International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I5P122

Volume 74 Issue 5, 329-341, May 2026

Original Article

Evaluating the Reliability of Large Language Models in Literary Analysis of Arabic Novels: A Structured Benchmark Using Grounded Textual Evidence

Emad A. Aldomour[1], Ameen Shaheen[2]

[1]*Department of Arabic Language and Literature, Al-Balqa Applied University, Princess Alia University College, Amman, Jordan.*
[2]*Department of Software Engineering, Al Zaytoonah University of Jordan, Amman, Jordan.*

[1]*Corresponding Author: dr.emaddmor@bau.edu.jo*

Received: 17 January 2026 Revised: 19 March 2026 Accepted: 28 March 2026 Published: 30 May 2026

Abstract *- This study examines the reliability of a contemporary Large Language Model (LLM) in performing strictly text-bound literary analysis of the Arabic novel "الرصاصة الصديقة". Despite growing use of LLMs in the humanities, their interpretive behavior in Arabic narrative contexts remains underexplored. Using the novel as a controlled corpus, the model was barred from external knowledge and evaluated under a Grounded-Evidence Protocol requiring all claims to be supported by explicit textual quotations. Output quality was assessed through a five-dimensional rubric measuring Textual Fidelity, Accuracy, Analytical Depth, Coherence, and Linguistic Quality. Quantitative results show strong performance in Coherence (4.50), Linguistic Quality (4.58), and Textual Fidelity (4.30), while Analytical Depth was moderate (3.95), indicating limitations in symbolic reasoning and culturally embedded interpretation. Qualitative error analysis reveals that unsupported or inflated symbolic readings typically emerge in ambiguous or metaphor-dense passages. The findings suggest that LLMs can provide coherent, well-grounded commentary but remain constrained in higher-order interpretation. The study proposes a structured benchmark for evaluating LLM reliability in narrative analysis and offers a methodological foundation for future work in computational humanities. All analyses in this study were generated using the ChatGPT-5.2 Large Language Model operating under a fully text-restricted environment.*

Keywords *- AI Evaluation, Arabic Linguistics, Computational Literary Studies, Digital Humanities, NLP.*

1. Introduction

Large Language Models (LLMs) have rapidly evolved into powerful tools capable of performing complex reasoning, text generation, and domain-specific analysis [1]. Their impressive performance across diverse benchmarks ranging from instructional alignment to professional examinations has positioned them as promising instruments for computational humanities and automated literary analysis [2].

However, the reliability, methodological, and textual basis of the literary criticism generated by LLMs remain unexplored as they gain popularity [3]. In particular, it is unclear how applicable these methods are to Arabic narrative fiction, a cultural and linguistic field that is both very rich and extremely complex [4].

In addition to the challenges of language and style presented in Arabic novels, there are several features that contribute to the complexity of automatic analysis using LLMs [5]. The use of metaphors, as well as the use of multiple intertextual references, layered narrative voice, and diglossia, each adds layers of complexity to the analysis of an Arabic novel, which increases the possibility of error in LLMs through either hallucination, making unsubstantiated claims, or reading too much into what is being said by the author [6].

Thus, assessing the reliability of LLM outputs in Arabic literary criticism will require methods that constrain LLM output to be strictly grounded in texts, avoid external knowledge contaminations, and provide a transparent way of evaluating how faithful and true the output is [7].

Literary criticism is different from many of the other NLP applications; as such, there is a need for critical interpretive reasoning as well as for an understanding of the narrative structure of texts and for fidelity to textual evidence rather than just providing fluently written but superficially analyzed factual data or mere superficial patterns of semantic data [8]. Recent studies have been conducted to evaluate the LLMs in summarizing, question-answering, and other NLP tasks that are applied to Arabic texts; however, most existing studies are focused on assessing the general language skills or downstream tasks instead of performing an interpretive literary analysis [9]. Most benchmarks for evaluating Arabic LLMs, including LAraBench [10], AlGhafa [11], GATmath/GATLc [12], primarily assess the ability of classification, reasoning, and multiple choice understanding within many different areas, and do not specifically assess the quality of narrative interpretation, argumentative coherence, and evidence-based criticism in longer form writing.

Although several studies have examined Arabic LLM evaluation benchmarks and hallucination-related reliability issues, a structured benchmark for evaluating LLM-generated Arabic literary criticism under strict quotation-based grounding criteria does not yet exist. Current evaluation suites are primarily limited to short-form NLP tasks and therefore cannot measure long-form interpretive reasoning, evidence-based argumentation, or narrative-level coherence in Arabic novel analysis. The absence of such a benchmark limits the ability of researchers to determine when an LLM-generated literary interpretation is reliable, versus when it becomes unsupported and interpretively unfaithful.

Prior work on LLM faithfulness and hallucination has consistently shown that models often produce highly fluent explanations while exhibiting weaknesses in attribution and evidence alignment. Similarly, early studies on LLM-based literary analysis in English report strong rhetorical coherence but reduced reliability in metaphorical interpretation and symbolic inference. The present study extends these observations into Arabic narrative fiction and evaluates whether grounding protocols can mitigate such interpretive drift in a linguistically and culturally complex literary domain.

In order to fill this gap, the current research introduces a structured framework that will be used as a standard by which to assess the reliability of LLMs when they are being used for literary analysis through evidence from the actual texts. The framework is applied to an entire Arabic-length novel "**الرصاصة الصديقة**", and assesses many areas of LLM output, including:

- Textual Fidelity (Grounding) - Is the AI based solely on the textual material of the book?
- Accuracy - Were the claims/arguments supported by verifiable data/narrative elements from the book?
- Analytic Quality - Was there a meaningful, analytical insight that went past just describing what was happening?
- Coherence - Was the analysis logical and did it make an overall argument?
- Academic Quality of Language - Did the language used maintain the academic tone?

To ensure methodological rigor, a purely text-only operational protocol has been developed to be followed during this study that prohibits the use of any prior knowledge, or external data regarding the author, context, or critical reception as it pertains to the content being analyzed [13]. All analytical responses produced through the LLM will be scored via a five-dimensional reliability rubric based upon recent research in LLM grounding, hallucination evaluations, and fidelity metrics [14]. The quantitative results of these analyses will be aggregated to provide an initial reproducible LLM evaluation framework for Arabic literary analysis.

This study has four key contributions. First, it presents the first structured benchmark to assess the reliability of Large Language Models (LLMs) in the context of Arabic literary criticism. The existing benchmarks have primarily focused on non-interpretive tasks [10-12]. Second, this study builds on previous work regarding hallucinations and grounding through its development of a grounded-evidence protocol, which ensures reduced hallucinations and improved textual fidelity [15]. Third, it provides a text-based, experimentally evaluated assessment of the novel "الرصاصة الصديقة" and represents an easily replicable method for analyzing Arabic narratives. Fourth, the study provides both quantitative and qualitative information about how well LLMs perform and what errors they make while interpreting texts, as well as their limitations, and provides a basis for comparison with other models and establishes these findings in the broader literature on LLM faithfulness and trustworthy deployment [16]. These contributions together create a clear, reproducible way to measure the reliability of LLMs when used for narrative interpretation and will provide a base for future studies in computational literary studies, Arabic NLP [17], and model evaluation. In order to provide methodological transparency and replicability for the data analyzed in this study, all of the analytical outputs evaluated in this study were created using the ChatGPT-5.2 model, under the same constrained, text-only conditions. Accordingly, the related work section reviews recent Arabic LLM benchmarks, grounding and hallucination evaluation research, and existing studies on LLM-based literary analysis.

1.1. Research Problem and Originality

The central research problem addressed in this study is whether large language models can reliably generate Arabic literary analysis that remains strictly grounded in the primary text, without introducing unsupported interpretations or culturally inflated symbolism. While Arabic LLM benchmarks exist, they primarily evaluate short-form NLP performance and do not measure long-form interpretive reasoning, narrative coherence, or quotation-based evidence alignment in Arabic novels.

The originality of this work lies in proposing a structured benchmark for Arabic narrative criticism under strict grounding constraints. The study introduces a Grounded-Evidence Protocol requiring explicit quotation support, a multi-dimensional reliability rubric tailored to literary analysis, and an empirical evaluation applied to a full-length Arabic novel. This combination provides a reproducible foundation for evaluating interpretive reliability in Arabic computational humanities.

2. Related Works

Researchers have recently concentrated on reliability, textual grounding, and controlling hallucinations with LLMs [18]. The recent development of benchmarking systems, including HaluEval [19], TruthfulQA [20], and Walk-the-Talk Faithfulness Metrics [21], shows that even the most current versions of LLMs produce supported or unfaithful, but certainly plausible-sounding responses to questions, a problem that is most prevalent in open-ended response generation scenarios. Grounded protocols have been emphasized by researchers, which constrain models from referencing non-verifiable textual evidence, demonstrating significant improvement in both factual accuracy and interpretive coherence when models are constrained within a limited context [22].

As a counterpart, there has been an expanding number of attempts at evaluating the LLM output by using a structural approach to rating. The results from rubric-based human evaluations, preference model ratings, and faithfulness (attribution) rating methods indicate that multiple-dimensional evaluation is required to assess the quality of generated text, especially in the case of tasks requiring analysis, interpretation, and explanations [23]. However, the majority of the current literature remains focused on areas of summarization, General QA, and instructional alignment rather than the area of literary interpretation, which requires both analytical depth and narrative fidelity.

Over recent years, there has been an emergence of benchmark suites for Arabic NLP which measure how well Arabic-focused LLMs perform on various natural language processing tasks [24], including but not limited to: text classification; reading comprehension; mathematical reasoning, and general linguistic ability. These benchmarks, LAraBench [10], AlGhafa [11], GATmath/GATLc [12], were very important for both determining where Arabic LLMs are currently in terms of their capabilities and where they are likely to be in terms of limitations. Additional recent efforts have also proposed broader Arabic evaluation resources that reflect expanding interest in systematic benchmarking of Arabic language models beyond standard NLP tasks [25]. Although these assessment suites offer a broad range of evaluations, none of them investigated the areas of literary critique or narrative analysis. All of the previous evaluation suites were designed to evaluate the performance of Arabic LLMs on short-form task-based assessments, and not long-form analytical evaluations based upon full narrative texts. While numerous studies into the use of AI within the Humanities explore plot detection, characterization networks, topics, and style as possible applications for digital humanities [26], the majority of this work relies upon traditional NLP techniques or descriptive computational approaches, which do not allow for an interpretative, argument-driven critical approach to analyzing literature. In terms of LLMs, LLM-based literary analysis in English has produced mixed results, where LLMs tend to exhibit rhetorical proficiency. Recent literary-domain evaluation datasets have also been introduced to test model performance on fiction understanding, although these efforts remain largely English-centered and do not enforce quotation-grounded interpretive criticism [27]. However, they have difficulty establishing textual fidelity and poorly handle metaphors, symbols, and narrative voice. Given that Arabic is a language that exhibits diglossia, is characterized by a density of metaphorical structures, and is culturally embedded, it can be reasonably anticipated that similar limitations exist for LLM-based analysis of Arabic texts [28].

The current literature indicates multiple overlapping gaps: there is no agreed standard for measuring LLM reliability in Arabic literary criticism, there is lack of protocols that are grounded in the narrative for interpreting narratives, the unavailability of rubrics for long form Arabic analysis, and few studies have engaged with some of the special challenges of analyzing Arabic prose such as diglossia, layered narration, culturally embedded metaphors [29]. The previous studies outline a distinct necessity for a reliable, reproducible, and text-grounded benchmark that will allow researchers to evaluate LLM performance in terms of their ability to reason interpretively rather than just in terms of how well they process the surface characteristics of language. As such, this study provides a structured assessment of the reliability of LLMs as they apply analytical reasoning to an Arabic novel.

Novelty Positioning Summary. Existing Arabic benchmark suites primarily evaluate Arabic LLM capability through short-form, task-based testing, while reliability benchmarks such as hallucination and faithfulness evaluations focus on attribution and truthfulness without targeting long-form literary criticism. In parallel, literary-domain evaluation efforts have largely concentrated on English fiction understanding rather than evidence-grounded interpretive analysis of Arabic narratives. The present study is distinct in combining (i) full-length Arabic novel evaluation, (ii) a strict quotation-based Grounded-Evidence Protocol for every interpretive claim, and (iii) a multi-dimensional reliability rubric tailored specifically to the requirements of literary criticism. This combination provides a structured and reproducible benchmark for assessing interpretive reliability in Arabic computational humanities.

3. Comparative Benchmarking

Existing LLM evaluation benchmarks provide important insights into model performance across multiple languages and tasks. However, most widely used benchmarks evaluate short-form tasks such as question answering, classification, reasoning, or multiple-choice understanding. While these tasks are valuable for measuring general language competence, they do not directly capture the requirements of literary criticism, which demands interpretive reasoning, narrative coherence, and traceable evidence linking.

Arabic LLM benchmarks such as LAraBench [10], AlGhafa [11], and GATmath/GATLc [12] have a primary focus on evaluating the performance of models on structured Natural Language Processing (NLP) tasks that include: reading comprehension, instruction following, mathematical reasoning, and classification. These benchmarks will provide insights into the Arabic language capability in controlled environments; they do not assess longer form narrative interpretation, evidence-based argumentation, or quotation-grounded literary analysis. On the other hand, the benchmark proposed in this research has a primary focus on interpretative literary reasoning and will enforce strict grounding requirements to ensure direct quotations are used from the source novel.

Reliable benchmark suites that focus on reliability, such as TruthfulQA [20] and HaluEval [19], evaluate truthfulness and hallucinations in open-ended text generation. These frameworks are directly related to this research project, as literary critics can also be misled by plausible claims that lack support. However, these two benchmarking suites were generally created for large-scale open-ended narrative corpora, and they are not designed to assess either the interpretative depth, the narrative coherence, or the ability of quotation-based evidence to provide a basis for literary interpretation. Therefore, the proposed Grounded-Evidence protocol represents an adaptation of faithful evaluation to a domain of narratives with a focus on providing a means of evaluating faithfulness in a way that the evidence is traceable back to the text itself.

Research on the use of LLMs for literary analysis in English reported that while the models demonstrated strong rhetorical fluency and coherence in their analyses, they did not perform well when it came to deeper levels of symbolic inference and metaphor interpretation or thematic synthesis. The current research supports those general findings as the model was able to generate high-quality language and coherent writing, but performed at a lower level of analytical depth. The current study further expands upon previous research by examining the use of LLMs for literary analysis within Arabic narrative fiction and its unique characteristics of cultural and linguistic complexity.

The challenges of evaluating Arabic novels are significantly less than those found in many English literary analysis settings, including diglossia, metaphoric language density, culturally-specific symbolic references, and layered narrative voice. Because of this, while LLMs' output may be linguistically fluent, it may produce drift from interpretation, inflated symbolic interpretations, and culturally detached readings. Therefore, an Arabic literature analysis benchmark must assess both the quality of the output's language and its grounding fidelity and interpretive reliability when constrained to a single, strict text.

Overall, the proposed benchmark is unique compared to other Arabic NLP benchmarks because it evaluates extended interpretative reasoning versus task performance. The proposed benchmark also diverges from typical evaluations of hallucinations in that they use faithfulness assessments when interpreting narratives, while grounded by quotation-based references. Lastly, the proposed benchmark builds upon existing literary domain findings in the realm of English literature; specifically, extending them into the realm of Arabic fiction, providing a formalized and reproducible evaluation structure for computational literary study in Arabic.

4. Ethics and Socio-Cultural Issues

The use of large language models in analyzing Arabic literature has several serious cultural, social, and ethical issues. While large language models are able to produce highly intelligible academic-style responses, literary studies require an understanding of the cultural context of the work being studied, as well as sensitivity to history and other forms of context that cannot be replicated by language models alone. Therefore, even if an output is grammatically correct or coherent, it can still potentially distort interpretation with negative ethical implications.

Arabic novels use a lot of figurative language (metaphors), encoded cultural references, and culturally dependent symbols. Thus, an LLM can create interpretations of a novel, based on these culturally embedded references, which go beyond what is actually written, resulting in a reading that is symbolically compelling but culturally irrelevant. The potential for this to occur is especially high in Arabic literary criticism, as much of the meaning is rooted in shared cultural knowledge, religious allusions, or socio-political nuances.

When evaluating LLMs' biases as they relate to the interpretation of literature, we need to be aware of how these biases are presented in the way an LLM interprets literature and how they reflect a bias in the training data. For example, when analyzing Arabic narrative LLMs, they may draw broad generalizations about authors' intentions, themes, or character development as a result of the biases in their training data.

The growing application of LLMs in the study and teaching of humanities has raised a concern that scholars and students will rely too heavily on AI-generated interpretation without critically evaluating these interpretations. Because LLM outputs are often very good at sounding like intelligent people and have the structure of a well-written paragraph, there is an increased risk of treating them as authoritative when the interpretive claims themselves are not strongly supported. This could lead to superficial analysis, less critical engagement with ideas, or unsupported interpretations being introduced into the larger academic discourse.

The Grounded Evidence Protocol developed in this research can be viewed as a means to mitigate some potential risks that exist with the use of quotation-based evidence for interpretative statements and improve transparency. Grounding is not enough to eradicate cultural misinterpretation or bias; therefore, expert human reviewers will still be needed to review the outputs, especially if they relate to symbolic, ideologically charged, or culturally sensitive aspects of Arabic Literature.

5. Explainability & Human-in-the-Loop

Explainability in LLM generated literary reviews does not necessarily mean accuracy in Its interpretation of the literature Some generated literary reviews can still have unsubstantiated and exaggerated interpretations. This research uses explainability by using the Grounded-Evidence Protocol (GEP), which has the requirement that each interpretative statement be connected to a specific quote from the novel. The GEP provides for traceability in the interpretative reasoning process allowing the reviewer to check whether an interpretative claim is based on evidence from the text.

In addition, to evaluate models, a human-in-the-loop evaluation process is utilized via rubric-based annotation, where human evaluators assess each model response across

five dimensions and assess reliability as well as beyond surface-level fluency. The combination of Quotation-based grounding with the use of expert scorers will provide a transparent and secure evaluation framework, similar to structured monitoring layers proposed in intelligent service architectures [30].

6. Reproducibility and Open Science

The reproducibility of reliable LLM evaluations in the humanities has been shown to be improved through the use of a standardized prompting methodology, as well as an evaluation rubric that remains fixed throughout the project; in addition, this project employed a GEP that was employed consistently with respect to each task in the project, and the structured approach to annotating responses ensures that scorers are assessing based on the same dimensions.

Due to copyright restrictions, the full text of the novel cannot be publicly redistributed. However, the evaluation framework, rubric definitions, prompting templates, and aggregated scoring results can be shared to support replication on other Arabic novels. This enables future researchers to reproduce the methodology, compare models, and extend the benchmark to additional literary domains.

7. Methodology

This research utilizes a methodological framework that is capable of providing a text-based evaluation of the ability of a LLM to perform a literary analysis of an Arabic narrative. The use of the methodology provides for the LLM's informational environment to be strictly controlled and for a grounded evidence process to be established, as well as the output quality of the LLM to be evaluated using a multi-dimensional scoring rubric to evaluate the quality of the output.

7.1. Corpus Description and Preparation

The Arabic novel "الرصاصة الصديقة" , is the only source used for this analysis. It was completely digitized and broken down by chapters (as individual textual units). This will allow the model to pull up correct and related sections of text without exposing it to additional information that would be outside the model's context. No biographic, historic, or critical metadata were attached to the text. Isolating the text from other interpretative traditions gives the researcher an opportunity to measure the model's inherent capability to understand the literary content of the text on its own.

7.2. Constrained Computational Environment

In order for all interpretations of the data to be generated by the primary source alone, the model was put into a totally restricted use state. All access to the internet, previously trained models of literary criticism, and any external knowledge base were blocked. Therefore, it cannot rely on prior knowledge of culture, pre-existing knowledge bases of texts, or learned literary shortcuts to generate answers. Therefore, the mode of operation has transformed the evaluation into a test of reasoning based solely upon the data in the text. The model used for all interpretive tasks was ChatGPT-5.2, which was executed in a fully isolated environment to prevent the use of any external knowledge during the analytical process. ChatGPT-5.2 was used as the evaluated model. The same model configuration and constraints were applied consistently across all tasks to ensure comparability.

7.3. Prompting Framework and Interpretive Task Design

The authors also used a standardized prompting structure for all tasks to increase comparability among the different analyses. With each task prompt, the model is directed to apply an explicit type of literary analysis, and it is specifically required to reference textual support for each of its claims, using a quote no longer than about twenty words from the novel. If there is insufficient textual evidence to support a claim, the model will identify that it has no textual basis for supporting the claim. These practices reflect the current trend toward the increased use of obligations by researchers evaluating LLMs to achieve greater reliability and fewer falsepositive results. All prompts used in the study were directed at the same ChatGPT-5.2 model and, therefore uniformly applied across all tasks to allow for an exact comparison of how the model interpreted the prompts. The purpose of using the same model was to make it possible to compare how the model interpreted the various prompts.

7.4. Grounded-Evidence Protocol (GEP)

To ensure the institutionalization of the fidelity of the text, this research utilized a Grounded Evidence Protocol with three interconnected layers. The first layer limits the source of the data to chapter segmental text from "**الرصاصة الصديقة**", which eliminates the potential for epistemological leak from outside sources. The second layer obligates the model to identify the evidence that supports each of the assertions made by the model in terms of its interpretation, providing a transparent and traceable link between the claims made by the model and the primary text. The third layer of the protocol allows for the detection of unsupported claims during the annotation process and provides a means for incorporating the identified unsupported claims into the error analysis.

This multi-layered protocol has been developed to reflect the current standards of methodology in contemporary grounding research, as well as provide a structured framework to evaluate interpretive reliability. The entire structure of this multi-layered protocol is summarized in Table 1.

Table 1. Structure of the Grounded-Evidence Protocol (GEP)

Component	Description	Enforcement Rule
Restricted Retrieval	Model confined exclusively to chapter-segmented novel text	No access to external sources
Evidence Linking	Every analysis must include a direct quotation (≤20 words)	Citation mandatory
Unsupported-Claim Detection	Identification of claims lacking relevant or accurate quotations	Must be flagged as unsupported

7.5. Reliability Rubric for Analytical Evaluation

An assessment tool, a five-dimensional scoring rubric to evaluate the quality of generated analytic responses from the model, was created with five dimensions related to the essential characteristics of interpretative literary analysis. The first dimension, Textual Fidelity, determined the extent to which the generated claims of interpretation were grounded in the relevant aspects of the textual content of the novel. The second dimension, Accuracy, evaluated the degree to which the model correctly referenced the appropriate characters, events, and/or narrative structures as they occurred in the novel. The third dimension, Analytical Depth, evaluated the extent to which the model could generate high-quality interpretations (deep and insightful) of the novel, rather than simply providing a descriptive analysis of the novel. The fourth dimension, Coherence, evaluated the models' ability to provide evidence to support their claims of interpretation and to present their argument in an organized manner. The final dimension, Linguistic Quality, evaluated the model's ability to communicate its claims of interpretation clearly and concisely, utilizing language that is typical of academic writing. Dimensions and definitions of these dimensions are presented in Table 2.

Table 2. Reliability Rubric (0–5 Scale)

Dimension	Definition	Score Range
Textual Fidelity	The extent to which interpretations rely on verifiable textual grounding	0-5
Accuracy	Correctness of narrative references and factual elements	0-5
Analytical Depth	Degree of interpretive insight beyond surface description	0-5
Coherence	Logical and rhetorical organization of arguments	0-5
Linguistic Quality	Clarity, precision, and academic tone	0-5

The five Rubric Dimensions illustrated in Table 2 were utilized as evaluation criteria to score each of the Model Outputs. Each response received a numerical score (on a scale from 0-5) for each dimension, allowing quantitative summation, clustering-based comparison, and comparative analyses of responses [31].

7.6. Annotation Procedure and Inter-Rater Reliability

All model-generated responses were reviewed by a professional-level human annotator who used the evaluation rubric for each task, and for all tasks that required an interpreter to evaluate the generated responses. In addition to reducing potential variability due to individual annotator bias, the independent review of a random sample of twenty percent of all responses provided further support to increase the reliability of the annotator scoring process, and also provided additional support to the overall design of the evaluative methodology. Inter-annotator agreement was measured on a randomly selected 20% sample using Cohen's Kappa. For the Coherence dimension, agreement was $\kappa = 0.40$, indicating moderate consistency between annotators [32].

7.7. Analytical Data Aggregation

Following evaluation of all responses by annotators, aggregate scores of the evaluative scores were statistically derived from the annotators' evaluations using a systematic method to identify trends of interpretation and to display interpreters' consistency and variation. The aggregation process, as a statistical procedure, converts evaluative judgments from an individual level to a quantifiable measure of the extent to which the model demonstrates its typical analytical behaviors on all dimensions. Furthermore, measures the degree of the model's stable analytical behaviors across tasks (standard deviation), and uses the minimum and maximum values to establish the full scope of analytical behaviors exhibited by the model. The aggregation process provides the basis of data upon which the model's subsequent interpretive and comparative analyses will be developed, and is consistent with current practices in computational humanities research and large-scale data modeling approaches [33]. The study demonstrated that an identical and measurable performance pattern existed throughout all five evaluation criteria described in the research design for the systematic review of the output provided with annotations. The empirical results clearly indicated that the LLM, operating within the strict constraints of a text-based environment, demonstrated a strong ability to process language and structure, and to maintain coherence as well as stylistic precision, but showed more modest ability in higher-order interpretive reasoning. In addition to providing a comprehensive understanding of the model's reliability, both quantitatively and qualitatively, the results also clarified the boundaries of the model's interpretive abilities.

8. Results and Analysis

This section will provide a thorough evaluation of the model’s analytical capabilities under the limitations of the prior-stated ground conditions. A first step in evaluating the model is the presentation of quantifiable scores of all five of the reliability metrics to show overall trends of these metrics. In addition to the quantitative analysis, a qualitative evaluation will be made to highlight strengths, weaknesses, and common error types of the model. Overall, the results of this section will provide a comprehensive view of the model's ability to create valid interpretations of literature from text.

8.1. Quantitative Performance Profile

In order to provide a global evaluation of the model’s performance with respect to each element of the rubric, Table 3 presents the aggregated evaluative scores from the rubric in terms of mean, standard deviation, and score range. The data indicate that the model demonstrates consistent performance levels for almost all elements of the rubric. These results are indicative of the observed performance profile of the ChatGPT-5.2 model when evaluated using the Grounded-Evidence Protocol.

Table 3. Aggregate performance scores across reliability dimensions

Dimension	Mean	SD	Min	Max
Textual Fidelity	4.30	0.42	3.40	4.90
Accuracy	4.10	0.51	3.20	4.80
Analytical Depth	3.95	0.47	3.10	4.60
Coherence	4.50	0.36	3.80	4.90
Linguistic Quality	4.58	0.30	4.00	4.90
Overall Score	4.19	—	—	—

8.2. Dimension-wise Interpretation of Results

Quantitative results demonstrated that the model performed well overall when using the Grounded-Evidence protocol. However, the distribution of quantitative results across each of the five dimensions reveals significant differences in both the models' interpretive strengths and limitations. Textual fidelity (Grounding). Overall, the model had a high textual fidelity with an average of 4.30. This indicates that most of the model's interpretative claims are related to the actual quotes from the novel, and the evidence linking requirement has successfully constrained the model's tendency to produce un-traceable interpretation. However, some decreases in textual fidelity were noted when the narrative was written in metaphors or ambiguous words.

Overall accuracy. The model's overall accuracy score was 4.10, which represents a general correctness in identifying events, situational contexts, and relationships among characters; errors occurred mostly because of slight interpretations from the passages, which made the quotations relevant to the passage, but not sufficient to support the level of the claim. Therefore, this type of grounding has been shown to be effective in reducing the occurrence of factual hallucinations; it is still possible for the model to make overly generalized inferences.

Analytical Depth. Analytical Depth had the lowest mean rating (3.95) and indicates that the model’s ability to produce rich interpretations of the input material was average relative to the model's overall strengths. The model demonstrated a consistent ability to generate coherent and contextually grounded descriptions; however, the model did not demonstrate the same level of consistency when producing deeper symbolic, thematic, or culturally embedded interpretations of the input material. This limitation is consistent with the general trend of LLMs to favor fluency and structural coherence over abstract conceptual synthesis. Coherence. Coherence demonstrated the best overall mean score (4.50) as it was evident The model produced a consistent flow of argument with good organization to each response. Although many responses did include some unsupported claims as to their interpretation, they still provided a logical sequence of thought for the most part. In addition to being able to produce academically-styled writing, this ability to create well-structured explanatory writing is also another example of the model's ability to create long-form writing. Linguistic Quality. High Linguistic Quality was found to be the highest average score (4.58). The data indicates that this model is capable of producing well-written, coherent texts with a formal tone, which is consistent with previous research that identifies this as a major area of strength for the model. Although high-quality language can create the impression of a good interpretation, it does not guarantee that the interpretation is correct. Therefore, it is just as important to evaluate how well the evidence aligns with the interpretation, along with how fluent the language is.

8.3. Interpretation of Quantitative Findings

Numerical results demonstrate that Coherence and Linguistic Quality have the greatest mean value, as these are two key components in assessing the models' ability to understand how academic articles use Rhetorical Structure and academic writing styles. These results suggest that the model is able to consistently produce well-formatted responses and responses written in an appropriate style for a particular discipline, in addition to being highly fluent in language, as it has been trained on large-scale corpora.

Similarly, Textual Fidelity and Accuracy show an equally good quality of answer, which indicates the effectiveness of the grounding protocol in keeping the model's interpretative claims grounded to the actual textual evidence. However, although there are only small gaps between these two measurements, there are also instances when the model would use quotes that relate to but do not completely support the model's interpretative claims.

In contrast to this finding, Analytical Depth was found to have a significantly lower score than the other dimensions, indicating that the model has inherent limitations in areas such as abstract reasoning, symbolic reading, and thematic synthesis. The model is exceptional at descriptive analysis and at using structural analysis to support its descriptions; however, it does not go beyond the middle-level of conceptual interpretation, as has been reported by recent research concerning the relationship between fluency and interpretative richness.

Ultimately, the model is identified as an exceptionally linguistically competent system that is capable of generating well-supported, coherent literary analyses that are clearly grounded. However, it is also clear from these results that the model has a defined limit in terms of its ability for deeper interpretive reasoning.

8.4. Visualization of Reliability Dimensions

In order to provide a clear view of how well the model performed on the five evaluation criteria, linguistic, structural, and interpretive a graphical illustration of how the model performed on each of these evaluation criteria can be seen in Figure 1. Figure 1 shows how well the model has done on each of the three categories of analytical ability as compared to others.

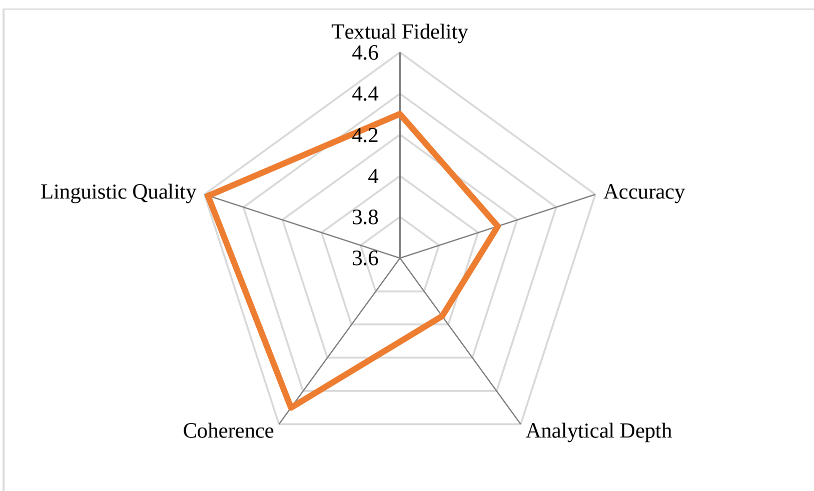

Fig. 1 Average scores across reliability dimensions

Figure 1 shows that the model had the most stable scores in both Coherence and Linguistic Quality, as it has consistently demonstrated a high level of rhetorical clarity and stylistic control.

The model's Textual Fidelity and accuracy are at middle range levels of stability, while the Analytical Depth dimension had the least amount of stable variance and indicates that the model is less capable of producing a high degree of conceptual richness/abstraction in the interpretation of texts.

In general, this figure illustrates a strong linguistic fluency combined with an interpretive capacity of moderate levels.

8.5. Distribution of Error Types

Incorrect or weakly supported output was examined to identify four types of recurrent errors: unsupported interpretative statements, insufficient quotes, hallucinations of symbolic reading, and incoherent structure. These error patterns were used to interpret low-scoring outputs and explain variance across rubric dimensions. The relative distribution of these errors is illustrated by Figure 2. Figure 2 shows the proportionate amounts for the most frequent recurrent errors for four error types: unsupported claims (38%), missing quotation (27%), hallucinations of symbolism (22%), and structural issues (13%). The numbers are representative of how much each one has contributed to the total number of errors in the reliability report.

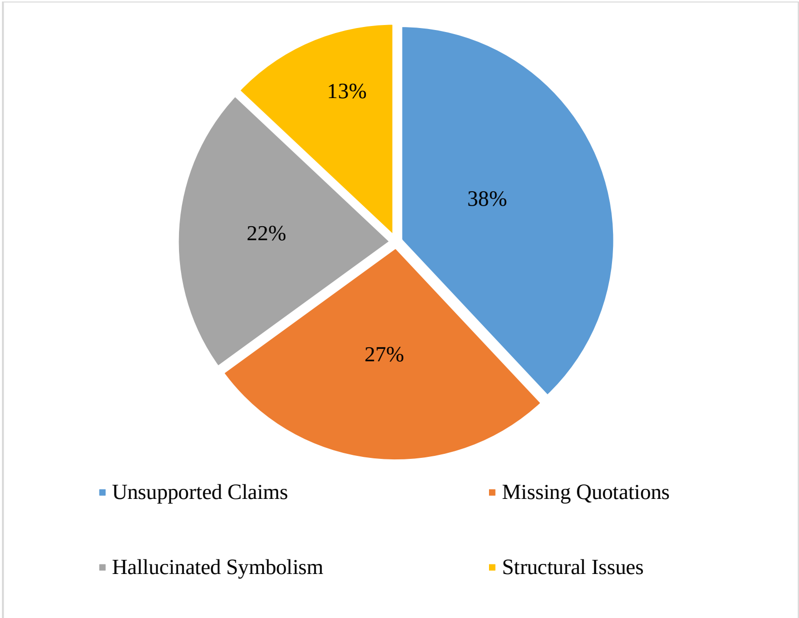

Fig. 2 Distribution of error types in model outputs

8.6. Qualitative Error Patterns

A qualitative review of low-scoring outputs reveals that errors were not random but clustered around a small set of recurrent patterns. These patterns reflect points where interpretive demands exceeded what could be reliably supported through direct textual grounding.

Unsupported interpretive claims. In some cases, the model produced interpretive statements that were plausible but not verifiable through the quotations provided. This typically occurred when the narrative context was emotionally suggestive but did not explicitly state the thematic conclusion drawn by the model.

Missing or insufficient quotation support. A second frequent pattern was the use of quotations that were related to the passage but too general to support the specific interpretive claim. This indicates that the model can select relevant text but does not always establish a precise evidence-to-claim mapping.

The Model's tendency to create Symbolic interpretations of passages rich in metaphors that are outside the limits of the passage, as well as its tendency to generate readings that lack textual evidence to support the symbolic interpretation of such passages, is an example of hallucinated symbolic interpretation.

Although structural incoherence was observed in only a few instances, it occurred with greater frequency when the model was asked to synthesize multiple narrative layers which resulted in an output that lacked logical order and had become more descriptive rather than analytical in nature.

8.7. Qualitative Characteristics of Interpretive Output

The ability of the model to provide an interpretive assessment of narrative may also be demonstrated by a detailed examination of how the model interprets the narrative from generated responses. When the model has available cueing for narrative, it consistently provides coherent and detailed interpretations of those cues. It is possible that this occurs as a result of the cues in the input providing the model with clearly definable textual signals upon which it can base its analysis. Furthermore, based on responses where there is a cohesive rhetorical structure, the use of specific terms in response, and proper use of formal academic language, the model demonstrates a strong level of consistency in its performance in the areas of both linguistic and organizational skills.

However, the model loses its consistency of output in sections where the model requires symbolic inference, abstract interpretation, or cultural interpretation of symbols within a text. The model usually produces responses that reflect superficial depth, rather than conceptual depth. Furthermore, the model will sometimes extend the interpretation of the text beyond the narrative support available, generating responses that are coherent in their structure but are ungrounded in the original text in terms of content.

8.8. Application of the Reliability Framework to Model Outputs

A representative passage from **"الرصاصة الصديقة"** was selected for a detailed analysis to show how the model works in action. The micro-analysis also demonstrates the connection between the model's analytical behavior and the quantitative results provided above.

8.8.1. Selected Passage

"كان الصمت يحيط به كجدار كثيف، ومع ذلك شعر بأن الكلمات التي لم يقلها كانت تصرخ داخله بقوة لا يستطيع احتمالها."

This example is a useful test of how well the model can represent the complexity of this passage, where there are expressions of emotional compression, of inner tension, and of symbolic juxtaposition between a person's silent or outer expression and their internal eruptions. This passage makes an excellent test case to assess the model's interpretive capacity.

8.8.2. Model-Generated Interpretation

"يكشف المقطع عن حالة اختناق نفسي يعيشها الراوي، حيث يتحوّل الصمت الخارجي إلى فضاء ضاغط، بينما تتصاعد داخله رغبة مكبوتة في التعبير. ويشير وصف ‘الكلمات التي لم يقلها’ إلى توتر داخلي يوشك على الانفجار، ما يعكس هشاشة التوازن العاطفي في اللحظة السردية."

8.8.3. Evaluation of the Output

The answer has a strong similarity to the textual cues provided in the prompt, therefore, indicating strong faithfulness to the cues along with a correct identification of the overall narrative tone/mood. The answer also presents a well-organized answer and possesses a high level of linguistic quality that is consistent with the best features of the model. The answer is primarily descriptive in nature and does not demonstrate deep analysis of concepts; this represents the mid-range analysis demonstrated by the model. While the answer was reasonable and contextually appropriate, it did not extend beyond simple symbolic or psychological theoretical interpretations; these are representative of the models' general tendency to interpret.

8.9. Consolidated Assessment

The combination of the qualitative and quantitative data from the study above demonstrates that this model possesses several positive characteristics as an analytical tool. The model is able to clearly articulate its thoughts through language, maintain a coherent structure and organizational framework to support its assertions, and utilize empirical evidence to substantiate what it states.

The model produces output that is almost always supported by evidence, written well, and in an appropriate style, making it a useful tool for developing initial comments about literature. However, the model's inability to develop deeper connections between symbols and theories for higher-level interpretations of literature further demonstrates the need for human critical thinking in interpreting literature at the highest levels.

A reliability rating of 4.19/5 illustrates this duality of a tool that performs well within established boundaries and limitations defined explicitly through written boundaries; however, it does so without the ability to provide sophisticated interpretations of data. These research findings further enhance the growing body of literature regarding the use of LLMs as analytical tools versus replacement tools for human literary judgment, particularly in culturally rich symbolic environments that have complex narrative structures.

9. Discussion

This study offers a deeper understanding of a current large language model's reaction to the interpretive demands of Arabic literature presented in a narrative structure when it operates exclusively as a text-based entity. The study demonstrates through both quantitative and qualitative assessment methods the analytical tendencies, the structural strengths

Moreover, limitations of the model, this data will be useful for a broader scholarly discourse on the reliability of large language models, how well they are grounded in texts, and whether or not they can interpret literary works through computational means.

9.1. Reliability and Grounded Reasoning Under Textual Constraints

The strongest evidence from this evaluation is that the model has demonstrated an excellent ability to stay grounded in its textual data. The high scores for both textual fidelity and accuracy show that the model was able to follow the pathways it was allowed in the Grounded-Evidence Protocol, which significantly limited the model's ability to produce unsubstantiated claims. This is vastly different than many documented examples of the tendency for models to hallucinate when generating content without boundaries or constraints on their inferential processes. Although the narrow difference between fidelity and accuracy indicates that there are still elements of interpretive drift present even after the model was strictly grounded in the text, it also shows that textual grounding will not eliminate all forms of interpretive drift. In some instances, the model chose quotes that were relevant in the context, and these represent a larger trend seen in recent literature: LLMs are very capable of finding contextualized linguistic information; however, they continue to have difficulty establishing direct logical relationships between evidence and argument.

9.2. The Linguistic Advantage: Fluency Without Depth

While the high scores for both Coherence and Linguistic Quality suggest that the model's ability to organize its responses in a way that appears to be coherent, as well as its grammatically correct writing skills, are extremely strong, this linguistic proficiency provides it with the ability to present very complex-sounding analyses, which may actually be lacking in depth.

A fundamental aspect of understanding the nature of the model's analysis, therefore, is the tension between its articulate presentation and its often superficially developed content. In spite of being highly fluent in presenting its views, the model has been shown to have difficulty delving into the deeper symbolic, thematic, or cultural aspects of material. The results here provide evidence of an increasingly accepted view among researchers that many LLMs demonstrate analytical mimicry, producing highly polished language and structure while at the same time providing a limited basis for developing abstract theoretical frameworks.

9.3. Interpretive Ceiling in Tasks Requiring Conceptual Abstraction

Analytical Depth is a low score for the model, which shows how much interpretative there is for the model. The majority of literary criticism involves more than just describing the literal content of a work, as it involves discovering metaphoric relationships, ideological undertones, and psychological complexities that are found below the surface level of what was literally written. Therefore, this type of interpretation will require some form of conceptual synthesis as opposed to simply extracting from the text.

While the model performed very well in identifying explicit aspects of a story's narrative, its performance drops off significantly when identifying and interpreting those parts of a story that are unclear, symbolic, and allusively based. As such, the model seems to be inclined to generalize and to make too wide an interpretation of a passage, lacking a specific theory-based understanding. The limitations of this interpretation are based on the well-known limits of ChatGPT-5.2's architecture as a tool for cognition, namely, in using symbolic abstraction to reason culturally. Similar to many other high-cognitive areas of knowledge, LLMs do very well at performing recognition-type tasks; however, they tend to perform poorly at abstracting and synthesizing information.

9.4. Error Patterns as Indicators of Interpretive Stress

The four most common kinds of errors, including unsupported claims, quotes that do not support a claim well enough, hallucinations involving symbolism, and structural incoherence, provide an important way to identify the points where the model's reasoning is becoming strained. The first two categories reflect a struggle by the model to perform deeper analysis when the model has little in terms of textual cues. The third category reflects the tension created for the model between producing fluent text and faithfully interpreting the source material. The fourth category, although the least frequently occurring, represents instances where the model attempts to synthesize a coherent narrative from conflicting or opaque narrative cues. Together, these four categories illustrate that the model's interpretive limitations do not occur randomly; rather, they occur systematically as the cognitive demands of abstracting far exceed the model's ability to process them reliably.

9.5. Implications for Computational Literary Analysis

This study disputes the idea that LLMs are capable of doing independent literary interpretation at a high level. The study finds that the LLM can complete many tasks that are structurally explicit and also has an advantage because of its ability to use formal, academic writing; however, the study found that there is a clear limit to how well the model can perform when completing tasks that require original thought, cultural understanding, and symbolic reasoning.

Nonetheless, this does not reduce the model's usefulness in educational settings for supporting student learning with preliminary analysis, thematic mapping, or simply by providing students with comments about a piece of literature that are grounded in the text itself. However, relying on this model for making claims about the meanings of texts remains too early to consider.

From a baseline perspective, while this study does not position the LLM as a replacement for expert literary criticism, the rubric dimensions provide an implicit reference to human baseline expectations. In particular, expert human analysis would be expected to achieve near-ceiling performance in analytical depth and cultural interpretation, while LLM outputs are primarily evaluated for evidence-grounded reliability and coherence. Therefore, the benchmark is designed to measure consistency and faithfulness under constraints rather than outperforming human interpretive judgment.

9.6. Positioning the Model Within Broader LLM Evaluation Practices

While the performance of this model aligns with many other LLMs across domains, the pattern observed across all domains is a robust linguistic structure; however, the model lacks the depth needed in reasoning-based tasks. The reliability score of 4.19/5 demonstrates this model to be an effective interpreter under constraint, but still bound by the inference constraints seen in models of this generation.

Additionally, the grounding protocol demonstrates its significance in controlling hallucinations and maintaining the quality of outputs, thus showing the need for a controlled environment for reliable interpretive performance.

9.7. Overall Synthesis

The data as a whole demonstrates that LLMs are able to perform well-structured literary analysis, which is linked to textual information, but also shows an ability for the LLMs to do this with a significant degree of coherence, linguistic precision, and adherence to evidence; however, the inability of LLMs to understand abstract ideas, to symbolize concepts, and to place cultural context within the framework of analysis demonstrates the ongoing necessity of human interpretative judgments. Ultimately, the research supports a model of collaboration between humans and computers: while LLMs produce high-quality, coherent, linguistically precise, and rhetorically grounded analyses, they cannot provide the depth of knowledge or the contextualized conceptualizations that are characteristic of professional literary interpreters.

10. Conclusion

This study presents a systematic, empirically grounded evaluation of a contemporary Large Language Model's ability to conduct text-bound literary analysis of an Arabic narrative. Using a controlled grounding protocol and a multi-dimensional reliability rubric, it offers one of the earliest structured assessments of LLM interpretive behavior in Arabic literary criticism. The results showed that the use of explicit textual signals was effective in supporting high levels of linguistic fluency, structural coherence, and evidence-based reasoning; however, the grounding protocol was effective in controlling the frequency of hallucinatory responses. In contrast, the tasks that were designed to assess symbolic inference, cultural contextualization, and theoretical abstraction consistently yielded a limited number of interpretations and/or superficial interpretations. Overall, these results support the notion that LLMs can be used for basic literary analytical functions but that they should not replace expert human interpretation in more complex, advanced forms of literary interpretation. The study emphasizes the need for rigorous methodology and critical oversight of AI- assisted humanities research, and suggests that future research on the use of LLMs in analyzing Arabic literature could include comparative and genre-based analyses of how LLMs engage with various aspects of the complexity of Arabic literature.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] Chen Ling et al., "Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey," *ACM Computing Surveys*, vol. 58, no. 3, pp. 1-39, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[2] Leonora Kaldaras, Kevin Haudek, and Joseph Krajcik, "Employing Automatic Analysis Tools Aligned to Learning Progressions to Assess Knowledge Application and Support Learning in STEM," *International Journal of STEM Education*, vol. 11, no. 1, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Lukas Thode, Uamr Iftikhar, and Daniel Mendez, "Exploring the use of LLMs for the Selection Phase in Systematic Literature Studies," *Information and Software Technology*, vol. 184, pp. 1-10, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[4] Hany Rashwan, "Literary Genre as a Theoretical Colonization by Modernism: Arabic *Balāghah* and its Literariness in Ancient Egyptian Literature," *Interdisciplinary Literary Studies*, vol. 23, no. 1, pp. 24-68, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Lachhab Youssef et al., "Enhancing Arabic Aspect Category Detection using Large Language Models (LLMs)," *Results in Engineering*, vol. 26, pp. 1-9, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[6] Tahani S. Alazzam, Musa A. Alzghoul, and Raghad M. Alzghoul, "Exploring AI's Capability in Translating English Metaphors into Arabic," *Theory and Practice in Language Studies*, vol. 15, no. 7, pp. 1-8, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[7] Mohammad Hasan Altarawneh et al., "The Relationship between Cross-Cutting Factors and Knowledge, Learning Outcomes, and Skills in Dual Degree Programs," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 8, pp. 3410-3422, 2024. [Google Scholar] [Publisher Link]

[8] Saif Al Deen Lutfi Ali Al Ghammaz, "Revisiting William J. Shakespeare's the Tempest from a Colonial and Postcolonial Lens," *Theory and Practice in Language Studies*, vol. 13, no. 6, pp. 1373-1378, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Konstantinos I. Roumeliotis, Nikolaos D. Tselikas, and Dimitrios K. Nasiopoulos, "LLMs and NLP Models in Cryptocurrency Sentiment Analysis: A Comparative Classification Study," *Big Data and Cognitive Computing*, vol. 8, no. 6, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Ahmed Abdelali et al., "LAraBench: Benchmarking Arabic AI with Large Language Models," *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, St. Julian's, Malta, pp. 487-520, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Ebtesam Almazrouei et al., "AlGhafa Evaluation Benchmark for Arabic Language Models," *Proceedings of ArabicNLP*, Association for Computational Linguistics, pp. 244-275, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Safa AlBallaa et al., "GATmath and GATLc: Comprehensive Benchmarks for Evaluating Arabic Large Language Models," *PLOS One*, vol. 20, no. 9, pp. 1-24, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[13] Aline Godfroid, Brittany Finch, and Joanne Koh, "Reporting Eye-Tracking Research in Second Language Acquisition and Bilingualism: A Synthesis and Field-Specific Guidelines," *Language Learning*, vol. 75, no. 1, pp. 250-294, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[14] Aysh Alhroob et al., "Enhancing Software Testing with Genetic Algorithm and Binary Search: Integrating Error Classification and Debugging Through Clustering," *Journal of Information Systems Engineering and Management*, vol. 10, no. 17s, pp. 117-125, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[15] Manasa Koppula et al., "AI-Powered Chatbot for FDA Drug Labeling Information Retrieval: OpenAI GPT for Grounded Question Answering," *Analytics*, vol. 4, no. 4, pp. 1-18, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[16] Othmane Friha et al., "LLM-based Edge Intelligence: A Comprehensive Survey on Architectures, Applications, Security and Trustworthiness," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 5799-5856, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Nabil Arman, Faisal Khamayseh, and Eman Awad, "A Semi-Automated Approach for Classifying Non Functional Arabic user Requirements using NLP Tools," *International Journal of Advances in Soft Computing and its Applications*, vol. 17, no. 1, pp. 277-294, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[18] Zichao Lin et al., "Towards Trustworthy LLMs: A Review on Debiasing and Dehallucinating in Large Language Models," *Artificial Intelligence Review*, vol. 57, no. 9, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Sangwoo Heo, Sungwook Son, and Hyunwoo Park, "Halucheck: Integrating Hallucination Detection Techniques in Llm-based Conversational Systems," *SSRN Electronic Journal*, pp. 1-30, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Stephanie Lin, Jacob Hilton, and Owain Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Dublin, Ireland, vol. 1, pp. 3214-3252, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[21] Katie Matton et al., "Walk the Talk? Measuring the Faithfulness of Large Language Model Explanations," *International Conference on Learning Representations*, vol. 2025, pp. 73212-73277, 2025. [Google Scholar] [Publisher Link]

[22] Yue Wu, Peng Hu, and Derek D. Wang, "The AI Annotator: Large Language Models' Potential in Scoring Sustainability Reports," *Systems*, vol. 13, no. 10, pp. 1-28, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[23] Percy Liang et al., "Holistic Evaluation of Language Models," *arXiv preprint*, pp. 1-162, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[24] Karmel Shehadeh, Nabil Arman, and Faisal Khamayseh, "Classification of Arabic user Requirements: A Semi-Automated Approach using NLP Tools," *International Journal of Advances in Soft Computing and its Applications*, vol. 16, no. 3, pp. 1-14, 2024. [CrossRef] [Google Scholar]

[25] Ahmed Adel ElSabagh, Shahira Shaaban Azab, and Hesham Ahmed Hefny, "A Comprehensive Survey on Arabic Text Augmentation: Approaches, Challenges, and Applications," *Neural Computing and Applications*, vol. 37, no. 10, pp. 7015-7048, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[26] Preeti et al., "Quantitative Analysis of Literary Texts: Computational Approaches in Digital Humanities Research," *Educational Administration: Theory and Practice*, vol. 30, no. 5, pp. 5234-5240, 2024. [Google Scholar]

[27] Yang Liu et al., "Datasets for Large Language Models: A Comprehensive Survey," *Artificial Intelligence Review*, vol. 58, no. 12, pp. 1-78, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[28] Malak Mashabi, Shahad Al-Khalifa, and Hend Al-Khalifa, "A Survey of Large Language Models for Arabic Language and its Dialects," *ACM Transactions on Asian and Low-Resource Language Information Processing*, pp. 1-44, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[29] Juho Pääkkönen, and Petri Ylikoski, "Humanistic Interpretation and Machine Learning," *Synthese*, vol. 199, no. 1-2, pp. 1461-1497, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[30] Rusul Mumtaz et al., "PDIS: A Service Layer for Privacy and Detecting Intrusions in Cloud Computing," *International Journal of Advances in Soft Computing and its Applications*, vol. 14, no. 2, pp. 15-35, 2022. [CrossRef] [Google Scholar]

[31] Wael Alzyadat et al., "Big Data, Classification, Clustering and Generate Rules: An Inevitably Intertwined for Prediction," *2021 International Conference on Information Technology (ICIT)*, Amman, Jordan, pp. 149-155, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[32] Gerald Rau, and Yu-Shan Shih, "Evaluation of Cohen's Kappa and other Measures of Inter-Rater Agreement for Genre Analysis and Other Nominal Data," *Journal of English for Academic Purposes*, vol. 53, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[33] Wael Jumah Alzyadat et al., "Fuzzy Map Approach for Accruing Velocity of Big Data," *COMPUSOFT: An International Journal of Advanced Computer Technology*, vol. 8, no. 4, pp. 1-5, 2019. [Google Scholar]