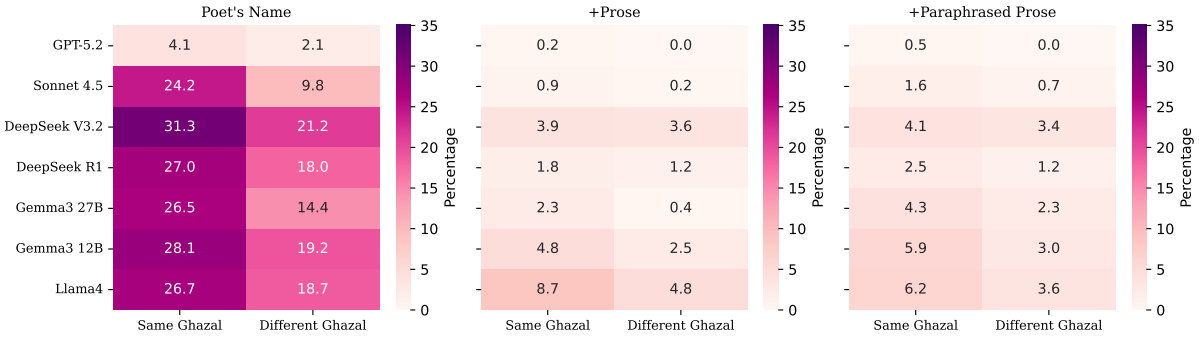

Figure 4: Percentage distribution of model selections among distractor options in the multiple-choice verse recognition task under varying contextual cues.