Mantiqu Tayr: Journal of Arabic Language, Vol. 6, No. 1, Januari 2026

ChatGPT vs Gemini: Which Digs Deeper into Arabic Semantics?

Nely Rahmawati Zaimah*[1], Chafidhoh Rizqiyah[2], Syamsul Hadi[3], Rifatul Muthiah[4], Wakhidati Nurrohmah Putri[5]

[1,3]Islamic Elementary School Teacher Education Program Sekolah Tinggi Agama Islam Al-Anwar Rembang, Indonesia.
[2]Islamic Religious Teaching Sekolah Tinggi Agama Islam Subang, Indonesia.
[4]Islamic Education Management Sekolah Tinggi Agama Islam Al-Kamal Rembang, Indonesia.
[5]Arabic Language Education Department Universitas Islam Negeri Salatiga, Indonesia.
Correspondence Address: nelyrahmawati@staialanwar.ac.id

Received: 17-12-2025	Revised: 06-02-2026	Accepted: 08-02-2026

Abstract

This study examined the performance of AI models in translating classical Arabic grammatical literature, focusing on *Alfiyah Ibn Mālik* and *Naẓm al-Imrīṭī*, two foundational texts marked by dense syntactic structures and strong pedagogical significance. ChatGPT and Gemini were evaluated in terms of translation accuracy, terminological precision, and contextual sensitivity. A panel of expert evaluators with more than fifteen years of experience in Arabic instruction assessed each model's capacity to apply syntactic rules, preserve semantic coherence, and maintain stylistic and didactic integrity. The aim and scope of the paper centred on measuring translation quality through a combined framework of METEOR scoring and human expert judgement. Qualitative evaluation further explored the models' adaptability to classical Arabic rhetorical patterns and instructional conventions. The results showed that ChatGPT achieved higher lexical alignment and word-level accuracy than Gemini according to METEOR scores; however, both models demonstrated notable limitations in rendering idiomatic expressions and conveying deeper grammatical and contextual meanings. Statistical analysis using the Mann–Whitney U test revealed no significant difference between the two models, underscoring the limited explanatory power of automated metrics when applied to highly structured classical texts. These findings underscored the ongoing need for expert validation beyond numerical scoring and supported the adoption of a hybrid translation framework, in which AI-generated outputs are systematically refined through scholarly review. Future research was suggested to broaden the textual corpus, incorporate additional AI models and evaluation metrics, and further strengthen expert-led validation to enhance the reliability of AI-assisted translation in advanced Arabic grammatical studies.

Keywords: Arabic AI Translation, ChatGPT and Gemini, METEOR Evaluation, Semantic Interpretation

ملخص

تناولت هذه الدراسة أداء نماذج الذكاء الاصطناعي في ترجمة الأدبيات النحوية العربية الكلاسيكية، مع تركيز خاص على «ألفية ابن مالك» و«نظم الإمرِيطِيّ»، بوصفهما متنين تعليميين يتميّزان بكثافة البنية النحوية وعمق الحمولة الدلالية، حيث تم تقييم نموذجي شات جي بي تي وجيميني من حيث دقة الترجمة، وضبط المصطلح النحوي، والحساسية السياقية. وقد اضطلعت لجنة من الخبراء المتخصصين، ممن تزيد خبرتهم على خمسة عشر عاماً في تدريس العربية وعلومها، بتقدير قدرة كل نموذج على تطبيق القواعد النحوية، والحفاظ على الاتساق الدلالي، وصون التكامل الأسلوبي والبعد الديداكتيكي. وتمحور هدف البحث حول قياس جودة الترجمة عبر إطار تقييمي مدمج يجمع بين درجات مقياس «ميتيور» والتقييم البشري الخبروي، بما يتيح الجمع بين المعطيات الكمية والتحليلات النوعية، كما استقصى التقييم النوعي مدى قابلية النماذج للتكيّف مع الأساليب البلاغية الكلاسيكية وممارسات التعليم التقليدية. وأظهرت النتائج أن شات جي بي تي حقق مستوى أعلى من المحاذاة المعجمية والدقة على مستوى المفردة مقارنة

DOI: 10.25217/mantiqutayr.v6i1.7146

بجيميني وفق درجات «ميتيور»، غير أن كلا النموذجين أظهرا محدوديات واضحة في نقل التعابير الاصطلاحية والدلالات النحوية والسياقية العميقة. كما بينت التحليلات الإحصائية باستخدام اختبار مان ويتني يو عدم وجود فروق ذات دلالة إحصائية بين النموذجين، مما أبرز محدودية القدرة التفسيرية للمقاييس الآلية عند تطبيقها على نصوص كلاسيكية معقدة البنية، وسلط الضوء على ضرورة استمرار التحقق الخبروي إلى جانب المؤشرات الرقمية، مع دعم اعتماد إطار ترجمة هجين تُنقَّح فيه مخرجات النماذج بصورة منهجية من خلال مراجعة علمية متخصصة، والتوصية بتوسيع متن النصوص المدروسة (المدوَّنة النصية)، وإدراج نماذج ومعايير تقييم إضافية، وتعزيز دور الخبراء لرفع موثوقية الترجمة المدعومة بالذكاء الاصطناعي في دراسات النحو العربي المتقدّم.

الكلمات المفتاحية: ألفية ابن مالك، ترجمة الذكاء الاصطناعي العربية، تقييم مِيتِيور، شات جي بي تي و جيميني، نظم الإِمْرِيطِيّ

Introduction

Integrating AI into academic environments holds tremendous potential but must be deployed within a framework of objectivity and integrity to avoid unforeseen issues.[1] This technology provides educators, students, and practitioners with instant access to interpretations, references, and clarifications, while also supporting the development of complex linguistic arguments, elucidating legal concepts, and deepening our understanding of theoretical and referential structures in various textsfrom historical narratives to religious documents.[2] Moreover, AI's versatility extends beyond text, enhancing images, audio, and video to improve efficiency across diverse sectors.[3]

Generative AI technologies such as ChatGPT and Gemini have unlocked new opportunities for translating and interpreting texts from various languages into users' native tongues, particularly Arabic.[4] With the emergence of large language models (LLMs), these tools not only produce and

[1] Muneer Alshater, "*Exploring the Role of Artificial Intelligence in Enhancing Academic Performance: A Case Study of ChatGPT*," SSRN Electronic Journal, ahead of print (AoP), 2022. DOI: https://doi.org/10.2139/ssrn.4312358; Nawaf N. Hamadneh et al., "*Using Artificial Intelligence to Predict Students' Academic Performance in Blended Learning*" Sustainability 14, no. 18 (January 2022): 18. DOI: https://doi.org/10.3390/su141811642

[2] Aqil M Azmi, Abdulaziz O Al-Qabbany, and Amir Hussain, "*Computational and Natural Language Processing Based Studies of Hadith Literature: A Survey,*" Artificial Intelligence Review 52 (2019): 1369–414. DOI: https://doi.org/10.1007/s10462-019-09692-w; Zena Abdulkader and Yousra Al-Irhayim, "*A Review of Arabic Intelligent Chatbots: Developments and Challenges*," Al-Rafidain Engineering Journal (AREJ) 27, no. 2 (September 2022): 178–89. DOI: https://doi.org/10.33899/rengj.2022.132550.1148

[3] Ashish Vaswani et al., *Image Transformer*, February 15, 2018: 5-7. https://openreview.net/forum?id=r16Vyf-0-

[4] Imtiaz Ahmed et al., *ChatGPT vs. Bard: A Comparative Study*, preprint (2023): 1-18. DOI: https://doi.org/10.36227/techrxiv.23536290.v1; Matthew K Hong et al., "*Planning for Natural Language Failures with the AI Playbook,*" nCHI Conference on Human Factors in Computing Systems (CHI '21), May8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, (2021): 1–11. DOI: https://doi.org/10.1145/3411764.3445735

comprehend text in a human-like manner but also significantly enhance cross-lingual understanding.[5] Advanced text interpretation, especially for translating global languages into Arabic or vice versa, necessitates a thorough assessment of the data's accuracy and reliability,[6] as this concern constitutes the primary background of this research.Despite their rapid development, transformer-based generative AI applications, such as ChatGPT and Gemini, continue to face inherent limitations in producing consistently optimal outputs. These limitations stem from computational constraints as well as challenges in managing complex parameters, interpretative protocols, and validation mechanisms necessary for robust linguistic analysis and reliable diagnostic evaluation.[7]

According to previous studies, many transformer models have shown resilience in addressing these challenges; initial concerns have emerged primarily in STEM, healthcare, and other fields.[8] Moreover, the use of AI raises ethical considerations and critical constraints in analysing scientific content, underscoring the importance of ensuring that AI-generated output remains authentic to the original intent of the texts.[9] This demand for accuracy is especially critical in advanced Arabic language studies, where classical texts require both scholarly expertise and innovative AI solutions.

Alfiyah ibn Malik and *Nadham al-Imrithy* are foundational works in the Arabic grammatical tradition, revered for their dense and nuanced exploration of syntax, morphology, and semantics within a poetic framework. Interpreting these texts presents significant challenges due to their reliance on intricate grammatical rules and context-sensitive meanings. For centuries, scholars and students have studied these works under the guidance of trained experts, as the interpretative process often involves navigating multiple layers of meaning.[10] Ensuring the authenticity of the

[5] Bruno Campello de Souza, Agostinho Serrano de Andrade Neto, and Antonio Roazzi, "*Are the New AIs Smart Enough to Steal Your Job? IQ Scores for ChatGPT, Microsoft Bing, Google Bard and Quora Poe,*" SSRN Scholarly Paper no. 4412505, Rochester, NY, April 7, (2023): 1-19. DOI: https://doi.org/10.2139/ssrn.4412505; Feyza Dalayli, "*Use of NLP Techniques in Translation by ChatGPT: Case Study,*" in Proceedings of the Workshop on Computational Terminology in NLP and Translation Studies (ConTeNTS) Incorporating the 16th Workshop on Building and Using Comparable Corpora (BUCC), ed. Amal Haddad Haddad et al. (Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, 2023), hlm. 19–25. https://aclanthology.org/2023.contents-1.3; Michael Lozano et al., "*Semantic Depth Redistribution in Large Language Models to Contextual Embedding Preservation,*" preprint, November 5, (2024): 1-10. DOI: https://doi.org/10.22541/au.173083529.98863661/v1

[6] Dalayli, "*Use of NLP Techniques in Translation by ChatGPT*": *Case Study,*" in Proceedings of the Workshop on Computational Terminology in NLP and Translation Studies (ConTeNTS) Incorporating the 16th Workshop on Building and Using Comparable Corpora (BUCC), ed. Amal Haddad Haddad et al. (Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, 2023), hlm. 19–25. https://aclanthology.org/2023.contents-1.3; Faten Khoshafah, "*ChatGPT for Arabic-English Translation: Evaluating the Accuracy,*" Ministry of Education, Yemen, ahead of print, Proceedings of the First ConTenNTS Workshop and the 16th BUCC workshop, 19–25, Verna, April 17, (2023): 5-11. DOI: https://doi.org/10.21203/rs.3.rs-2814154/v2

[7] Ahmed Mohammed Moneus and Yousef Sahari, "*Artificial Intelligence and Human Translation: A Contrastive Study Based on Legal Texts,*" Heliyon 10, no. 6 (March 2024): 1-14. DOI: https://doi.org/10.1016/j.heliyon.2024.e28106

[8] Lucila Carvalho et al., "*How Can We Design for Learning in an AI World?,*" Computers and Education: Artificial Intelligence 3 (2022): 100053. DOI: https://doi.org/10.1016/j.caeai.2022.100053; Douglas Johnson et al., *Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model*, Preprints, (2023): 1-14. DOI: https://doi.org/10.21203/rs.3.rs-2566942/v1

[9] Chhavi Chauhan, "*The Impact of Generative Artificial Intelligence in Scientific Content Synthesis for Authors,*" The American Journal of Pathology 194, no. 8 (August 2024): 1406–8. DOI: https://doi.org/10.1016/j.ajpath.2024.06.002; Gabriëlle Ras, Marcel Van Gerven, and Pim Haselager, *"Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges,"* in Explainable and Interpretable Models in Computer Vision and Machine Learning, ed. Hugo Jair Escalante et al., The Springer Series on Challenges in Machine Learning (Cham: Springer International Publishing, 2018): 19–36. DOI: https://doi.org/10.1007/978-3-319-98131-4_2

[10] Jonathan Porter Berkey, *The Transmission of Knowledge in Medieval Cairo: A Social History of Islamic Education*, Princeton, New Jersey: Princeton University Press, 2014; Anisatun Muthiah and Luqman Zain, *"KonsepIttishal Al-Sanad*

narrative diction (*matn*) and the intricate details of these Arabic texts requires principled alignment with precise AI augmentation and accurate interpretations. Moreover, the cultural sensitivities embedded within these texts necessitate that AI-generated interpretations avoid unintended distortions or misrepresentations of their context or meaning.[11] Because this line of reasoning serves as the primary conceptual foundation of the study, precision in both literal and contextual language analysis is essential, as it underpins the assessment of the reliability and validity of generative AI language models such as ChatGPT, one of the most widely used GAIs globally, and emerging systems like Gemini were evaluated for translation accurary, terminological precision, and contextual sensitivity.[12]

This study addresses a clear gap in the evaluation of Arabic language output, particularly in classical Arabic texts. While much previous research has focused on overall system performance or content adequacy, detailed assessment of linguistic accuracy in Arabic remains limited, especially when modern automated metrics are applied to highly structured grammatical and pedagogical texts. Most existing studies rely on general or modern corpora, which do not reflect the dense syntax, fixed terminology, and rhetorical patterns found in classical works. To respond to this limitation, the present study applies contemporary automated evaluation metrics—METEOR, BLEU, TER, and BERTScore—to classical Arabic grammatical texts, specifically *Alfiyah Ibn Mālik* and *Naẓm al-Imrīṭī*. It also introduces an Arabic-sensitive evaluation design that distinguishes between literal meaning, grammatical terminology, and contextual explanation, and combines automated scoring with expert judgement. In addition, the study explores adjustments to evaluation metrics to better accommodate Arabic-specific features such as diacritics, morphological variation, and compact syntactic structures, thereby strengthening language-centred evaluation for classical Arabic within a modern assessment framework.

This study focuses on two main questions: RQ1: Can ChatGPT and Gemini consistently understand the contextual nuances and variations in classical Arabic texts, as evaluated by literature reviews and expert assessments? RQ2: Is there a significant difference between ChatGPT and Gemini in their ability to interpret classical Arabic texts?

Method

This research employs a mixed-methods approach. Using the Collect-Measure-Repeat (CMR) framework developed by Inel et al.,[13] alongside insights from Raj et al.,[14] to provide a comprehensive

sebagai Syarat Kajian Kitab Kuning dalam Tradisi Pesantren An-Nahdliyyah Cirebon," Jurnal Studi Hadis Nusantara 2, no. 1 (2020): 11. DOI: https://doi.org/10.24235/jshn.v2i1.6746

[11] Islamiyah Sulaeman, Syuhadak Syuhadak, and Insyirah Sulaeman, *"ChatGPT as a New Frontier in Arabic Education Technology,"* Al-Arabi: Journal of Teaching Arabic as a Foreign Language 7, no. 1 (2023): 83–105. DOI: http://dx.doi.org/10.17977/um056v7i1p83-105

[12] Ghazala Bilquise, Samar Ibrahim, and Khaled Shaalan, *"Bilingual AI-Driven Chatbot for Academic Advising,"* International Journal of Advanced Computer Science and Applications 13, no. 8 (2022): 50-57. DOI: https://doi.org/10.14569/IJACSA.2022.0130808; Regina G Russell et al., *"Competencies for the Use of Artificial Intelligence–Based Tools by Health Care Professionals,"* Academic Medicine 98, no. 3 (2023): 348–56. DOI: https://doi.org/10.1097/ACM.0000000000004963

[13] Oana Inel, Tim Draws, and Lora Aroyo, *"Collect, Measure, Repeat: Reliability Factors for Responsible AI Data Collection,"* Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 11, no. 1 (November 2023): 1. DOI: https://doi.org/10.1609/hcomp.v11i1.27547

[14] Harsh Raj et al., *"Semantic Consistency for Assuring Reliability of Large Language Models,"* arXiv:2308.09138, preprint, arXiv, August 17, 2023: 6. DOI: https://doi.org/10.48550/arXiv.2308.09138

comparative evaluation of the semantic outputs generated by ChatGPT and Gemini. While reliability benchmarks proposed by Inel and colleaguesas well as other AI reliability framework stypically emphasise the evaluation of input datasets, this study shifts the analytical focus toward a qualitative comparison of AI-generated augmentations at the output level. The inclusion of human expert assessment provides a deeper and more nuanced basis for comparing the two models beyond automated scoring alone. Quantitatively, the comparative analysis adopted in this study aligns with the evaluation approach employed by Moneus and Sahari, particularly in its use of standardised metrics to benchmark model performance across comparable linguistic tasks.[15] evaluating the performance of ChatGPT and Geminiasdual subjects of analysis using the METEOR metric. By analysing the results from both AI models, the researchers aim to identify their strengths and limitations, offering insights into their applicability in advanced Arabic language studies.

This study conducts a comparative evaluation between the AI models ChatGPT 4.0 (T1) and Gemini 1.5 Flash (T2). Ten queries (Q1 to Q10) were formulated to probe specific aspects of Arabic structures and semantics in *Alfiyah ibn Malik* and *Nadham al-Imrithy*.

Table 1. Queries from the Books of *Alfiyah ibn Malik* and *Nadham al-Imrithy*

Query		Topic (Domain)	Original English Translation (Rephrased from Indonesian Translation by Fuad, 2010)
1	فَارْفَعْ بِضَمَ وَانْصِبَنْ فَتْحاً وَجُرْ ۞ كَسْـراً كَـذِكْرُ اللّهِ عَبْـدَهُ يَسُرْ	Signs of *I'rab* (Case Endings)	"Raise with a dammah, install with a *fatha*, and lower with a *kasra*, like the remembrance of Allah pleases His servant."
2	وَمِنْ ضَمِيْرِ الْرَّفْعِ مَا يَسْتَتِرُ ۞ كَافْعَلْ أوَافِقْ نَغتَبِطْ إِذْ تُشْكرُ	Concealed Pronouns (*Ism DhomirMustatir*)	"And among the pronouns of the nominative case is that which is concealed, like 'do,' 'I agree,' 'we rejoice,' when you are thanked."
3	وَمِنْهُ مَنْقُولٌ كَفَضْلٍ وَأَسَد ۞ وَذُو ارْتِجَال كَسُعَـادَ وَأُدَدْ	Transferred and Improvised Names (*Manqul and Murtajal*)	"And among them are transferred names like *'Fadl'* and *'Asad*,' and those improvised like *'Su'ad'* and *'Udad*.'"
4	الحال وصفّ فضلةّ منتصب ۞ مفهم في حال كفرداً أذهب	حال (Adverbial Modifier)	"The حال is a descriptive word, surplus, منصوب (accusative), indicating a

[15] Ahmed Mohammed Moneus and Yousef Sahari, *"Artificial Intelligence and Human Translation: A Contrastive Study Based on Legal Texts,"* Heliyon 10, no. 6 (March 2024): 1-14. DOI: https://doi.org/10.1016/j.heliyon.2024.e28106

			state, like 'I went alone.'"
5	وَتِلْوَ أَفْعَلَ انْصِبَنَّهُ كما ۞ أوْفَى خَلِيلَيْنَا وَأصْدِقْ بِهِمَا	صيغة التعجب (Expressi-ons of Wonder/Admiration)	"And after ' فعل ' (the verb form used for تعجب), make the object منصوب (accusative), as in 'How loyal are our two friends!' and 'How truthful are they!'"
6	وَلَفْظُهُ الْمَشْهُوْرُ فِيْهِ اَرْبَعُ۞نَفْسٌ وَعَيْنٌ ثُمَّ كُلٌّ اَجْمَعُ	توكيد (Emphasis) - Part 1	"And its well-known words are four: نفس (self), عين (eye), then كل (all) جمع (collective)."
7	وَغَيْرُهَا تَوَابِعٌ لِاَ ◌َجْمَعَ۞ مِنْ اَكْتَعٍ وَاَبْتَعٍ وَاَبْصَعَ	توكيد (Emphasis) - Part 2	"And others are followers for all, from أبتع ,أكتع, and أبصع."
8	هُوَ اسْمُ وَقْتٍ اَوْمَكَانِ نِ انْتَصَبْ ۞كُلٌّ عَلَى تَقْدِيْرِ فِى عِنْدَ اْلعَرَبْ	ظرف (Adverb of Time or Place)	"It is a noun of time or place that is منصوب (accusative). All of them are understood with the preposition 'في' (in) according to the Arabs."
9	وَلَفْظُ اْلاِسْتِثْنَا الَّذِى لَهُ حَوَى۞ اِلاَّ وَغَيْرُ وَسِوَى سُوًى سَوًا	استثناء (Exception)	"And the words of exception that it contains are: إلاّ (except), غير (other than), سوى (besides), سُوى (equivalent), سواءً (equal)."
10	وَهْوَ عَلَى تَقْدِيرٍ فِي أَوْلَامِ ۞ أَوْ مِنْ كَمَكْرِ اللَّيْلِ أَوْ غَلَامِ	ظرف (Adverb of Time or Place) - Additional Example	"And it (the adverb) is understood with the preposition 'في' (in), or 'لام' (for), or 'من' (from), like 'the deceit of the night' or 'my boy.'"

Table 1 illustrates that the queries were extracted from two classical Arabic texts (Q1-Q5 from *Alfiyah ibn Malik* and Q6-Q10 from *Nadham al-Imrithy*) characterised by high semantic density and linguistic complexity. Each query was repeated five times in every single prompt of ChatGPT and Gemini, and systematically structured into fifteen distinct Question Models (QMs): five focusing on literal translation (L), five on terminological explanation (T), and five on contextual adaptability (C). The responses generated from these QMs were subsequently treated as analytical variables to assess accuracy and consistency. Evaluation was conducted by three expert assessors with specialized expertise in Arabic language and grammar, ensuring an informed and reliable judgment of the AI-generated outputs.

The structured query formulation strategy used in this study was designed to examine the performance of AI models across three interrelated linguistic paradigms: literal translation, terminological proficiency, and contextual depth. Each paradigm comprised five carefully designed questions, and each question was repeated across five iterations to ensure consistency and reliability in the evaluation of how ChatGPT and Gemini interpreted and translated classical Arabic texts. Literal translation queries were designed to assess word-for-word accuracy and lexical precision through prompts that explicitly required direct translation of selected verses, phrases, or terms. Terminological proficiency queries aimed to measure the models' understanding of specialised grammatical and linguistic concepts by asking for definitions, explanations, or illustrative examples of key terms. Contextual depth queries extended beyond surface meaning to probe the models' ability to capture underlying meanings, rhetorical intent, and contextual significance, using prompts that required explanation or analysis of passages within their broader grammatical and didactic framework. This tripartite structure enabled a systematic comparison of the models' strengths and limitations across different layers of linguistic competence.

During the evaluation phase, numerical scores for each variable are gathered using a qualitative approach that incorporates cross-referencing and human assessment, ultimately yielding quantitative results. Each variable is rated on an ordinal scale from 0 to 5, and the METEOR metric is then calculated according to the following scheme:

The METEOR (Metric for Evaluation of Translation with Explicit ORdering) score in this study is a literal and terminological evaluation metric. This is similar to what Riina et al. did.[16] Although their study focused on LLMs' medical explanations, the steps involved in the METEOR evaluation are quite similar.

METEOR assesses Model-generated outputs against reference texts through a hierarchical matching process, incorporating exact matches, stemming, and synonym recognition, enabling a more nuanced linguistic assessment benchmarked by expert evaluations. Additionally, the metric applies a fragmentation penalty to ensure alignment coherence. The F-Mean score, which prioritizes recall, accounts for relevant yet overlooked words in the evaluation. However, human assessors remain essential in validating contextual nuances and advanced linguistic interpretations that AI metrics might miss.[17] To ensure a comprehensive assessment, scoring follows a 0-5 scale across all queries and prompting models (Question Models), covering five literal translations, five

[16] Nicholas Riina et al., *"An Evaluation of English to Spanish Medical Translation by Large Language Models,"* in Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations), ed. Marianna Martindale et al. (Chicago, USA: Association for Machine Translation in the Americas, 2024): 222–36. https://aclanthology.org/2024.amta-presentations.15/

[17] Simone Balloccu et al., *"Ask the Experts: Sourcing a High-Quality Nutrition Counseling Dataset through Human-AI Collaboration,"* Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, 2024, 11519–45. DOI: https://doi.org/10.18653/v1/2024.findings-emnlp.674

terminological assessments, and five contextual analyses for both ChatGPT and Gemini. With five iterations per query, this results in a total of 750 scores, representing the models' cumulative performance across all evaluation aspects. The METEOR formula framed for this study is:

$$\text{METEOR} = (1-\gamma).\frac{10\ .\ P\ .\ R}{R+9P} + \gamma \cdot Alignment\ Score$$

$$P = \frac{Matches}{AI\ Output\ Words},\ \text{R} = \frac{Matches}{Reference\ Words}$$

γ = Weight for synonym/stemming matches (default: 0.6)

AlignmentScore: Coherence of alignment (exact: 1, stem: 0.8, synonym: 0.6).

Fragmentation Penalty:

$$\text{Penalty} = 0.5 \cdot \left(\frac{\text{Chunks}}{Matches}\right)^3$$

F-Mean (Recall-Oriented):

$$\text{F-Mean} = \frac{(1+\beta^2)\ .\ P\ .\ R}{\beta^2\ .\ P\ .\ R},\ \beta=2$$

In this study, the scoring formula introduces weight factors (10P and 9P), diverging from the conventional parameter α = 0.9 used to balance precision and recall. This modification aims to emphasize precision-driven alignment in AI-generated outputs while still capturing recall for a more balanced assessment in classical Arabic translation and interpretation. Additionally, the alignment score term γ · Alignment Score is maintained to factor in word order and coherence.

Table 2. Evaluation Metrics (benchmarking)

Aspect	Metric	Human vs. Metric Weight
Literal Translation	METEOR Exact Matches	50% Human, 50% Metric
Terminological Mastery	METEOR Synonym + F-Mean	70% Human, 30% Metric
Contextual Adaptability	Human Rubric	100% Human

Table 2 summarises the evaluation framework used to benchmark model performance across three linguistic aspects. Literal translation combined human judgement and METEOR exact matches equally, terminological mastery relied more heavily on expert evaluation, while contextual adaptability was assessed exclusively through a human rubric due to its interpretative complexity.

Hypotheses:

1. H_0 (Null Hypothesis): There is no significant difference in accuracy between ChatGPT and Gemini in interpreting and translating classical Arabic texts.
2. H_1 (Alternative Hypothesis): There is a significant difference in accuracy between ChatGPT and Gemini.

To test these hypotheses, an independent t-test was conducted using JASP to compare the mean accuracy scores of both models. Key statistical measures, including mean differences, standard deviations, confidence intervals, and effect size (Cohen's d), were examined. Levene's test was applied to assess variance homogeneity before proceeding with the t-test.[18] If normality assumptions were violated, a Mann-Whitney U test was used as a robust alternative.[19] By incorporating the Brown-Forsythe test for equality of variances, this approach ensures that statistical analysis remains valid even when normality assumptions are violated.

Result and Discussion

A. Arabic Semantic Performance: Qualitative Expert Evaluations (Q1)

The qualitative phase of this study provided a comprehensive evaluation of the Arabic semantic performance of two AI models ChatGPT 4.0 (T1) and Gemini 1.5 Flash (T2) using a dataset derived from *Alfiyah ibn Malik* and *Nadham al-Imrithy*. The evaluation involved ten distinct queries (Q1–Q10), each repeated over five iterations, and categorized into three question models: literal translation (QM 1–5), terminological proficiency (QM 6–10), and contextual depth (QM 11–15).

1. Experts' Evaluation and Review

In the literal translation category, both models generally performed well by accurately preserving the structural and orthographic features of the source texts, including letters and diacritical marks (*harakat*). ChatGPT (T1) often consistently maintained these crucial elements more consistently than Gemini (T2). While both models maintained acceptable accuracy levels in Q1 and Q2, T1 consistently delivered responses that adhered more closely to the source's word-for-word integrity. This consistency is particularly significant given the sensitivity of classical Arabic texts to diacritical precision, where even minor deviations can alter the intended meaning.

In contrast, the terminological proficiency category, which assesses the models' ability to correctly interpret and explain specialized grammatical terms, demonstrated a mixed performance for both T1 and T2. Focused on defining technical vocabulary or providing examples of linguistic constructs (such as those related to specific grammatical phenomena in classical Arabic), both models exhibited variability in their responses. In several iterations, both T1 and T2 managed to capture the essence of the terminology accurately; however, there were noticeable instances where the models either offered vague definitions or included unsolicited examples that deviated from the strict terminological focus. T2 showed slight weaknesses in this area, occasionally scoring lower on queries that demanded precise definitions and clear terminological boundaries. These inconsistencies underscore the challenge inherent in translating domain-specific language, where the accuracy of specialized vocabulary is critical.

The third category—contextual depth—proved to be the most challenging for both models, as it requires the AI to not only translate text but also to capture the underlying cultural, historical,

[18] Eric-Jan Wagenmakers, Richard D. Morey, and Michael D. Lee, *"Bayesian Benefits for the Pragmatic Researcher,"* Current Directions in Psychological Science 25, no. 3 (June 2016): 169–76. DOI: https://doi.org/10.1177/0963721416643289

[19] John K. Kruschke, *"Bayesian Analysis Reporting Guidelines,"* Nature Human Behaviour 5, no. 10 (October 2021): 10. DOI: https://doi.org/10.1038/s41562-021-01177-7; Wei Liang and Hongsheng Dai, *"Bayesian Inference,"* in Quantum Chemistry in the Age of Machine Learning (Elsevier, 2023): 233–50. DOI: https://doi.org/10.1016/B978-0-323-90049-2.00005-6

and semantic nuances embedded within the source material. The experts emphasized that the contextual depth category was generally lower compared to the literal translation and terminological proficiency categories. This decline in scores is indicative of the difficulties both models face when attempting to convey deeper meanings and contextual subtleties. T1 tended to provide more detailed and coherent explanations when asked to interpret the underlying significance of a verse or analyse its broader contextual relevance. However, there were instances where T1's responses included unsolicited commentary or additional examples that, while well-intentioned, detracted from the precise answer expected by the evaluators. T2, on the other hand, sometimes produced responses that were either too brief or lacking in the necessary depth to fully capture the complexity of the source text.

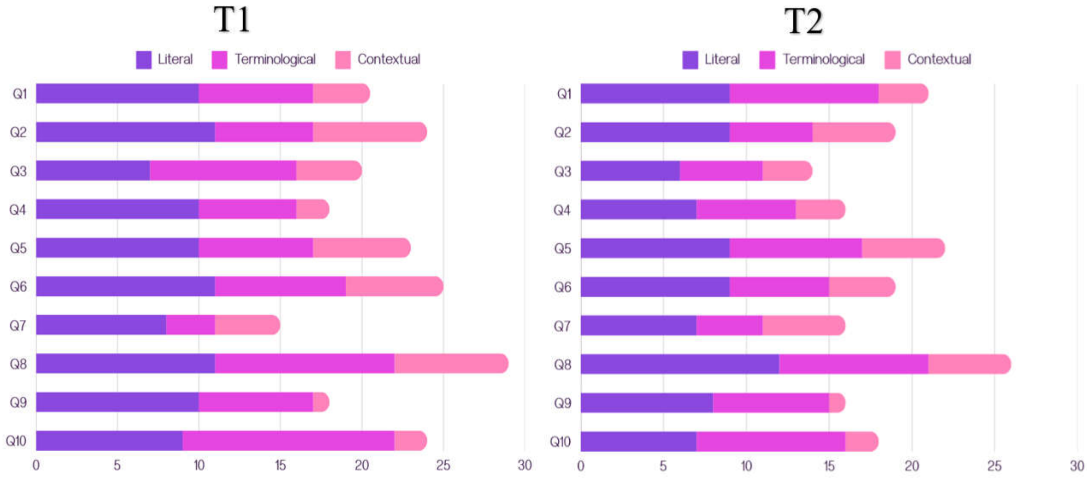

Figure 1. Comparative Performance of ChatGPT and Gemini in Arabic Interpretation Tasks

Figure 1 visualizes the comparison of the translation performance of two AI models, T1 (ChatGPT) and T2 (Gemini). Across ten queries using 15 question models — five literal, five terminological, and five contextual — reviewers assigned scores according to a detailed evaluation rublic. Between T1 and T2, noticeable differences emerge across various query types. In literal translations, both models exhibit similar performance with high accuracy and reliable syntactic fidelity. However, in terminological tasks, T1 outperforms T2, achieving higher reviewer scores by demonstrating a superior grasp of specialized vocabulary and grammatical terminology. In contextual queries, T1 shows a modest edge due to its ability to capture nuanced historical, cultural, and semantic subtleties, despite occasional inconsistencies. Conversely, T2, while more consistent in its output, often falls short in delivering the depth required for complex interpretative tasks. These differences underscore that T1 is generally more adaptable to intricate language features, where as T2 provides steadier if less nuanced, translations.

2. Metric for Evaluation of Translation with Explicit ORdering (METEOR) Alignment

This section presents a quantitative evaluation of METEOR scores to compare the alignment accuracy of the T1 and T2 models. By calculating precision, recall, and the F-Mean metric, we assess how effectively each model matches the reference text. The results offer insights into performance differences across multiple QM instances, highlighting which model demonstrates superior alignment. This quantitative phase provides an objective basis for further qualitative analysis and interpretation.

Table 3. Correlation Analysis: Expert Scores vs. METEOR

Category	Pearson Correlation (r)	Interpretation
Literal Translation	0.89	Strong alignment; METEOR reflects expert judgments on orthography.
Terminological Proficiency	0.62	Moderate alignment; METEOR misses the semantic accuracy of terms.
Contextual Depth	0.41	Weak alignment; METEOR fails to capture cultural nuances.

Table 3 presents the correlation between expert scores and METEOR across three translation quality dimensions. The high correlation (r = 0.89) in literal translation suggests that METEOR effectively mirrors expert evaluations in assessing orthographic accuracy, meaning it reliably identifies word-for-word correctness, including diacritics and letter structures. However, in terminological proficiency, the correlation drops to 0.62. It indicates that METEOR scoring struggles with semantic precision, particularly in distinguishing nuanced meanings of specialized terms. The weakest correlation (r = 0.41) is observed in contextual depth. These findings highlight the crucial role of human evaluation in advanced translation tasks, especially when dealing with nuanced interpretations, cultural references, and domain-specific terminology—areas where automated metrics struggle to assess meaning beyond mere lexical matching. The performance scores of the tested models are presented in the following table.

Table 4. Meteor Score

Model	Total Matches	Total METEOR Score
T1	217	3.122
T2	187	2.476

Table 4 indicates that T1 (ChatGPT) has a total METEOR score of 3.122, surpassing T2 (Gemini) with 2.476. In QM1, T1 (0.406) outperforms T2 (0.266) due to more precise word alignment. However, in QMs such as QM4, QM9, QM11, and QM14 (METEOR ≈ 0.098), both models struggle with specific terminology or complex contexts. In QM15, T1 (0.070) and T2 (0.084) produce responses that are either too brief or misaligned with the reference. T1 demonstrates greater consistency in QM6 (0.308) and QM10 (0.252), while T2 slightly surpasses T1 in QM10 (0.266) due to better keyword alignment.

T1 performs better in capturing word alignment with the reference compared to T2. T1 consistently demonstrates higher precision and recall than T2, as reflected in its higher METEOR scores across most QM instances. For example, in QM 1, 2, and 3, T1 significantly outperforms T2, indicating that T1 aligns words with the reference more effectively. However, there are certain QM cases where both models perform similarly or have only minor differences (e.g., QM 4 and QM 14). This suggests that in specific cases, both models exhibit comparable accuracy in aligning words with the reference.

In this context, the fragmentation penalty means that if the words in the translation do not align sequentially with the reference text, the score will be reduced. The value of 0.3 represents the penalty level applied for non-sequential word alignment. If this penalty level is increased to 0.5, the

METEOR scores for both models will decrease more significantly. However, even with the increased penalty, the T1 model is likely to maintain a better score compared to the T2 model. This indicates that T1 is superior in keeping the word order aligned with the reference.

These findings have implications for model selection. If word alignment accuracy is the primary concern, T1 is the better choice due to its higher average METEOR score. However, if fragmented alignments are considered relevant, the penalty parameter should be recalibrated to better fit the dataset's characteristics.

B. ChatGPT vs Gemini: Probing Statistical Significance (Q2)

Understanding the significance of performance differences between T1 and T2 goes beyond simple numerical comparisons. While METEOR scores and other metrics provide valuable data, statistical significance tests, such as the t-test, ensure that these differences aren't just random fluctuations. By conducting statistical significance tests, it becomes highly probable to confidently determine if one model consistently outperforms the other in translation tasks or other linguistic mechanisms.[20] This approach allows us to distinguish patterns, trends, and probabilities, providing a solid foundation for evaluating the model's overall effectiveness. It also highlights areas where each model may need further refinement and ensures that improvements are effective.[21] Like it or not, this rigorous analysis helps make informed decisions about model selection and deployment, responsibly optimizing the use of AI technology in the study of classical Arabic literature.[22] In this phase, the research delves into the t-test results to explore the statistical significance of the performance differences observed between T1 and T2. This follows the same principles and objectives as Ahmed et al.'s work but employs a different method.[23]

Table 5. Test of Equality of Variances (Brown-Forsythe)

	F	df_1	df_2	P
Total	0.282	1	28	0.600

The Brown-Forsythe test shown in table 5 assesses variance equality between groups, given that the data is not normally distributed. In Table 5, the F-value is 0.282, with degrees of freedom ($df_1 = 1$, $df_2 = 28$) and a p-value of 0.600. Since the p-value exceeds 0.05, the null hypothesis cannot be rejected, indicating no significant difference in variance. This implies homogeneous variance

[20] Philipp Koehn, Franz Josef Och, and Daniel Marcu, *"Statistical Phrase-Based Translation,"* Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL ' 03 1 (2003): 48–54. DOI: https://doi.org/10.3115/1073445.1073462

[21] Sui He, *"Prompting ChatGPT for Translation: A Comparative Analysis of Translation Brief and Persona Prompts,"* version 2, preprint, arXiv, 2024: 7-8. DOI: https://doi.org/10.48550/ARXIV.2403.00127; Nely Rahmawati Zaimah, Eko Budi Hartanto, and Fatchiatu Zahro, *"Acceptability and Effectiveness Analysis of Large Language Model-Based Artificial Intelligence Chatbot Among Arabic Learners,"* Mantiqu Tayr: Journal of Arabic Language 4, no. 1 (2024): 1. DOI: https://doi.org/10.25217/mantiqutayr.v4i1.3951

[22] Mohammad Awad AlAfnan, *"Artificial Intelligence and Language: Bridging Arabic and English with Technology,"* Journal of Ecohumanism 4, no. 1 (2025): 1. DOI: https://doi.org/10.62754/joe.v4i1.4961; Mahmoud Al-Ayyoub et al., "Deep Learning for Arabic NLP: A Survey," Journal of Computational Science 26 (2018): 522–31. DOI: https://doi.org/10.1016/j.jocs.2017.11.011

[23] Ahmed et al., *ChatGPT vs. Bard: A Comparative Study*, preprint (2023): 1-18. DOI: https://doi.org/10.36227/techrxiv.23536290.v1

across groups, supporting further statistical analysis that assumes equal variance. In this study, the non-parametric Mann-Whitney U test will be applied (Table 6).

Table 6. Independent Samples T-Test

					95% CI for Hodges-Lehmann Estimate	
	U	df	p	Hodges-Lehmann Estimate	Lower	Upper
Total	130.000		0.478	1.000	-3.000	8.000

Note. Mann-Whitney U test.

Table 6 explains everything about the performance of the Tested Models. The Mann-Whitney U test results indicate no statistically significant difference in performance between Model T1 (ChatGPT) and Model T2 (Gemini), with a U value of 130.000 and a p-value of 0.478. Since the p-value is well above the common significance threshold (α = 0.05), the null hypothesis cannot be rejected. It suggests that neither model consistently outperforms the other. The Hodges-Lehmann estimate of 1.000 implies a slight advantage for T1, but the wide confidence interval (-3.000 to 8.000) suggests high variability, making this difference negligible. Because the confidence interval includes zero, any observed differences may be due to random variation rather than actual superiority in language processing.

These findings align with METEOR score trends, where T1 generally performed better, though not to a statistically significant extent. This underscores the need for qualitative evaluation to capture nuances that quantitative metrics might overlook. Similarly, the comparative study by Farghal and Haedar remains consistent with this research, showing no significant differences.[24] T1 continues to outperform other models in terms of features and reliability, particularly for platforms that are widely recognized and used globally.

.

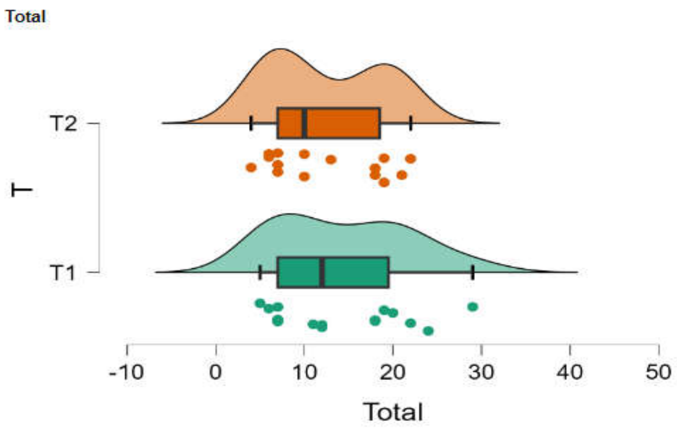

Figure 2. Raincloud Plots (descriptive)

[24] Mohammed Farghal and Ahmad S. Haider, *"Translating Classical Arabic Verse: Human Translation vs. AI Large Language Models (Gemini and ChatGPT),"* Cogent Social Sciences 10, no. 1 (December 2024): 2410998. DOI: https://doi.org/10.1080/23311886.2024.2410998

The raincloud plot, shown in figure 2, visually represents the distribution of scores for T1 (ChatGPT) and T2 (Gemini), combining the Density Plot (Cloud), Boxplot (Middle Section), and Scatter Plot (Rain) to provide a comprehensive view. The Density Plot (Cloud), represented by smooth kernel density estimation (KDE) curves above and below the boxplots, illustrates the overall distribution shape for each model. Both models exhibit similar median values, as seen in the Boxplot (Middle Section), which summarizes key statistical measures, including the median, interquartile range (IQR), and potential outliers. However, T1 appears to have a slightly wider spread, suggesting higher variance. The Scatter Plot (Rain) displays individual data points as dots, revealing the spread and clustering of values. T1 shows a longer tail toward lower values, indicating more dispersion, whereas T2 has a more concentrated distribution. These overlapping patterns align with the Mann-Whitney U test—surely reinforcing the need for deeper qualitative analysis beyond numerical metrics.

Both the visual and numerical data indicate that T1 outperforms T2 in several aspects, though overall, the models exhibit comparable performance. T1 shows greater variability in scores, while T2 maintains more stability in its distribution. These findings underscore the importance of qualitative evaluation to capture nuances that quantitative metrics might miss, especially for a more holistic context evaluation. Human-centered AI is a clear and undeniable principle.[25] Similarly, studies by Shahriar et al.,[26] and Ibrahim et al.,[27] emphasize the same point.

Closing

This study demonstrates that while generative AI models such as ChatGPT and Gemini show promise in handling surface-level translation tasks in classical Arabic texts, their performance remains constrained when confronted with deeper grammatical, cultural, and contextual complexities. Although ChatGPT exhibited slightly higher lexical alignment, statistical testing confirmed no significant performance difference between the two models, reinforcing findings from prior research that automated metrics alone are insufficient for evaluating linguistically dense and pedagogically sensitive texts. By integrating METEOR-based evaluation with expert assessment, this study contributes a more reliable and context-aware framework for assessing AI translation in advanced Arabic studies. The findings underscore the continued necessity of human scholarly validation and caution against overreliance on AI-generated outputs in academic and religious contexts. While the scope is limited to two evolving models, the study offers a replicable methodological foundation for future research, encouraging broader corpora, diversified metrics, and stronger expert involvement to advance responsible and verifiable AI-assisted scholarship.

[25] Simone Balloccu et al., *"Ask the Experts: Sourcing a High-Quality Nutrition Counseling Dataset through Human-AI Collaboration,"* Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, 2024, 11519–45. DOI: https://doi.org/10.18653/v1/2024.findings-emnlp.674; William J. Bingley et al., *"Where Is the Human in Human-Centered AI? Insights from Developer Priorities and User Experiences,"* Computers in Human Behavior 141 (April 2023): 107617. DOI: https://doi.org/10.1016/j.chb.2022.107617

[26] Sakib Shahriar et al., *"Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency,"* Applied Sciences 14, no. 17 (September 2024): 7782. DOI: https://doi.org/10.3390/app14177782

[27] Nourhan Ibrahim et al., *"A Survey on Augmenting Knowledge Graphs (KGs) with Large Language Models (LLMs): Models, Evaluation Metrics, Benchmarks, and Challenges,"* Discover Artificial Intelligence 4, no. 1 (November 2024): 76. DOI: https://doi.org/10.1007/s44163-024-00175-8

Acknowledgment

We extend our heartfelt gratitude to all the contributors and their distinguished colleagues for their support and contributions to this research. Additionally, we thank Al-Anwar Islamic College, Rembang, Indonesia, for their unwavering support and valuable reviews. Special thanks are extended to the professors, students, and all experts from Islamic boarding schools who dedicated their time and expertise to assist in this study. The authors would like to express their special gratitude to Chafidhoh Rizqiyah for providing substantial research assistance and technical support throughout the study; Syamsul Hadi for actively contributing to the drafting and structuring of the article; Rifatul Muthiah for strengthening the theoretical relevance and thoroughly revising both the original manuscript and the journal draft; and Wakhidati Nurrohmah Putri for her collaborative efforts in translating the manuscript into Arabic and English. Their contributions were invaluable to the completion and refinement of this work.

Bibliografi

Abdelhay, Mohammed, Ammar Mohammed, and Hesham A. Hefny. "*Deep Learning for Arabic Healthcare: MedicalBot.*" Social Network Analysis and Mining 13, no. 1 (April 2023): 71. DOI: https://doi.org/10.1007/s13278-023-01077-w

Abdulkader, Zena, and Yousra Al-Irhayim. "*A Review of Arabic Intelligent Chatbots: Developments and Challenges.*" Al-Rafidain Engineering Journal (AREJ) 27, no. 2 (September 2022): 178–89. DOI: https://doi.org/10.33899/rengj.2022.132550.1148

Ahmed, Imtiaz, Mashrafi Kajol, Uzma Hasan, and Partha Protim Datta. *ChatGPT vs. Bard: A Comparative Study*. Preprint. 2023. https://doi.org/10.36227/techrxiv.23536290.v1

AlAfnan, Mohammad Awad. "*Artificial Intelligence and Language: Bridging Arabic and English with Technology.*" Journal of Ecohumanism 4, no. 1 (2025): 1. DOI: https://doi.org/10.62754/joe.v4i1.4961

Al-Ayyoub, Mahmoud, Aya Nuseir, Kholoud Alsmearat, Yaser Jararweh, and Brij Gupta. "*Deep Learning for Arabic NLP: A Survey.*" Journal of Computational Science 26 (2018): 522–31. DOI: https://doi.org/10.1016/j.jocs.2017.11.011

Alshater, Muneer. "*Exploring the Role of Artificial Intelligence in Enhancing Academic Performance: A Case Study of ChatGPT.*" SSRN Electronic Journal, ahead of print, 2022. DOI: https://doi.org/10.2139/ssrn.4312358

Azmi, Aqil M, Abdulaziz O Al-Qabbany, and Amir Hussain. "*Computational and Natural Language Processing Based Studies of Hadith Literature: A Survey.*" Artificial Intelligence Review 52 (2019): 1369–414. DOI: https://doi.org/10.1007/s10462-019-09692-w

Balloccu, Simone, Ehud Reiter, Karen Jia-Hui Li, Rafael Sargsyan, Vivek Kumar, Diego Reforgiato, Daniele Riboni, and Ondrej Dusek. "*Ask the Experts: Sourcing a High-Quality Nutrition Counseling Dataset through Human-AI Collaboration.*" Findings of the Association for Computational Linguistics: EMNLP *2024*, Association for Computational Linguistics. 2024. 11519–45. DOI: https://doi.org/10.18653/v1/2024.findings-emnlp.674

Berkey, Jonathan Porter. *The Transmission of Knowledge in Medieval Cairo: A Social History of Islamic Education*. Princeton. New Jersey: Princeton University Press. 2014.

Bilquise, Ghazala, Samar Ibrahim, and Khaled Shaalan. "*Bilingual AI-Driven Chatbot for Academic Advising*." International Journal of Advanced Computer Science and Applications 13, no. 8 (2022): 34. DOI: https://doi.org/10.14569/IJACSA.2022.0130808

Bingley, William J., Caitlin Curtis, Steven Lockey, Alina Bialkowski, Nicole Gillespie, S. Alexander Haslam, Ryan K. L. Ko, Niklas Steffens, Janet Wiles, and Peter Worthy. "*Where Is the Human in Human-Centered AI? Insights from Developer Priorities and User Experiences*." Computers in Human Behavior 141 (April 2023): 107617. DOI: https://doi.org/10.1016/j.chb.2022.107617

Campello de Souza, Bruno, Agostinho Serrano de Andrade Neto, and Antonio Roazzi. "*Are the New AIs Smart Enough to Steal Your Job? IQ Scores for ChatGPT, Microsoft Bing, Google Bard and Quora Poe*." SSRN Scholarly Paper No. 4412505. Rochester. NY. April 7. 2023. DOI: https://doi.org/10.2139/ssrn.4412505

Carvalho, Lucila, Roberto Martinez-Maldonado, Yi-Shan Tsai, Lina Markauskaite, and Maarten De Laat. "*How Can We Design for Learning in an AI World?*" Computers and Education: Artificial Intelligence 3 (2022): 100053. DOI: https://doi.org/10.1016/j.caeai.2022.100053

Chauhan, Chhavi. "*The Impact of Generative Artificial Intelligence in Scientific Content Synthesis for Authors*." The American Journal of Pathology 194, no. 8 (August 2024): 1406–8. DOI: https://doi.org/10.1016/j.ajpath.2024.06.002

Dalayli, Feyza. "*Use of NLP Techniques in Translation by ChatGPT: Case Study*." In Proceedings of the Workshop on Computational Terminology in NLP and Translation Studies (ConTeNTS) Incorporating the 16th Workshop on Building and Using Comparable Corpora (BUCC), edited by Amal Haddad Haddad, Ayla Rigouts Terryn, Ruslan Mitkov, Reinhard Rapp, Pierre Zweigenbaum, and Serge Sharoff, 19–25. Varna, Bulgaria: INCOMA Ltd. Shoumen. Bulgaria. 2023. https://aclanthology.org/2023.contents-1.3

Farghal, Mohammed, and Ahmad S. Haider. "*Translating Classical Arabic Verse: Human Translation vs. AI Large Language Models (Gemini and ChatGPT)*." Cogent Social Sciences 10, no. 1 (December 2024): 2410998. DOI: https://doi.org/10.1080/23311886.2024.2410998

Hamadneh, Nawaf N., Samer Atawneh, Waqar A. Khan, Khaled A. Almejalli, and Adeeb Alhomoud. "*Using Artificial Intelligence to Predict Students' Academic Performance in Blended Learning*." Sustainability 14, no. 18 (January 2022): 18. DOI: https://doi.org/10.3390/su141811642

He, Sui. "*Prompting ChatGPT for Translation: A Comparative Analysis of Translation Brief and Persona Prompts*." Version 2. Preprint, arXiv. 2024. DOI: https://doi.org/10.48550/ARXIV.2403.00127

Hong, Matthew K, Adam Fourney, Derek DeBellis, and Saleema Amershi. "*Planning for Natural Language Failures with the Ai Playbook*." 2021. 1–11. DOI: https://doi.org/10.1145/3411764.3445735

Ibrahim, Nourhan, Samar Aboulela, Ahmed Ibrahim, and Rasha Kashef. "*A Survey on Augmenting Knowledge Graphs (KGs) with Large Language Models (LLMs): Models, Evaluation Metrics, Benchmarks, and Challenges.*" Discover Artificial Intelligence 4, no. 1 (November 2024): 76. DOI: https://doi.org/10.1007/s44163-024-00175-8

Inel, Oana, Tim Draws, and Lora Aroyo. "*Collect, Measure, Repeat: Reliability Factors for Responsible AI Data Collection.*" Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 11, no. 1 (November 2023): 1. DOI: https://doi.org/10.1609/hcomp.v11i1.27547

Johnson, Douglas, Rachel Goodman, J Patrinely, Cosby Stone, Eli Zimmerman, Rebecca Donald, Sam Chang, Sean Berkowitz, Avni Finn, and Eiman Jahangir. *Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model.* 2023. DOI: https://doi.org/10.21203/rs.3.rs-2566942/v1

Kataria, Pratik, Kiran Rode, Akshay Jain, Prachi Dwivedi, Sukhada Bhingarkar, and MCP India. "*User Adaptive Chatbot for Mitigating Depression.*" International Journal of Pure and Applied Mathematics 118, no. 16 (2018): 349–61. https://openreview.net/forum?id=r16Vyf-0-

Khoshafah, Faten. "*ChatGPT for Arabic-English Translation: Evaluating the Accuracy.*" Ministry of Education, Yemen, ahead of print. April 17. 2023. https://doi.org/10.21203/rs.3.rs-2814154/v2

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. "*Statistical Phrase-Based Translation.*" Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03 1 (2003): 48–54. DOI: https://doi.org/10.3115/1073445.1073462

Kruschke, John K. "*Bayesian Analysis Reporting Guidelines.*" Nature Human Behaviour 5, no. 10 (October 2021): 10. DOI: https://doi.org/10.1038/s41562-021-01177-7

Liang, Wei, and Hongsheng Dai. "*Bayesian Inference.*" In Quantum Chemistry in the Age of Machine Learning. 233–50. Elsevier, 2023. DOI: https://doi.org/10.1016/B978-0-323-90049-2.00005-6

Lozano, Michael, Stefan Winthrop, Cedric Goldsworthy, Artemas Leventis, and Alistair Birkenshaw. "*Semantic Depth Redistribution in Large Language Models to Contextual Embedding Preservation.*" Preprint, November 5. 2024. DOI: https://doi.org/10.22541/au.173083529.98863661/v1

Moneus, Ahmed Mohammed, and Yousef Sahari. "*Artificial Intelligence and Human Translation: A Contrastive Study Based on Legal Texts.*" Heliyon 10, no. 6 (March 2024): 55. DOI: https://doi.org/10.1016/j.heliyon.2024.e28106

Muthiah, Anisatun, and Luqman Zain. "*Konsep Ittishal Al-Sanad Sebagai Syarat Kajian Kitab Kuning Dalam Tradisi Pesantren An-Nahdliyyah Cirebon.*" Jurnal Studi Hadis Nusantara 2, no.1 (2020): 75. DOI: https://doi.org/10.24235/jshn.v2i1.6746

Raj, Harsh, Vipul Gupta, Domenic Rosati, and Subhabrata Majumdar. "*Semantic Consistency for Assuring Reliability of Large Language Models.*" arXiv:2308.09138. Preprint, arXiv, August 17, 2023. DOI: https://doi.org/10.48550/arXiv.2308.09138

Ras, Gabriëlle, Marcel Van Gerven, and Pim Haselager. "*Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges.*" In Explainable and Interpretable Models in Computer Vision and Machine Learning, edited by Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Baró, Yağmur Güçlütürk, Umut Güçlü, and Marcel Van Gerven, 19–36. The Springer Series on Challenges in Machine Learning. Cham: Springer International Publishing, 2018. DOI: https://doi.org/10.1007/978-3-319-98131-4_2

Riina, Nicholas, Likhitha Patlolla, Camilo Hernandez Joya, Roger Bautista, Melissa Olivar-Villanueva, and Anish Kumar. "*An Evaluation of English to Spanish Medical Translation by Large Language Models.*" In Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations), edited by Marianna Martindale, Janice Campbell, Konstantin Savenkov, and Shivali Goel, 222–36. Chicago, USA: Association for Machine Translation in the Americas. 2024. https://aclanthology.org/2024.amta-presentations.15/

Russell, Regina G, Laurie Lovett Novak, Mehool Patel, Kim V Garvey, Kelly Jean Thomas Craig, Gretchen P Jackson, Don Moore, and Bonnie M Miller. "*Competencies for the Use of Artificial Intelligence–Based Tools by Health Care Professionals.*" Academic Medicine 98, no. 3 (2023): 348–56. DOI: https://doi.org/10.1097/ACM.0000000000004963

Shahriar, Sakib, Brady D. Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool. "*Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency.*" Applied Sciences 14, no. 17 (September 2024): 7782. DOI: https://doi.org/10.3390/app14177782

Sulaeman, Islamiyah, Syuhadak Syuhadak, and Insyirah Sulaeman. "*ChatGPT as a New Frontier in Arabic Education Technology.*" Al-Arabi: Journal of Teaching Arabic as a Foreign Language 7, no. 1 (2023): 83–105. DOI: http://dx.doi.org/10.17977/um056v7i1p83-105

Vaswani, Ashish, Niki Parmar, Jakob Uszkoreit, Noam Shazeer, and Lukasz Kaiser. *Image Transformer*. February 15, 2018. https://openreview.net/forum?id=r16Vyf-0-

Wagenmakers, Eric-Jan, Richard D. Morey, and Michael D. Lee. "*Bayesian Benefits for the Pragmatic Researcher.*" Current Directions in Psychological Science 25, no. 3 (June 2016): 169–76. DOI: https://doi.org/10.1177/0963721416643289

Zaimah, Nely Rahmawati, Eko Budi Hartanto, and Fatchiatu Zahro. "*Acceptability and Effectiveness Analysis of Large Language Model-Based Artificial Intelligence Chatbot Among Arabic Learners.*" Mantiqu Tayr: Journal of Arabic Language 4, no. 1 (2024): 1. DOI: https://doi.org/10.25217/mantiqutayr.v4i1.3951