Fann or Flop: A Multigenre, Multiera Benchmark for Arabic Poetry Understanding in LLMs

Wafa Alghallabi[1†] Ritesh Thawkar[1†] Sara Ghaboura[1†] Ketan More[1†]
Omkar Thawakar[1†] Hisham Cholakkal[1] Salman Khan[1,2] Rao Muhammad Anwer[1,3]
[1]Mohamed bin Zayed University of AI, [2]Australian National University, [3]Aalto University
{wafa.alghallabi, sara.ghaboura, omkar.thawakar}@mbzuai.ac.ae
https://mbzuai-oryx.github.io/FannOrFlop/

Abstract

Arabic poetry is one of the richest and most culturally rooted forms of expression in the Arabic language, known for its layered meanings, stylistic diversity, and deep historical continuity. Although large language models (LLMs) have demonstrated strong performance across languages and tasks, their ability to understand Arabic poetry remains largely unexplored. In this work, we introduce *Fann or Flop*, the first benchmark designed to assess the comprehension of Arabic poetry by LLMs in 12 historical eras, covering 14 core poetic genres and a variety of metrical forms, from classical structures to contemporary free verse. The benchmark comprises a curated corpus of poems with explanations that assess semantic understanding, metaphor interpretation, prosodic awareness, and cultural context. We argue that poetic comprehension offers a strong indicator for testing how good the LLM understands classical Arabic through Arabic poetry. Unlike surface-level tasks, this domain demands deeper interpretive reasoning and cultural sensitivity. Our evaluation of state-of-the-art LLMs shows that most models struggle with poetic understanding despite strong results on standard Arabic benchmarks. We release *Fann or Flop* [1] along with the evaluation suite [2] as an open-source resource to enable rigorous evaluation and advancement for Arabic language models.

Figure 1: **Chronological Wheel of Arabic Poetic Eras.** This circular taxonomy visualizes the evolution of Arabic poetry across 12 major historical eras, from the Pre-Islamic and Transitional periods through the Abbasid, Andalusian, and Mamluk dynasties, up to the Modern era. The layout reflects both temporal flow and the rich cultural shifts that shaped poetic expression. Detailed taxonomy by genre, meter, and notable poets presented in Table 2.

1 Introduction

Arabic is among the world's most lexically rich languages, with a vocabulary exceeding 12.3 million words—far surpassing that of most modern languages (AlSuyuti, 15th Century; Andrews, 2024). A single word can convey multiple meanings, varied pronunciations, and diverse interpretations, reflecting the language's profound semantic complexity. Despite its official status in 27 countries—ranking third in global geopolitical presence (wikipedia, 2025)—only a fraction of this lexicon remains in common use today.

To unify communication across its many dialects, Modern Standard Arabic (MSA) emerged in the late 19th and early 20th centuries as a formal register (oussama, 2024). Today, it is the primary language of education, media, and governance in the Arab world. Although linguists distinguish Classical Arabic (CA) from MSA, native speakers generally view them as a unified formal variety (wikipediaArabic, 2025). Nevertheless, even the most comprehensive Arabic dictionaries—such as Lisan al-Arab (Manzur, 14th Century), Taj al-Lugha (al Jawhari, 10th Century), and

[1]https://huggingface.co/datasets/omkarthawakar/FannOrFlop
[2]https://github.com/mbzuai-oryx/FannOrFlop
[†]Equal contribution.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 20224–20244
November 4-9, 2025

Feature	AQMAR	Tafsir	Ashaar	Ara Bench	Arabic SQuAD	ARCD	AraBERT Collection	CAMeL Corpus	Tash keela	PADIC	MADAR	Fann or Flop
Dialectal Variety	✗	✗	✗	✓	✗	✗	✓	✗	✗	✓	✓	✓
Poetic Device Annotation	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Verse/Sentence-Level Annotation	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓
Temporal/Historical Context	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓
QA-Style Task Format	✗	✗	✗	✗	✓	✓	✓	✗	✗	✗	✗	✓
Open-Source	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: **Comparison of key Arabic NLP datasets.** Existing Arabic NLP resources typically address isolated features such as dialectal coverage, QA formats, or classical text processing. In contrast, **Fann or Flop** uniquely integrates multiple underrepresented dimensions (i.e. dialectal diversity, poetic device annotation, verse-level granularity, temporal grounding, and a QA-style evaluation format) positioning it as the first comprehensive benchmark for Arabic poetry understanding. AQMAR (Mohit et al., 2012), Tafsir (Ahmed et al., 2022), Ashaar (Alyafeai et al., 2023), AraBench (Sajjad et al., 2020), Arabic-SQuAD (Mozannar et al., 2019), ARCD (Mozannar et al., 2019), AraBERT Collection (Antoun et al., 2020), CAMeL Corpus (Abdul-Mageed et al., 2020; Khalifa et al., 2018), Tashkeela (Zerrouki and Balla, 2017), PADIC (Meftouh et al., 2015), MADAR (Bouamor et al., 2018).

al-Mu'jam al-Mu'asir (alsharekh, 2019; Ar-Riyadh, 2025)—cover only a small portion of the historical corpus, revealing the inherent challenges of Arabic lexicography and sociolinguistic narrowing of usage. Within this broader linguistic context, Arabic poetry has served as a repository of cultural and intellectual expression from the older era to the modern time. Poetic forms such as long odes (qasida), lyrical love poems (ghazal), elegies (ritha'), strophic songs (muwashsha), and vernacular verse (zajal) are marked by distinct metrical, rhetorical, and performative characteristics. While contemporary poets explore free verse and modernist motifs, classical forms continue to exert a strong literary and cultural influence.

Recent advances in LLM, such as GPT (Chen et al., 2025), LLaMA (Touvron et al., 2023), AceGPT (Huang et al., 2023), Jais (Sengupta et al., 2023), and Falcon (Malartic et al., 2024), have demonstrated impressive multilingual capabilities, including Arabic. However, most Arabic natural language processing (NLP) benchmarks focus on tasks such as sentiment analysis, question answering, or recognition of named entities (Antoun et al., 2020; Abdul-Mageed et al., 2021; Obeid et al., 2020), typically in MSA or dialectal prose. These benchmarks often miss the linguistic depth and cultural nuances that are inherent in Arabic poetry. As LLMs are increasingly evaluated for their ability to handle complex linguistic phenomena, such as metaphor, figurative language, and stylistic nuance, their limitations become evident (Liu et al., 2022; Bisk et al., 2020). The FLUTE benchmark (Chakrabarty et al., 2022) and the FigLang 2024 workshop (FIGLANG202, 2024) have reaffirmed that non-literal language understanding remains a significant challenge. This challenge is particularly acute in Arabic, where poetry is densely layered with intertextuality and cultural symbolism. Arabic poetry thus provides a uniquely demanding testbed for assessing deep linguistic in language models.

To address this gap, we introduce *Fann or Flop*, the first benchmark dedicated to evaluating LLMs' understanding of Arabic poetry. Our benchmark comprises 6,984 poem-explanation pairs curated from 12 distinct historical poetic eras (see Figure 1), which can be broadly seen as spanning three major historical periods: pre-Islamic, classical, and contemporary. It covers 14 poetic genres and includes a range of metrical forms, as detailed in Table 1. Each sample is manually verified by native Arabic speakers with domain knowledge to ensure linguistic authenticity and interpretive accuracy. This rich and diverse collection makes Fann or Flop a reliable benchmark for evaluating deep cultural and literary reasoning in Arabic NLP. Figure 2 represents the examples from our proposed Fann or Flop dataset, showcasing the diversity of eras, genres, and poetic styles covered.

Our goal is to provide a diagnostic on how well your language model understands and interprets genuine Classical Arabic. Unlike general text, poetry requires sensitivity to rhetorical devices, metrical patterns, and sociohistorical context, making it a rigorous and culturally grounded indicator of language proficiency. Our benchmark serves as a clear indicator of whether a model has truly been exposed to and internalized high-quality Arabic content, offering a focused lens into its cultural and linguistic depth. We evaluate a range of open-source and commercial LLMs using Fann or Flop and find that, despite strong performance on con-

Era	Approx. Years	Genres (Theme)	Meter	Notable Poets
Pre-Islamic (Jahiliyyah)	Until 610 CE	Satire, Separation, Wisdom	At-Tawil, Al-Kamel, Al-Basit	Imru al-Qays, Antarah ibn Shaddad, Zuhayr ibn Abi Sulma
Transitional Poets (Mukhadramun)	Late 6th – Early 7th c.	Praise, Apology, Religious	Ar-Rojz, Ar-Ramel	Hassan ibn Thabit, Labid ibn Rabi'a, Al-Khansa
Islamic	610–661 CE	Religious, Wisdom, Patience	Al-Madid, Al-Kamel	Abu Sallama Al-Arhabi, Onayf Ibn Kitra
Umayyad	661–750 CE	Love, Satire, Political	At-Tawil, Al-Wafer, As-Sari'	Jarir, al-Farazdaq, al-Akhtal
Abbasid	750–1258 CE	Praise, Elegy, Wisdom	Al-Basit, Kamel, Al-Monsareh, Al-Moktadab	Abu Nuwas, al-Mutanabbi, al-Buhturi, Abu Tammam
Fatimid	909–1171 CE	Religious, Praise, Sadness	Ar-Rojz, Al-Mutakareb	Ibn Hayus, Abu al-Ala al-Ma'arri
Andalusian	756–1492 CE	Love, Longing, Wisdom	Mowachah, Al-Mowaliya, Al-Mohtath	Ibn Sahl Al-Andalusi, Ibn Zaydun, Ibn Khafaja
Ayyubid	1171–1250 CE	Religious, Praise, Elegy	Al-Kamel, Al-Khafif	Ibn al-Farid, Mohyiddine Bin Arabi
Mamluk	1250–1517 CE	Wisdom, Praise, Religious	Al-Wafer, Ar-Rojz	Bahaa'eddine Zuhair, Safiyueddine Alhilli
Between the Two Dynasties	1258–1517 CE	Religious, Wisdom, Reproach	Al-Mutadarek, Ar-Ramel	Bashar bn Burd
Ottoman	1517–1800 CE	Religious, Love, General	Al-Kamel, Al-khafif	Bnt Al-Shahna, Ibn Razka
Modern	19th c. – Present	Nationalism, Love, Social Justice	Free Meter	Ahmad Shawqi, Hafeth Ibrahim

Table 2: **Taxonomy of Arabic Poetic Eras with Genre and Meter Coverage.** This table provides a structured overview of 12 major eras in Arabic poetic history, detailing their approximate chronological spans, the most prominent poetic themes (genres) representative of each era, the dominant metrical patterns (Arabic *buḥūr*) used in poetic composition, and notable poets who exemplify the literary character of their time. The genre column highlights recurring thematic concerns such as satire, elegy, love, nationalism, and religious devotion, while the meter column showcases the classical metrical forms like *At-Tawil*, *Al-Kamel*, and *Ar-Rojz*, along with innovations such as free verse in the modern period. This taxonomy reflects the dynamic interplay between form, content, and historical context in shaping Arabic poetic expression.

ventional Arabic tasks, most models struggle with the interpretive depth required by poetry. These findings highlight the need for culturally informed benchmarks that better reflect the depth and diversity of Arabic. We release Fann or Flop as an open-source resource to support the development and evaluation of Arabic-capable language models.

2 The Fann or Flop Dataset

2.1 Dataset Taxonomy

To capture the linguistic, historical, and thematic richness of Arabic poetry, we construct an expert-verified taxonomy that organizes poems across both form and era. As illustrated in Figure 1 and detailed in Table 2, the taxonomy traces 12 distinct poetic eras, from the pre-Islamic period to modern times, encompassing 14 genres that capture the dominant styles, concerns, and historical contexts of each era. It illustrates how poetic expression evolved over the centuries.

This structured framework was carefully reviewed and validated by scholars specializing in Arabic language and literature to ensure both linguistic accuracy and contextual relevance. Their expertise helped align the taxonomy with established literary traditions while accommodating the nuances of classical and modern poetic forms. Beyond its utility for literary and philological analysis, the taxonomy serves as a robust foundation for

Figure 2: **Representative Poetic Samples Across Arabic Literary Eras.** This figure presents curated excerpts from Arabic poems spanning key historical eras, illustrating the evolution of language, themes, and stylistic expression. The Pre-Islamic sample reflects tribal valor and rhetorical precision; the Umayyad excerpt captures satire and social commentary; the Abbasid example highlights philosophical reflection and refined metaphorical use; the Transitional era showcases a poetic voice confronting injustice and advocating moral clarity; and the Andalusian selection reveals emotional openness and psychological depth through lyrical expression. Together, these samples provide insight into how Arabic poetry has adapted to diverse historical, cultural, and ideological contexts. Refer to Appendix F.1, Figure 9 for the GPT-4o-generated English translations of the Arabic poetic samples.

computational modeling. It enables more precise automatic genre classification and facilitates temporal contextualization across different eras of Arabic poetry, thereby supporting culturally informed and interpretable Arabic NLP research.

2.2 Data Collection

We curated Arabic poems from a well-established digital archive[3], which hosts a broad spectrum of poets, genres, and historical periods. A custom web scraper was developed to extract the poem texts along with associated metadata, including

[3] https://arabic-poetry.net

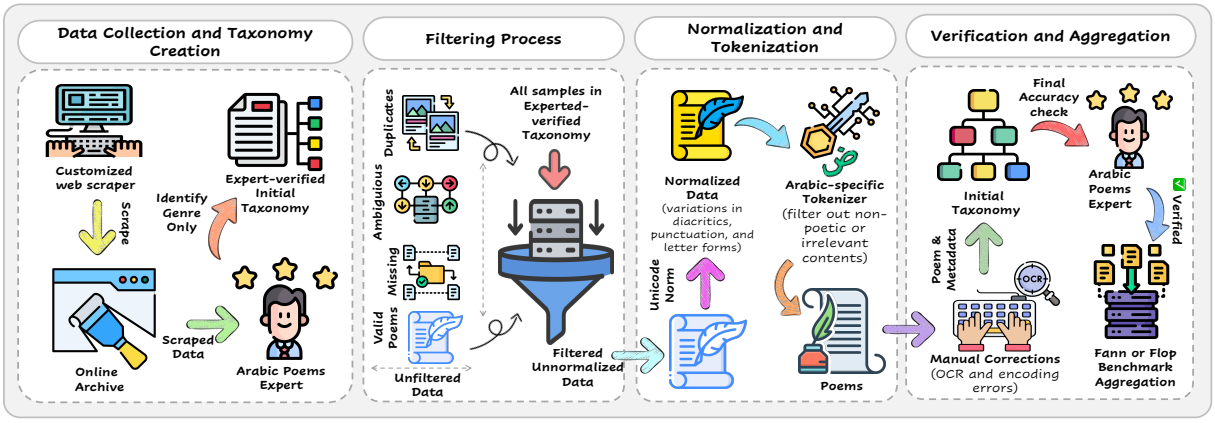


Figure 3: **Fann or Flop Pipeline.** Fann or Flop is built out of the multi-stage pipeline. It begins with scraping Arabic poems from a trusted online archive using a custom web scraper. Extracted poems are matched to an initial expert-verified taxonomy and filtered to remove duplicates, ambiguous metadata, and invalid entries. The filtered texts then undergo normalization (e.g., unifying diacritics, punctuation, and letter forms) and Arabic-specific tokenization, with non-poetic or irrelevant content excluded. Manual corrections are applied to fix OCR and encoding errors. In the final stage, linguistic experts verify each sample to ensure proper alignment with genre and era labels.

the poet's name, historical era, genre, meter, and poem's title. The resulting dataset extends across 12 distinct eras, from the pre-Islamic period to the modern era, and reflects a diverse range of poetic forms and styles. All entries were curated following our expert-verified taxonomy (see Table 2), ensuring consistency across genre and era classifications. This structured approach preserves both the linguistic richness and historical specificity of Arabic poetry, providing a valuable resource for research in both Arabic NLP and digital literary studies. By aligning each sample with a well-defined literary context, the dataset enables more accurate model evaluation and supports culturally grounded language understanding.

2.3 Data Filtering and Verification

To ensure data quality, consistency, and alignment with the expert-defined taxonomy, we applied a multi-step filtering and verification pipeline, illustrated in Figure 3. The process consisted of the following stages:

- **Duplicate and Metadata Filtering:** Starting with a collection of over 10,000 Arabic poems, we removed duplicate entries and discarded those with missing or ambiguous metadata, such as unknown poets or unspecified historical eras, resulting in a curated dataset of 6,984 high-quality poems.

- **Unicode Normalization:** All poems were standardized using Unicode normalization to address orthographic inconsistencies common in Arabic, including variations in diacritics, punctuation, and letter forms (e.g., alternate representations of *alif* and *ta marbuta*).

- **Text Tokenization and Content Filtering:** We applied an Arabic-specific tokenizer to segment the text accurately. Non-poetic or irrelevant content, such as editorial comments, footnotes, and prose fragments, was automatically excluded.

- **Manual Correction of Encoding Errors:** A sample subset of poems was manually reviewed to correct common OCR and encoding issues that were not resolved through automated preprocessing.

- **Expert Validation of Labels:** All genre and era annotations were reviewed by Arabic language and literature experts. This validation step ensured that each poem was accurately categorized in accordance with the taxonomy introduced in Section 2.1.

3 Fann or Flop Benchmark Evaluation

Evaluation Framework. To assess the quality of LLM-generated explanations for Arabic poetry, we adopt a multi-layered evaluation pipeline. This framework integrates (i) automatic lexical metrics, (ii) semantic and entailment modeling, (iii) human expert annotation (interpretive depth), and (iv)

	Model	BLEU	chrF(++)	BERTScore	Textual Entailment	Faithfulness/ Consistency	Fluency/ Grammaticality	Interpretive-Depth
Closed	GPT-4o-2024-08-06 (OpenAI, 2024)	0.0395	**0.2882**	**0.6410**	0.6775	3.92 (± 0.99)	4.96 (±0.20)	**7.52**
	GPT-4o-mini-2024-07-18 (OpenAI, 2024)	0.0395	0.2542	0.6124	0.4383	2.91 (±0.75)	4.28 (±0.57)	7.50
	Gemini-2.5-Flash (AI, 2025b)	0.0153	0.2618	0.6319	**0.7475**	**4.25** (±1.00)	**4.98** (±0.16)	7.22
	Gemini-2.0-Flash (AI, 2025a)	0.0395	0.2618	0.6393	0.7154	3.99 (±1.04)	4.95 (±0.22)	6.50
	Gemini-1.5-Pro (Reid et al., 2024)	0.0395	0.2618	0.6333	0.6180	3.59 (±1.00)	4.80 (±0.41)	5.38
	Fanar-Star (Team et al., 2025)	0.0138	0.1538	0.5677	0.6468	2.16 (±0.92)	3.40 (±0.76)	2.88
Open	Deepseek-V3 (Liu et al., 2024)	0.0395	0.2771	0.6335	0.5117	3.36 (±0.91)	**4.98** (±0.16)	4.75
	Deepseek-R1 (Guo et al., 2025)	0.0395	0.2771	0.6335	0.5117	3.38 (±0.92)	**4.98** (±0.16)	4.25
	Llama-3.3-70B (Meta AI, 2024)	0.0153	0.2618	0.6393	0.5364	2.51 (±0.90)	3.37 (±0.73)	7.20
	Qwen-3 (Team, 2025)	0.0296	0.2837	0.6158	0.6468	3.98 (±0.90)	4.73 (±0.45)	6.50
	Aya-Expanse (Dang et al., 2024)	0.0329	0.2771	0.6328	0.6468	3.76 (±0.90)	4.68 (±0.47)	5.88
	Jais (Sengupta et al., 2023)	0.0312	0.2698	0.6245	0.6023	3.21 (±0.88)	4.35 (±0.52)	5.35
	ALLaM-7B (Bari et al., 2024)	0.0119	0.0463	0.5375	0.5997	1.32 (±0.62)	2.11 (±0.89)	3.12
	AceGPT-v2-70B-Chat (Huang et al., 2023)	**0.0402**	0.0412	0.5759	0.6061	2.52 (±0.91)	3.46 ±0.95)	4.12

Table 3: **Comparison of closed and open-source models on the Arabic poem understanding task using both automatic and human evaluations.** BLEU, chrF(++), and BERTScore capture lexical and semantic similarity with reference explanations, while textual entailment assesses factual alignment. Human evaluation includes interpretive depth, while faithfulness and fluency are automatically judged using GPT-4o as a reference grader. Closed models like GPT-4o and Gemini-2.5-Flash achieve strong overall performance, while open models such as Deepseek-V3 and Aya-Expanse show promising consistency and interpretability. This benchmark highlights the potential of open models and the need for deeper cultural reasoning in Arabic poetic understanding.

	Model	Pre-Islamic	Transitional	Early Islamic	Umayyad	Abbasid	Fatimid
Closed	GPT-4o-2024-08-06 (OpenAI, 2024)	0.6285	0.6304	0.6341	0.6285	0.6421	0.6398
	GPT-4o-mini-2024-07-18 (OpenAI, 2024)	0.5980	0.6060	0.6134	0.5998	0.6125	0.6127
	Gemini-2.5-Flash (AI, 2025b)	0.6245	0.6264	0.6286	0.6253	0.6326	0.6282
	Gemini-2.0-Flash (AI, 2025a)	0.6290	0.6303	0.6326	0.6312	0.6404	0.6373
	Gemini-1.5-Pro (Reid et al., 2024)	0.6255	0.6293	0.6223	0.6278	0.6338	0.6307
	Fanar-Star (Team et al., 2025)	0.5694	0.5749	0.5695	0.5696	0.5720	0.5666
Open	Deepseek-V3 (Liu et al., 2024)	0.6225	0.6303	0.6311	0.6263	0.6313	0.6330
	Deepseek-R1 (Guo et al., 2025)	0.6271	0.6296	0.6321	0.6247	0.6324	0.6359
	Llama-3.3-70B (Meta AI, 2024)	0.5705	0.5703	0.5701	0.5668	0.5831	0.5719
	Qwen-3 (Team, 2025)	0.6111	0.6152	0.6129	0.6136	0.6164	0.6145
	Aya-Expanse (Dang et al., 2024)	0.6214	0.6232	0.6220	0.6232	0.6343	0.6294
	Jais (Sengupta et al., 2023)	0.6172	0.6218	0.6241	0.6183	0.6285	0.6239
	ALLaM-7B (Bari et al., 2024)	0.5786	0.5826	0.5917	0.5790	0.5862	0.5799
	AceGPT-v2-70B-Chat (Huang et al., 2023)	0.6194	0.6246	0.6329	0.6213	0.6261	0.6225

	Model	Andalusian	Ayyubid	Mamluk	Between Dynasties	Ottoman	Modern
Closed	GPT-4o-2024-08-06 (OpenAI, 2024)	0.6386	0.6440	0.6563	0.6440	0.6510	0.6487
	GPT-4o-mini-2024-07-18 (OpenAI, 2024)	0.6151	0.6167	0.6273	0.6176	0.6202	0.6140
	Gemini-2.5-Flash (AI, 2025b)	0.6297	0.6340	0.6421	0.6336	0.6415	0.6341
	Gemini-2.0-Flash (AI, 2025a)	0.6346	0.6409	0.6533	0.6414	0.6504	0.6441
	Gemini-1.5-Pro (Reid et al., 2024)	0.6313	0.6349	0.6409	0.6355	0.6443	0.6387
	Fanar-Star (Team et al., 2025)	0.5746	0.5684	0.5569	0.5831	0.5586	0.5392
Open	Deepseek-V3 (Liu et al., 2024)	0.6337	0.6404	0.6482	0.6393	0.6404	0.6368
	Deepseek-R1 (Guo et al., 2025)	0.6353	0.6404	0.6509	0.6408	0.6423	0.6373
	Llama-3.3-70B (Meta AI, 2024)	0.5791	0.5755	0.5935	0.5854	0.5797	0.5794
	Qwen-3 (Team, 2025)	0.6153	0.6163	0.6189	0.6160	0.6242	0.6149
	Aya-Expanse (Dang et al., 2024)	0.6289	0.6366	0.6475	0.6367	0.6393	0.6398
	Jais-30B-v3 (Sengupta et al., 2023)	0.6279	0.6321	0.6413	0.6307	0.6348	0.6316
	ALLaM-7B (Bari et al., 2024)	0.5876	0.5925	0.6004	0.5884	0.5933	0.5864
	AceGPT-v2-70B-Chat (Huang et al., 2023)	0.6168	0.6280	0.6466	0.6212	0.6205	0.6265

Table 4: **Era-wise Evaluation using BERTScore.** Model-wise performance breakdown using BERTScore evaluation across different Arabic poetic eras, evaluating understanding and generation quality within historical and stylistic contexts. The eras span from Pre-Islamic to Modern periods, offering a fine-grained analysis of model capabilities across evolving linguistic and cultural expressions. This table highlights gaps in temporal generalization and cultural grounding, motivating the need for era-aware training and evaluation in Arabic literary modeling.

LLM-as-Judge scoring. Together, these complementary layers capture both surface-level fidelity and the deeper interpretive demands of poetic understanding.

3.1 Automatic and Semantic Metrics

For automatic evaluation, we compute BLEU (Papineni et al., 2002) and chrF(++) (Popović, 2017) scores to quantify semantic and character-level overlap between model outputs and actual poem explanation references. While useful for consistency checks, these metrics are limited in capturing the nuanced variation allowed in literary interpretation.

To assess semantic alignment, we employ BERTScore (Zhang et al., 2019), leveraging Arabic-pretrained transformers such as AraBERT (Antoun et al., 2020) to quantify the semantic similarity between model-generated explanations and human-authored references. In addition, we incorporate Textual Entailment (TE) analysis using mDeBERTaV3 (He et al., 2021), a multilingual model fine-tuned for Natural Language Inference (NLI). This enables us to evaluate whether the reference explanation logically entails the generated output. Unlike surface-level similarity metrics, this approach provides a deeper measure of semantic consistency, capturing whether the generated interpretation remains faithful to the intended meaning of the expert-authored reference, even when expressed using different lexical or syntactic forms.

3.2 Human Evaluation and Inter-Annotator Agreement

To capture interpretive and literary nuance beyond automated metrics, we performed human evaluation on model-generated explanations. Annotators used a rubric-based scale (0–10) grounded in poetic analysis, comprising the following criteria:

- **Literal Comprehension (0–1):** Does the explanation correctly reflect the surface meaning of the poem?
- **Thematic and Emotional Depth (0–2):** Does it convey underlying themes, sentiment, or tone (e.g., longing, satire, mysticism)?
- **Cultural and Historical Appropriateness (0–2):** Does it demonstrate awareness of cultural, religious, or historical context?
- **Stylistic Sensitivity (0–3):** Does it acknowledge rhetorical and literary features such as metaphor, figurative language, rhythm, or imagery?
- **Expressiveness and Coherence (0–2):** Is the explanation clear, well-articulated, and stylistically appropriate in Arabic?

Annotators. Three native Arabic speakers with expertise in Arabic literature and linguistics voluntarily applied the evaluation rubric, reflecting a shared commitment to advancing Arabic linguistic research and strengthening LLM evaluation in culturally specific domains.

Rubric Guidelines. The evaluation rubric was collaboratively refined by the annotators to capture the cultural, literary, and linguistic dimensions of Arabic poetry understanding, ensuring a balanced and consistently applied framework (see Table 17 for a detailed breakdown).

Inter-Annotator Agreement (IAA). To evaluate the reliability of the interpretive depth scoring, 10% of the data was randomly sampled and Krippendorff's α was computed across the three annotators, resulting in an IAA of 87.7%. This level of agreement underscores the consistency of the scoring process and validates the robustness of the interpretive depth metric.

3.3 LLM-as-Judge Evaluation

Based on human-annotated interpretive depth scores (Section 3.2), GPT-4o (OpenAI, 2024) demonstrated the strongest performance among all open- and closed-source models (Table 3), thereby motivating its adoption as the LLM-as-judge. In this role, it evaluates explanation quality across two key dimensions: *Faithfulness/Consistency*—whether the explanation accurately reflects the poem without introducing hallucinated or misleading content—and *Fluency/Grammaticality*—whether it is grammatically correct, clear, and natural in Arabic.

4 Results and Analysis:

Table 3 presents the performance of both closed and open-source models on Arabic poem understanding using a combination of automatic metrics (BLEU, chrF(++), BERTScore, Textual Entailment, faithfulness and fluency) and human evaluation such as interpretive depth analysis. These metrics collectively assess the quality, relevance, and clarity of model-generated explanations when interpreting Arabic poetry.

Figure 4: **Qualitative Comparison of Model-Generated Explanations for a Single Arabic Poem.** This figure presents a representative Arabic poem alongside its original human-written explanation and corresponding verse-by-verse explanations generated by four different language models. The comparison highlights how each model interprets the poem's rhetorical devices, imagery, and thematic depth relative to the gold explanation. This qualitative analysis illustrates variations in faithfulness, fluency, and literary sensitivity, offering insight into each model's ability to handle nuanced Arabic poetic language and convey its intended meaning.

Overall, closed models such as GPT-4o and Gemini-2.5-Flash achieve consistently strong scores across both automatic and human evaluations. Notably, Gemini-2.5-Flash attains the highest textual entailment score (0.7475), along with high fluency and faithfulness scores, indicating strong alignment with poetic content and natural language clarity. GPT-4o also performs well across all dimensions, with the highest BERTScore and a strong balance of semantic coherence and linguistic quality. Among open models, Deepseek-V3, Aya-Expanse, and Qwen-3 show competitive performance, especially in fluency and textual entailment. However, models like ALLaM-7B and AceGPT-v2 lag significantly in both lexical and semantic overlap, as well as in human-judged fluency and consistency.

A key insight from this evaluation is that most state-of-the-art models perform well on content expressed in Modern Standard Arabic (MSA) but struggle with the classical forms and linguistic intricacies present in historical and poetic Arabic. Despite high scores in generic semantic metrics, many models fail to capture deeper cultural and metaphorical meanings embedded in traditional Arabic poetry. Our analysis highlights the importance of domain-specific evaluation for literary and cultural tasks. It also underscores the need for

building or fine-tuning models that are more sensitive to classical Arabic forms. The gap between fluency and interpretive depth in some models suggests that future research should focus not just on surface-level correctness but also on deeper reasoning and cultural grounding. Such efforts are essential for advancing Arabic NLP in creative and heritage-preserving applications.

Table 4 shows era-wise performance of closed and open-source models on Arabic poem understanding using BERTScore, which captures semantic similarity with human explanations. Closed models like GPT-4o and Gemini variants perform consistently well, especially on modern and recent historical eras. In contrast, open models such as Deepseek-V3 and Aya-Expanse perform reasonably on some eras but struggle with older poetic forms like Pre-Islamic and Umayyad due to their complex language and cultural depth. This highlights that while current models are effective on MSA, they face challenges with classical Arabic. A complementary analysis using Textual Entailment is included in the Appendix (refer Table 18), further supporting these findings.

Additionally, Figure 4 shows a qualitative comparison of model-generated explanations for a classical Arabic poem. It compares outputs from GPT-4o, Gemini 2.5 Flash, Qwen 3, and Aya-Expanse against a human-written explanation. The figure highlights differences in faithfulness, fluency, and interpretive depth, showing how well each model captures the poem's meaning, style, and literary richness. This example clearly illustrates the strengths of advanced models like GPT-4o in understanding nuanced poetic language.

5 Conclusion

Arabic poetry represents one of the richest and most culturally nuanced forms of expression within the Arabic language, characterized by layered meanings, stylistic diversity, and deep historical roots. In this paper, we introduced Fann or Flop, the first benchmark specifically developed to evaluate the capabilities of LLMs in understanding Arabic poetry across 12 historical eras, spanning from pre-Islamic to contemporary periods, and encompassing a broad spectrum of poetic genres and metrical forms. Our benchmark includes carefully curated diagnostic questions aimed at assessing semantic comprehension, metaphorical interpretation, prosodic awareness, and sensitivity to cultural contexts. Through extensive evaluation, we demonstrated that despite strong performances on standard Arabic language tasks, state-of-the-art LLMs consistently struggle with the interpretative and culturally embedded dimensions of Arabic poetic texts. By releasing Fann or Flop as an open-source resource, we aim to encourage further research, promote rigorous assessment methodologies, and support advancements in linguistically and culturally rich Arabic language modeling.

6 Limitations and Societal Impact

While Fann or Flop provides a rigorous framework for evaluating LLMs' understanding of Arabic poetry, it has several limitations. The benchmark covers only a portion of the broader Arabic poetic tradition, as some poems could not be included due to missing metadata, unclear authorship, or lack of reliable era or genre annotations. Additionally, poetry often invites multiple valid interpretations, which current evaluation metrics may not fully capture, even with expert-curated references. Expanding the dataset to include more diverse annotations, as well as dialectal and regional poetic forms, remains a key area for future work.

On the societal front, this benchmark contributes to the preservation and computational accessibility of Arabic literary heritage by positioning poetry as a meaningful testbed for language understanding. By promoting the development of culturally informed and linguistically grounded models, Fann or Flop encourages more inclusive and context-sensitive NLP. Nonetheless, as with any system trained on culturally rich and potentially sensitive material, there is a risk of misinterpretation or misuse. Ensuring transparency, human oversight, and responsible deployment is essential to safeguard the ethical impact of this work, especially in educational, literary, and public-facing applications.

References

Muhammad Abdul-Mageed, Shady Elbassuoni, Jad Doughman, AbdelRahim Elmadany, El Moatez Billah Nagoudi, Yorgo Zoughby, Ahmad Shaher, Iskander Gaba, Ahmed Helal, and Mohammed El-Razzaz. 2021. DiaLex: A benchmark for evaluating multidialectal Arabic word embeddings. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 11–20, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert &

marbert: Deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Sajawel Ahmed, Rob van der Goot, Misbahur Rehman, Carl Kruse, Ömer Özsoy, Alexander Mehler, and Gemma Roig. 2022. Tafsir dataset: A novel multi-task benchmark for named entity recognition and topic modeling in classical Arabic literature. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3753–3768, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Google AI. 2025a. Gemini 2.0 flash. Large language model, accessed May 20, 2025.

Google AI. 2025b. Gemini 2.5 flash. Large language model (Preview), accessed May 20, 2025.

Abu Nasr al Jawhari. 10th Century. Taj al-lugha wa sihah al-arabiya - Wikipedia — en.wikipedia.org. https://en.wikipedia.org/wiki/Abu_Nasr_al-Jawhari. [Accessed 06-05-2025].

alsharekh. 2019. Al-mujam al muaser– lexicon.alsharekh.org. https://lexicon.alsharekh.org/. [Accessed 06-05-2025].

AlSuyuti. 15th Century. Al-mizhar fi eulum allughat wa'anwaeiha - Wikipedia — en.wikipedia.org. https://en.wikipedia.org/wiki/Al-Suyuti. [Accessed 06-05-2025].

Zaid Alyafeai, Maged S Al-Shaibani, and Moataz Ahmed. 2023. Ashaar: Automatic analysis and generation of arabic poetry using deep learning approaches. *arXiv preprint arXiv:2307.06218*.

Toni Andrews. 2024. Is Arabic The Richest Language In Words? - Interpreters & Translators, Inc. — ititranslates.com. https://ititranslates.com/is-arabic-the-richest-language-in-words/. [Accessed 06-05-2025].

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Mujam Ar-Riyadh. 2025. Mujam ar-riyadh— dictionary.ksaa.gov.sa. https://dictionary.ksaa.gov.sa/. [Accessed 06-05-2025].

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, Ahmed Abdelali, Jeril Kuriakose, Abdalghani Abujabal, Nora Al-Twairesh, Areeb Alowisheq, and Haidar Khan. 2024. Allam: Large language models for arabic and english. *Preprint*, arXiv:2407.15390.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. Flute: Figurative language understanding through textual explanations. *arXiv preprint arXiv:2205.12404*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2025. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

FIGLANG202. 2024. FigLang2024 — sites.google.com. https://sites.google.com/view/figlang2024. [Accessed 07-05-2025].

Giuseppe Gallipoli and Luca Cagliero. 2025. It is not a piece of cake for gpt: Explaining textual entailment recognition in the presence of figurative language. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9656–9674.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, et al. 2023. Acegpt, localizing large language models in arabic. *arXiv preprint arXiv:2309.12053*.

Salam Khalifa, Nizar Habash, Fadhl Eryani, Ossama Obeid, Dana Abdulrahim, and Meera Al Kaabi. 2018. A morphologically annotated corpus of emirati Arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. *arXiv preprint arXiv:2204.12632*.

Quentin Malartic, Nilabhra Roy Chowdhury, Ruxandra Cojocaru, Mugariya Farooq, Giulia Campesan, Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Ankit Singh, Maksim Velikanov, Basma El Amel Boussaha, et al. 2024. Falcon2-11b technical report. *arXiv preprint arXiv:2407.14885*.

Ibn Manzur. 14th Century. Lisan al-arab — en.wikipedia.org. https://en.wikipedia.org/wiki/Ibn_Manzur. [Accessed 06-05-2025].

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on padic: A parallel arabic dialect corpus. In *Proceedings of the 29th Pacific Asia conference on language, information and computation*, pages 26–34.

Meta AI. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A Smith. 2012. Recall-oriented learning of named entities in arabic wikipedia. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 162–173.

Hussein Mozannar, Karl El Hajal, Elie Maamary, and Hazem Hajj. 2019. Neural arabic question answering. *arXiv preprint arXiv:1906.05394*.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the twelfth language resources and evaluation conference*, pages 7022–7032.

Susanna Olivero. 2024. *Figurative Language Understanding based on Large Language Models*. Ph.D. thesis, Politecnico di Torino.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

oussama. 2024. Modern Standard Arabic – The Missing Glossary - — blog.jarrousse.org. https://blog.jarrousse.org/2024/03/27/modern-standard-arabic-the-missing-glossary/. [Accessed 06-05-2025].

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. Arabench: Benchmarking dialectal arabic-english machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, Mus'ab Husaini, Soon-Gyo Jung, Ji Kim Lucas, Walid Magdy, Safa Messaoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan Naeem, Mourad Ouzzani, Dorde Popovic, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, Yifan Zhang, Ahmed Ali, Yassine El Kheir, Xiaosong Ma, and Chaoyi Ruan. 2025. Fanar: An arabic-centric multimodal generative ai platform. *Preprint*, arXiv:2501.13944.

Qwen Team. 2025. Qwen3.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

wikipedia. 2025. ar.wikipedia.org. https://ar.wikipedia.org/wiki/%D8%A7%D9%84%D9%84%D8%BA%D8%A9_%D8%A7%D9%84%D8%B9%D8%B1%D8%A8%D9%8A%D8%A9. [Accessed 06-05-2025].

wikipediaArabic. 2025. Varieties of Arabic - Wikipedia — en.wikipedia.org. https://en.wikipedia.org/wiki/Varieties_of_Arabic. [Accessed 06-05-2025].

Taha Zerrouki and Amar Balla. 2017. Tashkeela: Novel corpus of arabic vocalized texts, data for auto-diacritization systems. *Data in brief*, 11:147.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Cheng Zhao, Bin Wang, and Zhen Wang. 2024. Understanding literary texts by llms: A case study of ancient chinese poetry. *arXiv preprint arXiv:2409.00060*.

A Appendix

This appendix provides supplementary material to support our study of Arabic poetry understanding in language models. It includes four key sections: (1) a brief overview of related work in Arabic NLP, highlighting recent progress in benchmark development and the specific gaps our work addresses; (2) detailed dataset statistics, including token distribution, genre coverage, and temporal representation across poetic eras; (3) additional details on the prompts used for model generation and evaluation; and (4) a selection of qualitative examples from the *Fann or Flop* benchmark that illustrate its richness and the interpretive challenges it presents. Together, these components underscore the linguistic, historical, and cultural depth of our dataset and evaluation framework.

B Related Work

Understanding Arabic poetry computationally intersects with multiple subfields of NLP, including language modeling, data set construction, figurative language interpretation, and the evaluation of cultural knowledge. To contextualize our contribution, we review prior work across two key domains: Arabic NLP benchmarks and poetry understanding in LLMs.

B.1 Arabic NLP Benchmarks

Over the past decade, Arabic NLP has advanced considerably with the introduction of large-scale benchmarks such as SOQAL (Arabic-SQuAD and ARCD) (Mozannar et al., 2019), AraBench (Sajjad et al., 2020), and the AraBERT Collection (Antoun et al., 2020). These benchmarks cover essential tasks such as sentiment analysis, named entity recognition (NER), and question answering, and typically support both MSA and dialectal varieties. However, they largely overlook CA, which remains underrepresented in the main resources. Consequently, while models trained on these datasets perform well on surface-level tasks, they lack the depth to assess cultural, rhetorical, and literary understanding, especially in classical poetic contexts.

Additional resources such as the CAMeL corpus (Abdul-Mageed et al., 2020; Khalifa et al., 2018), Tashkeela (Zerrouki and Balla, 2017), PADIC (Meftouh et al., 2015), and MADAR (Bouamor et al., 2018) have enriched the field through morphologically annotated corpora, diacritized texts, and dialectal content. However, these datasets are primarily designed for structural tasks such as morphological disambiguation or dialect identification, without engaging the semantic or figurative dimensions of the poetic language.

More recently, efforts have extended Arabic NLP to the literary and religious domains. The Tafsir dataset (Ahmed et al., 2022) introduces a benchmark derived from *Tafsir al-Tabari*, including NER and topic modeling in CA. AQMAR (Mohit et al., 2012) targets recall-oriented NER in Arabic Wikipedia, offering annotations across standard and domain-specific entity types. Although both datasets engage with classical Arabic and semantic granularity, they do not address poetry or the interpretive challenges it poses.

Among the most directly relevant efforts is Ashaar (Alyafeai et al., 2023), the first large-scale Arabic poetry dataset. It includes tasks such as meter classification, era identification, and poet recognition, along with descriptive metadata. Despite its contributions to computational poetics, Ashaar lacks verse-level annotation, rhetorical device modeling, question-answer style interpretation, and historical contextualization, limiting its ability to evaluate deeper poetic reasoning in language models.

B.2 Poetry Understanding in NLP

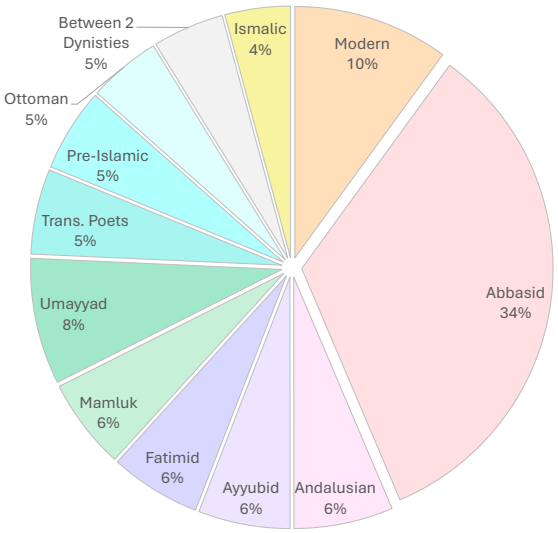
Outside Arabic, poetry and figurative language have emerged as valuable testbeds for assessing the reasoning of LLM (Liu et al., 2022; Bisk et al., 2020; Olivero, 2024). Benchmarks like FLUTE (Chakrabarty et al., 2022) and the FigLang shared tasks (FIGLANG202, 2024) reveal persistent challenges in handling metaphor, simile, and symbolic expression. Recent works (Gallipoli and Cagliero, 2025; Zhao et al., 2024) further expose the limitations of LLMs in interpreting literary texts, including complex poetic structures and non-literal meaning. Despite Arabic's longstanding poetic legacy, this evaluation line remains largely unexplored for Arabic, leaving a notable gap in culturally grounded reasoning tasks.

Fann or Flop addresses this gap by combining a chronological taxonomy of Arabic poetry with interpretive question-answering. It spans 12 eras and integrates dialectal variation, rhetorical analysis, historical context, and verse-level annotation. As summarized in Table 1, no existing benchmark offers this breadth of poetic features, positioning Fann or Flop as the first comprehensive diagnostic tool for evaluating Arabic poetic understanding in LLMs.

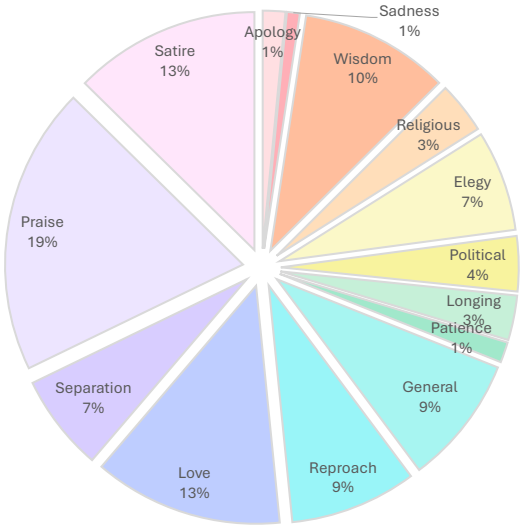
C Fann or Flop Data Statistics

To better characterize the distributional properties of our curated Arabic poetry dataset, we present a series of descriptive statistics that cover both historical and thematic dimensions.

These include the distribution of poems across major eras (Figure 5a), the overall distribution of poetic genres (Figure 5b), and a genre-by-era breakdown (Figure 6).



(a) **Distribution of poems by historical era.** The chart shows the proportion of poems collected from each era. Abbasid, Modern, and Andalusian periods are the most represented, reflecting their central role in Arabic literary production.



(b) **Distribution of poems by genre.** This chart shows the proportion of poetic genres across the dataset. Praise, Satire, and Love dominate the distribution, while genres such as Apology and Sadness appear less frequently.

Figure 5: **Era and Genre Statistics.** Subfigure (a) displays the distribution of poems across historical eras, while subfigure (b) shows the overall genre distribution across the dataset.

Complementing these visualizations, we also include detailed per-era tables listing the most represented poets and the number of poems attributed to each. Together, these statistics contextualize the coverage of the dataset and support downstream applications such as genre classification, diachronic literary analysis, and poet-specific modeling.

Tables 5 to 16 provide a breakdown of the number of poems attributed to prominent Arabic poets across different historical eras. Each table is dedicated to one era:
Table 5: Pre-Islamic era; Table 6 Transitional (Early Islamic) period; Table 7: Modern era; Table 8: Islamic era; Table 9: Umayyad era; Table 10: Abbasid era; Table 11: Between 2 Dynasties; Table 12: Fatimid Dynasty; Table 13: Andalusian era; Table 14: Ayyubid era. Table 15: Mamluk Dynasty; Table 16: Ottoman era.

Era	**Pre-Islamic**
Poet	**Poems**
Aws ibn Hajar	35
al-Samaw'al	12
al-Sulayk ibn al-Sulaka	7
Imru' al-Qais ibn Hujr	34
Zuhayr ibn Abi Sulma	48
Salama ibn Jandal	14
Tarfah ibn al-Abd	26
Urwah ibn al-Ward al-Absi	31
Ubayd ibn al-Abras	40
Amr ibn Qami'a	21
Amr ibn Kulthum	24
Antarah ibn Shaddad	82
Total	**374**

Table 5: Poem counts for major poets from the Pre-Islamic era.

Era	**Transitional Poet**
Poet	**Poems**
Al-Hadira	7
Al-Hutay'a	95
Al-Khansa	92
Hassan ibn Thabit	74
Amir ibn al-Tufayl	41
Amr ibn Barraqa	5
Labid ibn Rabi'a	69
Total	**383**

Table 6: Poem counts for major poets from the Early-Islamic Transitional period.

Era	**Modern**
Poet	**Poems**
Ahmed Shawqi	460
Hafiz Ibrahim	240
Total	**700**

Table 7: Poem counts for major poets from the Modern era.

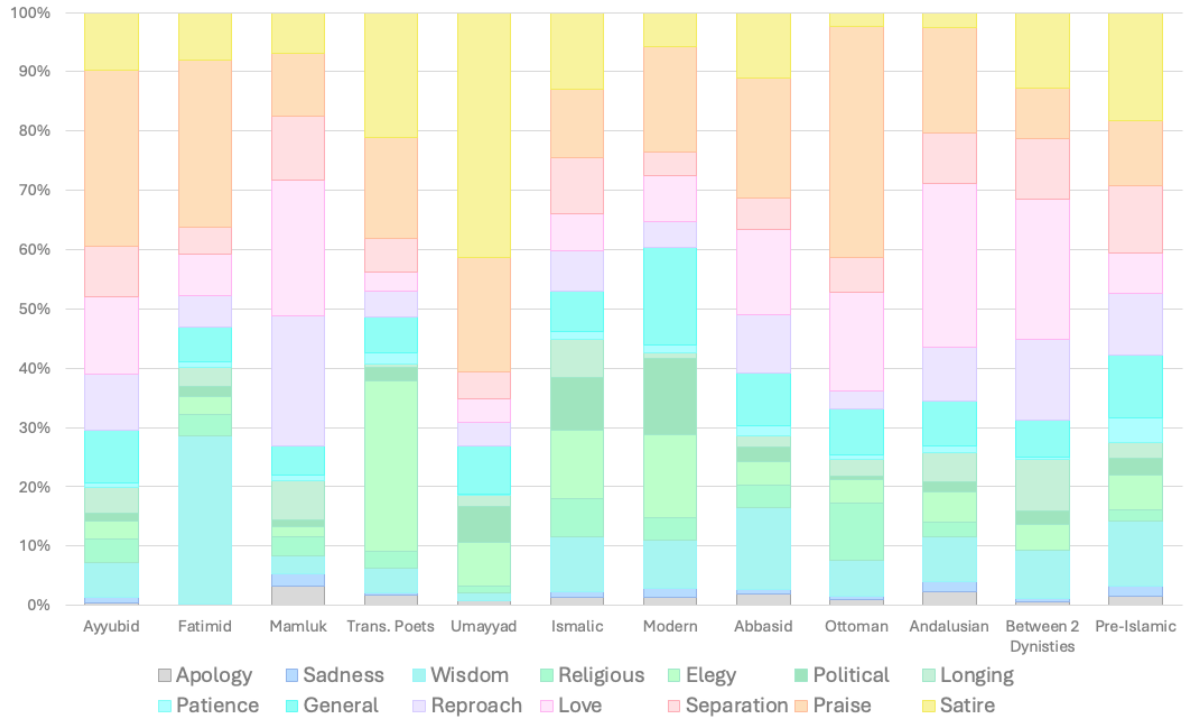


Figure 6: **Genre distribution across historical eras.** This stacked bar chart illustrates how poetic themes evolved across different dynasties. It highlights patterns such as the prominence of Praise and Satire during the Abbasid and Umayyad eras, and the diverse thematic expression in Modern poetry.

Era	**Islamic**
Poet	**Poems**
Abu Muhammad al-Faq'asi	28
Al-Akraa bin Muath Al-Kushairi	16
Asmaa Bin Kharja El-Fazari	4
Aasha Taroud	4
Khuzaima ibn Thabit al-Ansari	14
Khalid ibn al-Walid	7
Az-Zzubayr bin Al-Awam	4
As-Samhari Al-Okliy	11
Al-Ghitamish Al-Dabbi	4
Abd al-Rahman ibn Abi Bakr al-Siddiq	3
Jubaiha' al-Ashja'i	7
Habib ibn Khidrah al-Hilali	6
Ka'b ibn Mashhur al-Makhbali	13
Mas'ud al-Mazini	3
Satira al-Usaybiyya	4
Ziyad ibn Abihi	3
Ziyad ibn Hanzala al-Tamimi	4
Murrah ibn Junada	3
Atika bint Zayd	6
Abd al-Aziz ibn Zararah al-Kalabi	6
Urwa ibn Hizam	6
Ali ibn al-Husayn	6
Amr ibn al-'As	26
Amra bint Mirdas	4
Other Poet with 1 or 2 poems	102
Total	**294**

Table 8: Poem counts for major poets from the Islamic era.

Era	**Umayyad Era**
Poet	**Poems**
Al-Akhtal	136
Jarir	239
Al-Farazdak	178
Ubaydallah ibn al-Ruqayyat	15
Total	**568**

Table 9: Poem counts for major poets from the Umayyad Era.

Era	**Abbasid**
Poet	**Poems**
Abu al-Atahiya	362
Abu Firas al-Hamdani	129
Abu Nuwas	701
Abu Tammam	29
Ibn al-Rumi	227
Imam al-Shafi'i	20
al-Buhturi	601
al-Mutanabbi	273
Total	**2342**

Table 10: Poem counts for major poets from the Abbasid era.

Era	**Between Dynasties**
Poet	**Poems**
Bashar bin Bord	321
Total	**321**

Table 11: Poem counts for major poets from the Between 2 Dynasties.

Era	Fatimid Dynasty
Poet	**Poems**
Abu al-Ala al-Ma'arri	183
Ibn Hayyus	120
Arqala al-Kalbi	106
Total	**409**

Table 12: Poem counts for major poets from the Fatimid Dynasty.

Era	Andalusian
Poet	**Poems**
Abu Ishaq al-Albiri	38
Ibn Khafaja	225
Ibn Zaydun	146
Ibn Sahl al-Andalusi	37
Total	**446**

Table 13: Poem counts for major poets from the Andalusian era.

Era	Ayyubid Dynasty
Poet	**Poems**
Ibn al-Farid	35
Sibt Ibn al-Tawawidhi	291
Muhyiddin Ibn Arabi	87
Total	**413**

Table 14: Poem counts for major poets from the Ayyubid Dynasty.

Era	Mamluk Dynasty
Poet	**Poems**
Baha al-Din Zuhayr	368
Safiyy al-Din al-Hilli	40
Total	**408**

Table 15: Poem counts for major poets from the Mamluk Dynasty.

Era	Ottoman
Poet	**Poems**
Abu al-Ma'ali al-Talawi	75
Ibn Razka	19
Ibn Matuq al-Musawi	74
al-Kawkabani	2
Bint al-Shuhna	2
Abd al-Rahman al-Musili	58
Muhammad al-Isba'i	31
Muhammad al-Sharafi al-Safaqsi	65
Total	**326**

Table 16: Poem counts for major poets from the Ottoman era.

Prompt

```
You are an academic expert in Arabic literature and
poetry analysis. Your task is to provide a deep
linguistic, rhetorical, and literary explanation for
every verse of the Arabic poem provided below,
regardless of its style or period.

The poem to be analyzed is: <POEM_CONTENT>

Please adhere strictly to the following guidelines:
1. Carefully read the entire poem to understand its
   deep meaning and global message, general theme,
   purpose, emotional tone, and cultural or
   historical context, which should inform your
   verse-by-verse analysis.

2. For each verse (or each paired line, if the poem
   follows a two-hemistich structure), write one
   cohesive paragraph in formal Arabic that
   integrates:
   - The literal meaning (المعنى الحرفي):  What the
     poet is directly saying.
   - The figurative, symbolic, or rhetorical meaning
     (المعنى المجازي): Emotional or intellectual
     connotations, artistic imagery, and rhetorical
     devices, such as التشبيه، الاستعارة، الكناية،
     الطباق، الجناس، التورية، etc. You must
     explicitly name these forms and explain their
     function within the verse's meaning and effect.

3. The explanation must be:
   - Unified : Do not separate the literal and
     figurative meanings. Present them in a single
     integrated paragraph.
   - Focused on the verse alone : Do not reference
     other verses explicitly within your
     explanation. However, you may consider the
     overall poem global message and meaning
     implicitly for interpretive accuracy.

4. Avoid generic commentary (e.g., "the poet
   expresses love or sadness"). Instead, your
   analysis must refer directly to the language and
   structure of the verse and explain how the
   meaning is conveyed.

5. If there are multiple possible interpretations ,
   choose the clearest and most textually supported
   one , without mentioning alternative readings.

6. Each paragraph should be between 2 to 4 well-
   formed sentences , reflecting literary insight and
   academic precision.

7. Ensure that every verse of the poem is
   explained. Do not skip or summarize verses.

8. Generate a global explanation of the whole poem
   piece to support your verse explanation and refer
   to it to ensure consistency and connectivity and
   coherence from start to end. The explanation
   should not exceed a few lines.

9. Return your output in the following strict and
   valid JSON format only:
   • No extra text, no explanations, no markdown:
     ```json
       {
         "explanation": "<full poem summary>",

         [{
         "verse": "<full text of verse 1>",
         "explanation": "<full Arabic explanation
                        for verse 1>"
         },

        {
         "verse": "<full text of verse 2>",
         "explanation": "<full Arabic explanation
                        for verse 2>" },

       //... continue for all remaining verses ...
         ]
       }
```

Figure 7: **The verse-level explanation prompt used for evaluation.** This prompt instructs the model to produce detailed verse-by-verse explanations in Arabic. It guides the model to integrate both literal and figurative interpretations, explicitly name rhetorical devices (e.g., metaphor, personification, paronomasia). The prompt enforces coherence, academic rigor, and structural consistency by requiring output in a strict JSON format.
```

D Annotator Rubric Guidelines

Table 17 presents the scoring rubric used by annotators. Each annotator assigns a single value for every criterion, and the criterion scores are summed to produce an individual interpretive depth score. The final reported score is the average of these values across annotators.

Criterion	Score	Guideline
Literal Comprehens.	0	Misinterprets or ignores the poem's literal meaning; adds unrelated content.
	1	Correctly reflects the surface meaning of the poem.
Thematic & Emotional Depth	0	Fails to capture themes, sentiment, or tone.
	1	Identifies a theme or emotion only superficially.
	2	Clearly conveys central themes and emotions with nuance.
Cultural & Historical Appropriateness	0	Culturally/historically inaccurate; ignores context.
	1	Shows partial or limited awareness of context.
	2	Demonstrates accurate and insightful cultural/historical awareness.
Stylistic Sensitivity	0	No recognition of rhetorical/literary features.
	1	Mentions some features, but inaccurately or superficially.
	2	Correctly identifies key features, but lacks depth.
	3	Nuanced discussion of literary devices (e.g., metaphor, rhythm, imagery).
Expressiveness & Coherence	0	Incoherent, disorganized, or ungrammatical.
	1	Understandable but lacks fluency or cohesion.
	2	Clear, well-structured, and stylistically appropriate in Arabic.

Table 17: Annotator evaluation rubric of model-generated explanations. Each criterion is scored independently, with higher values indicating better performance.

E Prompts Used

E.1 Model Generation Prompt

To generate verse-level explanations suitable for evaluating both open- and closed-source models, we developed a carefully optimized generation prompt. The prompt design followed an iterative and augmented process. Initially, we used a simple bilingual (in Arabic and English) prompt asking for explanations. Based on early outputs, which tended to capture local semantic meaning but lacked coherence and global context, we progressively refined the prompt to elicit more structured and connected responses.

Through multiple rounds of testing, expert evaluation, and prompt engineering, we incorporated explicit instructions to address both local (verse-specific) and global (poem-wide) interpretive elements as support. This enhancement significantly improved the quality of the generated explanations, resulting in outputs that were more coherent, context-aware, and semantically aligned with the original verses.

After extensive comparison, expert reviewers favored the English version of the prompt over its Arabic counterpart, as it more consistently achieved local-global alignment and produced well-connected, high-quality explanations. This final version of the English prompt (Figure 7) was adopted for all subsequent evaluations.

E.2 Model Evaluation Prompt

To ensure consistent and reliable automatic LLM-Judge evaluation of model-generated poem explanations, we designed a clear and structured system prompt (see Figure 8). The prompt positions the evaluator as an expert Arabic linguist and literary critic, responsible for assessing AI-generated verse-by-verse explanations against ground-truth references.

Each poem is evaluated on two key dimensions: Faithfulness/Consistency, which measures how accurately the explanation reflects the verse's intended meaning, and Fluency/Grammaticality, which assesses the quality of the generated text in Modern Standard Arabic. Annotators assign a score from 1 to 5 for each criterion based on the overall performance across all verses, without providing per-verse feedback or open-ended commentary.

The prompt ensures simplicity, objectivity, and high inter-annotator agreement, making it well-suited for evaluating poetic reasoning in culturally rich and linguistically nuanced contexts like Arabic poetry.

F Additional Examples: Qualitative, Translated, and Quantitative Insights

In the following section, we present a more detailed evaluation of the Textual-Entailment (Refer

System Prompt used to evaluate the poem explanation for Faithfulness and Fluency metric

```
You are an expert Arabic linguist and literary evaluator.

Your task is to **evaluate a full Arabic poem's verse-by-verse explanations**.
You will compare **ground-truth** (human-written) explanations with
**generated** explanations from an AI model.

You will judge each verse explanation based on the following two criteria:

---

### Evaluation Criteria (per verse)

1. **Faithfulness / Consistency**:
   Is the generated explanation consistent with the meaning of the verse?
   - Score 5: Deeply faithful to the verse's content
   - Score 3: General alignment but loses poetic imagery
   - Score 1: Misinterprets or invents meaning

2. **Fluency / Grammaticality**:
   Is the generated explanation well-formed Modern Standard Arabic?
   - Score 5: Fluent, grammatically correct
   - Score 3: Understandable with minor issues
   - Score 1: Awkward, incomplete, or ungrammatical

---

### What You Will Receive

You will receive for each poem:
- `poem_title`
- "ground_truth": a list of objects { "v": <int>, "text": <string> }
- "generated": a list of objects with the **same v indices**

---

### What You Must Do

- Compare all verses together and assign a single score of 1-5 for each
    criterion.
- Do **not** provide per-verse scores or any comments.


Then:
- Calculate average scores for the whole poem
- Provide an `overall_score` (1-5) that reflects your judgment across all
    verses

---

Do NOT provide any comments or rationale.
Respond with valid JSON **only** in this format:

### Output Format (in JSON)

{
  "faithfulness_score": <1-5>,
  "fluency_score": <1-5>,
  "overall_score": <1-5>
}
```

Figure 8: System prompt used for LLM-Judge evaluation of verse-by-verse poem explanations. LLM (OpenAI, 2024) compare AI-generated outputs with original explanations and assign overall scores for faithfulness to meaning and fluency in MSA, following clearly defined criteria. The structured format ensures consistency and reliability across evaluations.

	Model	Pre-Islamic	Transitional	Early Islamic	Umayyad	Abbasid	Fatimid
Closed	GPT-4o-2024-08-06 (OpenAI, 2024)	0.6425	0.6502	0.7116	0.6166	0.6699	0.7050
	GPT-4o-mini-2024-07-18 (OpenAI, 2024)	0.4355	0.4789	0.5436	0.4200	0.4266	0.4532
	Gemini-2.5-Flash (AI, 2025b)	0.7275	0.7308	0.7527	0.7112	0.7417	0.7542
	Gemini-2.0-Flash (AI, 2025a)	0.6908	0.7156	0.7458	0.6798	0.7033	0.7462
	Gemini-1.5-Pro (Reid et al., 2024)	0.6004	0.6372	0.6497	0.6312	0.6035	0.6502
	Fanar-Star (Team et al., 2025)	0.6142	0.6354	0.6621	0.5900	0.6413	0.6717
Open	Deepseek-V3 (Liu et al., 2024)	0.5066	0.5875	0.6174	0.5482	0.4736	0.5581
	Deepseek-R1 (Guo et al., 2025)	0.5066	0.5875	0.6174	0.5482	0.4736	0.5581
	Llama-3.3-70B (Meta AI, 2024)	0.5456	0.5469	0.5747	0.5211	0.5341	0.5387
	Qwen-3 (Team, 2025)	0.6142	0.6354	0.6621	0.5900	0.6413	0.6717
	Aya-Expanse (Dang et al., 2024)	0.6142	0.6354	0.6621	0.5900	0.6413	0.6717
	ALLaM-7B (Bari et al., 2024)	0.5619	0.5630	0.6037	0.5844	0.5848	0.6158
	Jais (Sengupta et al., 2023)	0.6124	0.6289	0.6482	0.6047	0.6295	0.6421
	AceGPT-v2-70B-Chat (Huang et al., 2023)	0.5851	0.5656	0.6104	0.5770	0.6119	0.6095

	Model	Andalusian	Ayyubid	Mamluk	Between Dynasties	Ottoman	Modern
Closed	GPT-4o-2024-08-06 (OpenAI, 2024)	0.7128	0.6774	0.7393	0.6656	0.7379	0.6843
	GPT-4o-mini-2024-07-18 (OpenAI, 2024)	0.4869	0.4303	0.4507	0.4240	0.4836	0.3988
	Gemini-2.5-Flash (AI, 2025b)	0.7778	0.7416	0.7866	0.7398	0.7994	0.7544
	Gemini-2.0-Flash (AI, 2025a)	0.7527	0.7320	0.7698	0.7164	0.7585	0.6951
	Gemini-1.5-Pro (Reid et al., 2024)	0.6710	0.6074	0.6377	0.5971	0.6441	0.5965
	Fanar-Star (Team et al., 2025)	0.6749	0.6454	0.7105	0.6342	0.7151	0.6429
Open	Deepseek-V3 (Liu et al., 2024)	0.5927	0.5065	0.5448	0.4929	0.5226	0.4705
	Deepseek-R1 (Guo et al., 2025)	0.5927	0.5065	0.5448	0.4929	0.5226	0.4705
	Llama-3.3-70B (Meta AI, 2024)	0.5873	0.5221	0.5849	0.5129	0.5712	0.4897
	Qwen-3 (Team, 2025)	0.6749	0.6454	0.7105	0.6342	0.7151	0.6429
	Aya-Expanse (Dang et al., 2024)	0.6749	0.6454	0.7105	0.6342	0.7151	0.6429
	ALLaM-7B (Bari et al., 2024)	0.5892	0.6044	0.6736	0.5905	0.6556	0.6302
	Jais (Sengupta et al., 2023)	0.6540	0.6399	0.6812	0.6183	0.6625	0.6348
	AceGPT-v2-70B-Chat (Huang et al., 2023)	0.6215	0.6131	0.6683	0.5681	0.6273	0.6044

Table 18: **Era-wise Evaluation using Textual Entailment (TE).** Era-wise performance of closed and open-source models on the Arabic poem understanding task, measured using the Textual Entailment metric. This metric evaluates how well the model-generated explanation logically aligns with the original poem content. The results are grouped across key historical eras, from Pre-Islamic to Modern, allowing a fine-grained view of model strengths and limitations across time periods. Closed models such as GPT-4o and Gemini variants demonstrate consistently high entailment across most eras, while select open models like Deepseek-V3 and Aya-Expanse show promising results in specific historical contexts. This analysis highlights the importance of temporal generalization and cultural grounding in building robust Arabic literary reasoning models.

Table 18) metric across the 12 historical eras, comparing both open-source and closed-source models on this dimension. To support comprehensive engagement and a clearer understanding of the data evaluated, we also include selected English translations of Arabic poetic samples, as well as additional qualitative Arabic examples. These additions offer deeper insight into the linguistic diversity, thematic range, and overall quality of the dataset used in our analysis.

F.1 English Translated Qualitative Samples

To support accessibility and improve cross-linguistic understanding, we translated selected Arabic poetry samples shown in Figure 2 into English. For this task, we used GPT-4o, which provided deeper context-aware translations that more accurately capture the figurative and rhetorical nuances of the original verse, outperforming basic literal tools such as Google Translate. These translations allow non-Arabic speakers to more easily engage with the literary richness, emotional depth, and stylistic variety explored in our research.

F.2 Additional Qualitative Samples

To further showcase the dataset's richness, we present additional qualitative samples spanning diverse historical periods, poetic genres (e.g., satire, elegy, political verse), and metrical patterns (See Figure 10). These examples were selected to demonstrate the stylistic, thematic, and rhetorical variety encountered in our evaluation.

Figure 9: **Translated Samples.** This figure presents English translations of the Arabic samples shown in Figure 2. The translations are included to facilitate understanding and accessibility for non-Arabic speakers, allowing broader engagement with the poetic content without requiring prior knowledge of Arabic.

Figure 10: **Fann or Flop Samples by Genre.** Additional representative examples from the Fann or Flop benchmark, illustrating the diversity of genres covered, including Love (Ghazal), Praise (Madḥ), Wisdom (Hikma), Satire (Hijā'), Elegy (Rithā'), Reproach ('Itāb), Political Poetry, and Longing (Shawq). Each example showcases a poetic excerpt alongside an interpretive breakdown highlighting figurative language, rhetorical devices, and thematic nuances. These curated samples reflect the benchmark's aim to evaluate models' nuanced understanding of Arabic poetic tradition.