LISANIA: Journal of Arabic Education and Literature
P-ISSN: 2614-4425 E-ISSN: 2580-1716
Vol.9, No.1, 2025, pp.41-59
DOI: http://dx.doi.org/10.18326/lisania.v9i1.41-59

Assessing Copilot's Semantic Depth in Classical Arabic: A Mixed-Methods Evaluation Using *Alfiyah ibn Malik* and *Nadham Al-Imrithy*

Nely Rahmawati Zaimah
Sekolah Tinggi Agama Islam Al-Anwar Rembang, Indonesia
nelyrahmawati@staialanwar.ac.id

Syamsul Hadi
Sekolah Tinggi Agama Islam Al-Anwar Rembang, Indonesia
syamsulhadi@staialanwar.ac.id

Chafidloh Rizqiyah
Sekolah Tinggi Agama Islam Subang, Indonesia
rizqi.umiaqila@gmail.com

Risty Kamila Wening Estu
Sekolah Tinggi Agama Islam Al-Anwar Rembang, Indonesia
ristykamila@staialanwar.ac.id

Akhmad Roja Badrus Zaman
Albert-Ludwigs-Universität Freiburg, Germany
akhmad.zaman@email.uni-freiburg.de

ENGLISH ABSTRACT
It was rather surprising that Windows users readily embraced Copilot, even trusting it with translation projects. Surely, not many users would trust its accuracy in providing cross-language explanations for prompts solely based on the developer's claims. Building on that, this research aimed to test it in a manner distinct from other assessments. Researchers evaluated how accurately Copilot interpreted and understood the advanced Arabic prose from the intricate works of *Alfiyah ibn Malik* and *Nadham Al-Imrithy*. The aim was to understand Copilot's strengths and weaknesses in terms of literal accuracy, terminological-analogical mastery, and contextual depth. Using a mixed-method approach under the Collect-Measure-Repeat (CMR) framework of Responsible AI, the researchers conducted qualitative performance assessments with three experts and quantitative evaluations using METEOR (Metric for Evaluation of Translation with Explicit Ordering). The results showed that although Copilot had no issues comprehending and translating simple Arabic commands, especially word-for-word, it struggled with contextual understanding for many of the complex texts and displayed

numerous inconsistencies when the instructions were vague. Copilot's performance issues in context saturation were evident during iterative phases. This led to the conclusion that, while Copilot is competent enough to attempt the challenging task of interpreting complex linguistic structures, it still needs human assistance and cross-references.

Keywords: Copilot's Performance, Semantic Interpretations, Arabic Translations, METEOR Scores, AI's Contextual Depth

INDONESIAN ABSTRACT

Sangat mengejutkan bahwa pengguna Windows menerima Copilot begitu saja, bahkan mempercayainya untuk membantu dalam proyek-proyek terjemahan. Tentunya tidak banyak pengguna yang mencoba menguji sebarapa akurat penjelasan lintas-bahasanya ketika mengajukan permintaan (prompts). Atas dasar hal itu, penelitian ini bertujuan untuk mengevaluasi dengan cara yang berbeda dengan pengujian yang lain. Para peneliti mengevaluasi seberapa akurat Copilot menginterpretasi dan memahami bait-bait Arab yang sulit disadur dari kitab Alfiyah ibn Malik dan Nadham Al-Imrithy yang terkenal sangat rumit. Penelitian ini ditergetkan untuk memahami kekuatan dan kelemahan Copilot dalam hal akurasi literal, penguasaan terminologi dan analogi, serta kedalaman konteks. Menggunakan pendekatan mix-method dalam prinsip Responsible AI (RAI) di bawah kerangka Collect-Measure-Repeat (CMR) dengan pengukuran kinerja secara kualitatif dari tiga pakar sebagai tolok ukur, dan penilaian kuantitatif menggunakan METEOR (Metric for Evaluation of Translation with Explicit ORdering). Hasilnya menunjukkan bahwa meskipun Copilot tidak memiliki masalah dalam memahami dan menerjemahkan perintah Arab yang sederhana—terutama kata per kata—ia kesulitan dengan pemahaman kontekstual untuk banyak teks yang rumit dan menunjukkan banyak inkonsistensi ketika instruksi tidak jelas. Masalah kinerja Copilot dalam saturasi konteks terlihat selama fase iteratif. Ini mengarah pada kesimpulan bahwa, meskipun Copilot cukup kompeten untuk mencoba tugas menantang dengan menginterpretasikan struktur linguistik Arab yang kompleks, ia tetap masih membutuhkan bantuan manusia dan referensi silang.

***Kata Kunci**: Performa Copilot, Interpretasi Semantik, Terjemahan Bahasa Arab, METEOR Skor, Pemahaman Konteks AI*

Introduction

The application of Generative AI, particularly transformer-based models, remains limited in delivering user responses due to their computational architecture and inherent constraints in handling parameters or protocols for supporting augmented data analysis and valid diagnostics (Ahmed et al., 2023; Gemini Team et al., 2024). Many transformer models have shown resilience in addressing these challenges, with initial concerns arising in exact science fields such as STEM, healthcare, and many others (Carvalho et al., 2022; Johnson et al., 2023). At the same time, AI raises new ethical considerations, boundaries, and critical concerns in analyzing scientific content and

contexts (Perkins, 2023; Ras et al., 2018). In popular perception, Microsoft Copilot, a built-in feature of the Microsoft operating system integrated into Windows OS devices and similar platforms, has become a familiar and user-friendly instant assistant for users worldwide (Stratton, 2024). For most users, it's simply another tool that makes everything easier, including cross-language translation projects. But convenience doesn't always mean accuracy, especially when it comes to complex languages like Arabic. So, who can truly guarantee the precision of these translations? In this study, researchers systematically tested Microsoft Copilot and comprehensively analysed its accuracy using human experts' evaluation and METEOR (Metric for Evaluation of Translation with Explicit Ordering) scoring.

Certainly, the researchers selected verses from the works of *Alfiyah ibn Malik* and *Nadham Al-Imrithy,* as these texts are not only canonical in the Arabic grammatical tradition but are also highly regarded for their intricate and nuanced treatment of syntax, morphology, and semantics within a poetic framework (Fodhil & Hanifah, 2022; Inas, 2024). For centuries, scholars and students have approached these texts under the guidance of trained experts, as the interpretive process often involves navigating multiple layers of meaning (Berkey, 2014; Muthiah & Zain, 2020). The authenticity of the narrative diction (*matn*) and the intricate details of these Arabic texts must be principled and aligned with precise AI augmentation and accurate interpretations. Additionally, the cultural sensitivities embedded in these texts require that AI-generated interpretations avoid unintended distortions or misrepresentations of their context or meaning (Lozano et al., 2024; Sulaeman et al., 2023).

The precision of literal and contextual analysis in Arabic is vital (Alruqi & Alzahrani, 2023; Hidayatullah & Fauji, 2023). It forms the basis for assessing the reliability and validity of generative AI language models like Copilot, one of the most popular generative AI models globally. Proper citation and attribution of AI-augmented content reflect respect for scholarly traditions and are equally important in maintaining academic integrity (Bilquise et al., 2022; Russell et al., 2023). The broader accessibility provided by Generative AI offers opportunities but also raises critical questions: Firstly (RQ1), what are experts' evaluations of Copilot's performance in interpreting complex Arabic texts? Secondly (RQ2), how is the accuracy of its interpretations measured using

the METEOR (Metric for Evaluation of Translation with Explicit Ordering) metric regarding literal, analogical, and contextual dimensions?

In evaluating machine translation for classical Arabic texts such as *Alfiyah ibn Malik* and *Nadham Al-Imrithy*, it is important to combine metrics like METEOR with the human experts' evaluations for measuring semantic accuracy. Early research by Lavie and Agarwal (2007) established a solid theoretical foundation by incorporating lexical matching, synonyms, and penalties for word order differences to achieve evaluations that closely resemble human judgments, which are also included in the assessment criteria. This work was later expanded by Denkowski and Lavie (2014), which adapted METEOR for various languages—including those with agglutinative structures such as Arabic—by integrating morphology and lemmatisation processes. This strong theoretical base, combined with cross-linguistic adaptations, has provided a robust framework for assessing translation quality, especially for texts rich in linguistic nuances, just as envisioned and schemed by Dahia & Belbacha (2024), He (2024), and Zhang (2024) on ChatGPT. By merging METEOR with expert benchmarking and these new techniques, the performance evaluation of systems like Microsoft Copilot on test texts from *Alfiyah ibn Malik* and *Nadham Al-Imrithy* can be carried out more comprehensively.

Methods

This mixed-methods research adopts the Collect-Measure-Repeat (CMR) framework model developed by Inel et al. (2023), alongside insights from Raj et al. (2023), to provide a comprehensive evaluation of semantic accuracy. Even though the validity benchmarks from Inel and colleagues, like other AI validity tests, focus on input datasets, this research will attempt to measure the opposite: AI augmentation (output). This research, along with all reviewers' qualitative responses, will assume the same standard, where a correct response is marked with a score of 1 and an incorrect one is marked with a score of 0. Quantitatively, this study also integrates the METEOR measurement system—selected for its proven correlation with human evaluation and its sensitivity to paraphrases and word-order variations—to assess lexical matching, synonym use, and penalties for word-order differences, thereby complementing expert judgment with an objective, reproducible metric. Similar to the approach by Hameed et

al. (2022), who measured performance for Russian-Arabic translations, METEOR has been further developed for various languages through the integration of morphological and lemmatisation processes. This aids in evaluating the quality of interpretations from Arabic to Indonesian.

Copilot is the sole model evaluated in this study. Ten queries (Q1 to Q10) were formulated to probe specific aspects of Arabic grammar 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)	Benchmark
1	فَارْفَعْ بِضَمَ وَانْصِبَنْ فَتْحَاً وَجُرْ ۞ كَسْـــرَاً كَـذِكْرُ اللَّهِ 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 نَغْتَبِطْ إذْ تُشْكَرُ	Concealed Pronouns (*Ism Dhomir Mustatir*)	"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 وَأَسَدْ ۞ وَذُو ارْتِجَال كَسُعَـــادَ 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 وصفٌ فضلةٌ منتصب ۞ مفهم في حال كفردا ً أذهب	حال (Adverbial Modifier)	"The حال is a descriptive word, surplus, منصوب (accusative), indicating a state, like 'I went alone.'"
5	وَتِلْوَ أَفْعَلَ انْصِبَنَّهُ كما ۞ أَوْفَى خَلِيلَيْنَا وَأَصْدَقْ بِهِمَا	صيغة التعجب (Expressions 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	وَلَفْـظُهُ الْمَشْهـُوْرُ فِيْهِ اَرْبَعُ ۞ نَفْسٌ وَعَيْنٌ ثُمَّ كُـلٌّ اَجْـمَـعُ	توكيد (Emphasis) - Part 1	"And its well-known words are four: نفس (self), عين (eye), then كل (all) جمع (collective)."
7	وَغَـيْرُهَا تَـــوَابِـعٌ لِأَجْمـَعَ ۞ مِـــنْ اَكْتَـــعَ وَاَبْتَـــعَ وَاَبْـصَـعَ	توكيد (Emphasis) - Part 2	"And others are followers for all, from أكتع, أبتع, and أبصع."
8	هُوَ اسْمُ وَقْتٍ اوَمَكَانٍ نِ انْتَصَبْ ۞ كُلٌّ عَلَى تَقْدِيْرِ فِى عِنْدَ الْعَرَبْ	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 الْإِسْتِثْنَا الَّذِي لَهُ حَوَى ۞ اِلاَّ وَغَـيْرُ وَسِـــوًى سُوًى سَوَا	استثناء (Exception)	"And the words of exception that it contains are: إلاّ (except), غير (other than), سوى (besides), سُوى (equivalent), سواءً (equal)."
10	وَهُوَ عَلَى تَقْدِيرِفِي أَوْ لاَمِ ۞ أَوْ مِنْ كَمَكْرِ اللَّيْلِ أَو غَلاَمِي	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.'"

Each query was repeated five times, packaged into 15 different question models (QMs): five literal translations (L), five terminological explanations (T), and five contextual adaptability (C). The resulting responses were then analysed as variables for accuracy and consistency. Three expert assessors specialising in Arabic language and grammar evaluated the responses, scoring them from 1 to 5 after five task regenerations, culminating in a total of 750 outputs.

As the final part of the evaluation method, this study integrates a specialised lexical database to enhance the accuracy of machine translation quality measurements from Arabic to Indonesian. Without diminishing the value of expert evaluations, the METEOR score will be used as a supporting (or opposing) perspective to the initial qualitative assessment results. In this approach, in addition to applying the previously explained METEOR formula, where the F-Mean score is calculated using the following equation:

$$F_{mean} = \frac{10 \times P \times R}{R + 9P}$$

The fragmentation penalty is calculated using the formula:

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

The Final score is calculated using the following formula:

$$\text{Final Score} = F_{mean} \times (1 \times Penalty)$$

The researchers used the pymeteor Python library to evaluate Copilot's translations. For each translated sentence, the METEOR function compares Copilot's output against human references, applying stemming, synonym matching, and paraphrase recognition to assess accuracy. To ensure consistency, this process is repeated five times, though METEOR itself provides stable scores in a single run.

Unlike basic word-for-word comparisons, METEOR incorporates lexical resources such as WordNet and multilingual databases, enabling it to recognize synonyms, domain-specific terms, and contextual variations. By balancing precision and recall through a harmonic mean, applying a fragmentation penalty to discourage disjoint word order, and leveraging rich lexical matching techniques, METEOR provides a comprehensive evaluation of translation quality. This is particularly valuable for assessing idioms, cultural expressions, and nuanced meanings that go beyond literal

translation. Pymeteor code simulation will be provided in the appendix page for reference.

Copilot's Interpretations and Experts' Opinions (RQ1)

The testing results of Copilot generally showed literal translations with some variations in emphasis and explanation. Researchers verified each translation and the patterns generated by the model using translations of *Alfiyah ibn Malik* and its commentaries compiled by Fuad (2010) and Haq (2022). These translations demonstrated an understanding of Arabic linguistic constructions, particularly by identifying letters and harakat (diacritical marks) as shown in queries Q2, Q3, and Q7. However, they lacked the additional elaborations seen in earlier iterations. Notably, a single repetition in query 3 was responded to correctly. Despite these shortcomings, the translations conveyed the meaning of individual words within the verses.

When evaluating the overall sentence structure, it proved necessary to vary the prompts—such as requesting a complete verse translation and an explanation—due to inconsistent quality in word-for-word translations. Although some repetitions of queries were answered correctly, there were also several instances of errors. For the 11th Question Model (QM), Q3, Q4, Q6, Q7, Q9, and Q10 were entirely incorrect in all repetitions. While the vocabulary was accurate, the combined sentences often deviated from the intended context. However, translating the Arabic language goes beyond mere word-for-word translation. It requires understanding the correlation between words, grammar, and the nuances of context to build an accurate interpretation (Anwar et al., 2023).

When asked for a detailed explanation, Copilot provided contextually appropriate responses with occasional errors in meaning; nevertheless, the reviewers still considered it a mistake. At this stage, the accuracy of word-for-word translation was very good, with only minor errors in translation and contextual alignment when no specific task instructions were provided. Researchers noted some biases and errors in examples, citing occasional misinterpretations of word structures that led to translation errors in Q4, Q7, and Q9 in Question Model 9. The issue arose when the prompt was regenerated or repeated, producing contextually different and inconsistent results in the development of contextual examples across several iterations.

In the explanation of four verses (Q3-Q6 with simple Question Model 12), Copilot attempted to provide a comprehensive and systematic understanding of their meaning and grammatical context in Arabic. Copilot detailed each word and concept presented in the verses, explaining their meanings as well as their interactions within the sentences (word-to-word). Although Copilot's explanations tended to be literal, it occasionally offered practical examples to give readers a clearer understanding of the intended message. The answers focused on analyzing the verses structurally and semantically, proving that each element was well understood. However, this response is ultimately an answer to a simple question.

Feedback from experts indicates that AI translation still faces considerable limitations in fully comprehending advanced Arabic texts, such as the rich literary works of *Alfiyah ibn Malik* and Nadham al-Imrithi. All reviewers argue that the depth and intricacy of classical Arabic literature often remain challenging for Copilot to grasp completely. The translations produced tend to be surface-level, unable to capture the deeper nuances and meanings embedded within these texts.

Table 2. Distribution of Prompted Responses

Terms	Question Model	Iteration	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Literal Translation (L)	1	5x	2	2	1	2	3	2	2	3	4	4
	2	5x	2	2	2	3	2	4	1	3	2	2
	3	5x	1	0	0	1	2	1	0	2	1	2
	4	5x	2	2	0	0	2	0	0	3	0	0
	5	5x	1	1	2	3	1	3	1	2	3	1
Analogical Interpretation (T)	6	5x	2	3	1	2	2	1	1	3	3	4
	7	5x	1	1	1	1	1	1	1	2	1	4
	8	5x	1	1	2	2	2	3	1	2	2	2
	9	5x	1	1	2	0	0	1	0	1	0	1
	10	5x	2	1	3	2	2	2	0	3	1	2
Contextual Adaptability (C)	11	5x	1	1	0	0	2	0	0	2	0	0
	12	5x	3	3	4	4	4	4	1	4	1	1
	13	5x	1	2	2	0	2	2	1	3	0	0
	14	5x	1	1	0	1	3	2	2	3	1	1
	15	5x	1	1	0	0	1	1	0	1	0	0

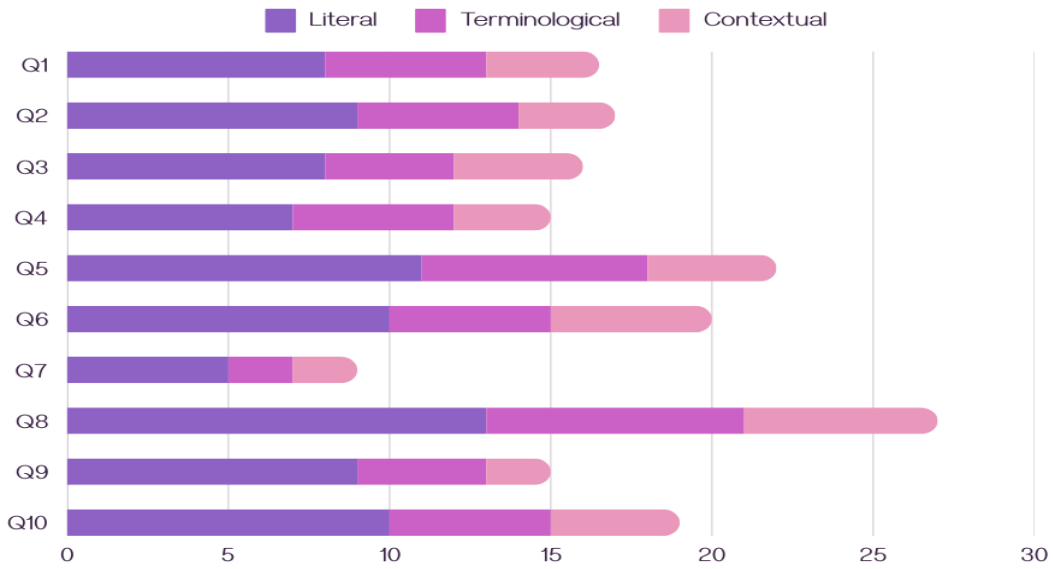
Source: Final scores from the experts

The distribution table (Table 2) reflects the reviewers' ratings of Copilot's responses across ten queries (Q1 to Q10) concerning three key semantic examination

methods (Terms). Scores were assigned based on the proximity or accuracy of the response to a predefined benchmark. If the repeated results are augmented differently, especially if they deviate from the context outlined by the benchmark, they are deemed incorrect (0). A score of 1 indicates a correct or contextually appropriate response, while a higher score reflects consistency across multiple iterations.

The scores from expert reviewers indicate that none of the ratings reached 30%, with some falling below 10%. Even upon repetition, the results varied significantly and still failed to achieve correct contextual answers. According to the reviewers, *Alfiyah ibn Malik* and *Nadham Al-Imrithy* cannot be comprehended in their contextual depth by Copilot. The figure illustrates the scores after multiple iterations for ten different questions (Q1 to Q10). Each bar represents the score for a specific question, with the y-axis showing the percentage score ranging from 0.00% to 30.00%. The highest score is for Q8, which is slightly above 25%. This is due to the simpler grammatical structure of the verse compared to other verses in this test. Unfortunately, the lowest score is for Q7, which is below 5%. This visualization highlights the variability and generally low performance of the scores, emphasizing the difficulty in achieving high accuracy with complex texts.

Figure 1. Reviewer's Score of Copilot's Performance



Moreover, experts emphasised the importance of cross-referencing translations with additional sources, such as commentaries and explanatory texts. They also highlighted the critical role of human expertise in studying classical Arabic literature.

While AI technology provides convenience in translating and analyzing classical texts, the involvement of experts with deep knowledge of classical Arabic remains indispensable. These experts can offer cultural context, interpretive depth, and richer insights into the texts. Collaboration between AI, cross-referenced literature, and human expertise is essential for developing a well-rounded and verified understanding of classical Arabic literature.

Therefore, experts strongly advocate for an integrated approach that combines AI technology, cross-referenced literature, and human expertise. Through this collaboration, AI's limitations can be addressed, the scope of research expanded, and a deeper, more validated understanding of classical Arabic literary heritage achieved.

Accuracy and Consistency in Numbers (RQ2)

Building on qualitative insights, this research advances to a structured quantitative analysis to expand previous findings and enrich the existing discourse. The primary objective of this phase is to systematically evaluate the reviewers' assessment records, consisting of scores organised in rows and columns of various variables. This approach is expected to provide a deeper understanding of the model's capabilities in handling complex linguistic constructs and its limitations, while also examining consistency and interrelatedness.

Drawn from verses of two classical Arabic texts (*Alfiyah ibn Malik* and *Nadham Al-Imrithy*), ten meticulously formulated queries (Q1 to Q10) were designed to address critical aspects of Arabic grammar with varying degrees of contextual depth. These queries are designed to test challenges such as literal translations (L), terminological interpretations (T), and contextual adaptability (C). For each query, benchmark responses were established to define the expected answers and to explore how they might be expanded into contextualised examples. Each query was then crafted with 5 different prompt models (Question Models/QMs), with each model regenerated 5 times to ensure repetition and measure the accuracy or proximity to the accuracy of the repeated responses. Based on those schemes, the outputs from each prompt model have been collected and evaluated using the METEOR metric to measure the translation's alignment with the established benchmarks. This measurement encompasses lexical accuracy, terminology usage, and contextual adaptation capabilities, enabling a thorough

assessment of the translation system's performance. By repeating the regeneration process five times per model, representative data have been obtained to test the consistency of the output. The following table presents the METEOR scores for each question (Q1 to Q10) in the literal translation category, which serves as a key parameter in this evaluation analysis.

Table 3. Meteor Scores on Copilot's Literal Translations

Q	Precision (P)	Recall (R)	F-Mean (α=0.9)	Penalty Frag.	Final Score
Q1	0.840	0.744	0.753	0.032	0.728
Q2	0.840	0.753	0.761	0.032	0.736
Q3	0.853	0.760	0.768	0.032	0.743
Q4	0.813	0.725	0.733	0.032	0.709
Q5	0.836	0.740	0.749	0.032	0.725
Q6	0.849	0.749	0.758	0.032	0.733
Q7	0.857	0.766	0.775	0.032	0.749
Q8	0.845	0.744	0.754	0.032	0.729
Q9	0.847	0.740	0.751	0.032	0.726
Q10	0.864	0.762	0.778	0.032	0.752

Table 3 shows that Copilot's literal translations are generally strong. Its precision values, ranging from 0.813 to 0.864, indicate that the words it picks are mostly correct when compared to the reference texts. Although recall values are slightly lower, between 0.725 and 0.766, this suggests that some words from the reference might be missing in the translation. The F-mean scores, which combine both precision and recall with a bit more weight on recall, range from 0.733 to 0.778, reflecting a good balance between accuracy and completeness. The constant penalty for fragmented word order of 0.032 implies that the matched words are mostly kept in a coherent order, so there isn't much disruption in the sentence structure. After accounting for this penalty, the final scores range from 0.709 to 0.752, showing that overall, Copilot produces translations that are not only accurate but also maintain a well-ordered structure.

Table 4. Meteor Scores on Copilot's Analogical and Terminological Performance

Q	Precision (P)	Recall (R)	F-Mean (α=0.9)	Penalty Frag.	Final Score
Q1	0.680	0.620	0.625	0.032	0.605
Q2	0.720	0.650	0.682	0.032	0.660
Q3	0.690	0.630	0.657	0.032	0.636
Q4	0.650	0.580	0.608	0.032	0.589
Q5	0.710	0.640	0.673	0.032	0.652
Q6	0.670	0.610	0.636	0.032	0.616
Q7	0.730	0.660	0.692	0.032	0.670
Q8	0.640	0.570	0.600	0.032	0.581
Q9	0.680	0.620	0.625	0.032	0.605
Q10	0.700	0.630	0.662	0.032	0.641

Table 4 demonstrates that Copilot's performance on analogical and terminological tasks is only moderate. Precision scores range from approximately 0.640 to 0.730, indicating that while the system sometimes selects the correct words, it does not always do so perfectly. Recall scores, which fall between 0.570 and 0.660, suggest that Copilot often misses some crucial details from the reference texts. The F-mean scores, which balance precision and recall with a slight emphasis on recall, range from 0.600 to 0.692. After applying a small fragmentation penalty of 0.032 to account for minor disruptions in word order, the final scores range from 0.581 to 0.670.

This overall performance highlights several key issues: basic errors in analogies, inconsistent terminology usage, and ambiguous or unclear analogies. Simply put, Copilot struggles to clearly and accurately convey analogical relationships and consistently uses the correct terminology. Researchers consider this evaluation overly rigid, as exemplifying or analogising sentences cannot be entirely quantified with numbers. Although the structural order of the output is fairly maintained, these deficiencies in understanding and conveying deeper meaning limit the system's effectiveness in handling more complex analogical and terminological content.

Table 5. METEOR Scores on Copilot Contextual Depth

Q	Precision (P)	Recall (R)	F-Mean (α=0.9)	Penalty Frag.	Final Score
Q1	0.780	0.690	0.726	0.032	0.702
Q2	0.791	0.710	0.743	0.032	0.719
Q3	0.770	0.680	0.717	0.032	0.693
Q4	0.750	0.660	0.694	0.032	0.671
Q5	0.802	0.720	0.752	0.032	0.727
Q6	0.761	0.670	0.706	0.032	0.683
Q7	0.810	0.730	0.763	0.032	0.738
Q8	0.741	0.651	0.685	0.032	0.663
Q9	0.770	0.700	0.727	0.032	0.703
Q10	0.822	0.740	0.772	0.032	0.746

The METEOR scores for Copilot's contextual depth (Table 5) show a moderate performance, with final scores ranging from 0.671 to 0.746. While the system achieves decent Precision (between 0.741 and 0.822) and Recall (from 0.651 to 0.740), it still struggles to fully capture the nuanced meaning of the text. One significant weakness is its tendency to ignore implicit details, leading to an overly simplified understanding of the context. For instance, when personal names such as "*fadl*" appear in Q3, Copilot sometimes misinterprets them and occasionally turns these names into unrelated terms that do not fit the intended context. This misinterpretation highlights a major flaw in its ability to handle context-specific information correctly. Moreover, the system often fails to grasp subtle elements like irony or metaphorical language, resulting in translations that lack depth and accuracy. Although the low fragmentation penalty indicates that the correct order of words is generally maintained, the issues with oversimplification and misinterpretation reduce the overall quality of the translation. In summary, while Copilot can perform adequately in clear, straightforward contexts, its inability to accurately interpret more complex or nuanced content calls for further improvements in its contextual understanding capabilities.

Figure 2. METEOR Scores of Copilot's Performance

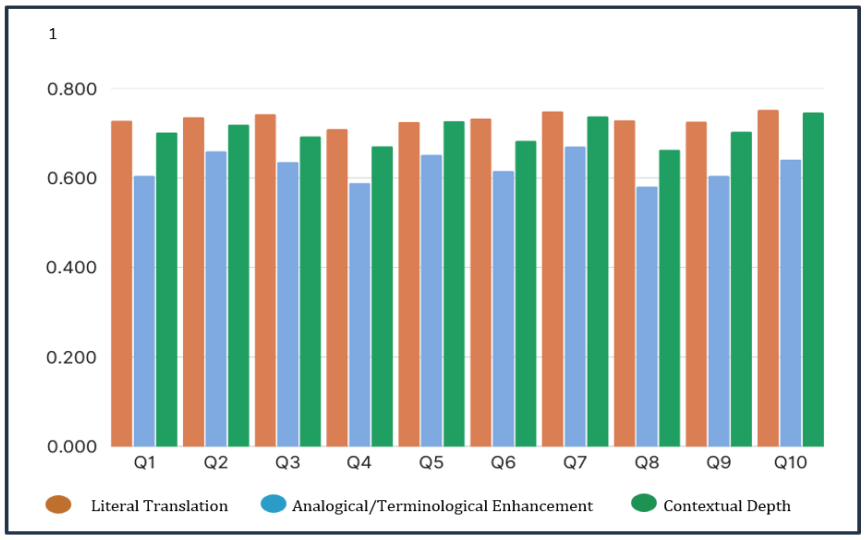


Figure 2 shows the METEOR scores achieved by Copilot for each of the ten questions (Q1 to Q10), represented by three bars across three categories. This pattern demonstrates the extent of Copilot's performance in handling various linguistic demands. Overall, the literal category tends to yield moderate to high scores, reflecting Copilot's relative ease in matching words on a one-to-one basis. Meanwhile, the terminological category often exhibits lower scores, due to specialised vocabulary or specific terminology posing greater challenges, especially when exact word matching or more nuanced definitions are required, or when Copilot, unprompted, provides erroneous example sentences. In contrast, the contextual category ranges from moderate to fairly high. It indicates that although Copilot can capture the overall meaning and maintain coherent word order, it still struggles considerably with deeper inference or implicit nuances such as figurative language or extended references (particularly when continuity across stanzas is needed but not explicitly requested).

A closer look at certain questions reveals that items like Q4 or Q8 display lower performance in some categories, pointing to structural complexity or specific terminology that Copilot fails to map accurately. Although this slightly differs from human experts' findings, items such as Q7 or Q10 show better alignment, possibly due to simpler phrasing or more direct contextual cues from the reference database. Overall, the figure highlights the importance of balancing literal accuracy, correct use of terminology, and broader contextual understanding. It also pinpoints areas where Copilot excels (for instance, straightforward literal translations) as well as those needing

further enhancement (such as advanced terminology handling, analogies and examples, and deeper contextual interpretation).

Overall, this study demonstrates that while Copilot excels in maintaining structural consistency and lexical accuracy in literal translations, it still faces significant challenges in capturing more complex linguistic nuances, such as contextual interpretation, relevant references, and the consistent delivery of specialised terminology. Copilot is capable of producing acceptable translations for simple texts, but its performance diminishes when it must process implicit contexts or deeper analogies, as Esfandiari & Allaf-Akbary (2024) have noted. These findings align with previous research on AI models like ChatGPT, Gemini, and other transformer-based translation systems, which similarly excel at basic tasks yet exhibit limitations when confronted with multi-interpretative complexities (Farghal & Haider, 2024; Khoshafah, 2023). A more holistic approach—including enhancing domain-specific training data and developing algorithms capable of capturing subtle nuances and deeper context—is increasingly deemed a critical need in current literature (Chaturvedi et al., 2024; Olsher, 2014). Consequently, this evaluation not only reaffirms the potential of AI technology (in this case, Microsoft Copilot) in processing classical texts but also highlights key areas for improvement to achieve a more comprehensive and accurate understanding. These results make a significant contribution to our overall understanding of the capabilities and limitations of AI in handling complex linguistic contexts, resonating with trends and extending the findings of previous studies.

For Windows users who have already placed their trust in Copilot for translation projects, these findings highlight the need to recognise both the strengths and limitations of today's AI technology. While Copilot performs well with literal translations—keeping the structure consistent and the words accurate—it still struggles to capture detailed contextual nuances, deliver specialised terminology consistently, and correctly interpret subtle, implicit references. These issues echo earlier research on AI, which stresses the importance of being critical of AI outputs, even though AI can simplify many tasks. Given these shortcomings, it is vital to blend cross-disciplinary literature reviews with human evaluations in the assessment process. Although human judgment is subjective, it offers essential contextual insights and sensitivity to subtle meanings that automated metrics often miss. As Balloccu et al. (2024) and Chen et al. (2023)

pointed out, it's crucial to keep objectively assessing all linguistic elements, whether through human evaluations or metrics, because AI progress must align with its credibility. This integrated approach not only provides a fuller understanding of translation quality but also guides improvements in training data and algorithm refinement. Ultimately, this dual framework is crucial for enhancing the accuracy of translation systems and ensuring that AI-generated translations are both contextually appropriate and dependable in real-world applications.

Conclusion

Copilot, which holds a special place popular among Windows users, indeed demonstrates strong ability in handling literal translations, consistently matching structure and word accuracy directly. However, when faced with complex classical Arabic texts, particularly those requiring specialised terminology and nuanced contextual interpretation, the system shows significant limitations. The METEOR scores reveal that literal translations score higher than those in the terminological and contextual categories. This suggests that while Copilot can produce adequate translations for straightforward sentences, it tends to struggle with capturing implicit nuances, subtle references, and deeper analogies. Expert assessments further indicate that, although the basic structure is maintained, there are errors in conveying the correct context and inconsistencies in term selection. This underscores the need for a comprehensive evaluation approach that combines quantitative metrics, like METEOR, with qualitative expert feedback, acknowledging the inherently subjective nature of human evaluation. These findings echo earlier research on AI models such as ChatGPT and Gemini, which excel at simple tasks yet encounter challenges in more complex, multi-interpretative contexts.

The researchers recommend enhancing training data by incorporating a specialised corpus of classical Arabic texts and refining the algorithms to better detect contextual nuances and handle specialised terminology. A hybrid approach that integrates quantitative metric evaluation with qualitative expert feedback is also strongly advised for future studies, as this combined method is essential to significantly improve the overall quality of AI translations.

Acknowledgment

A heartfelt thank you to the diligent efforts of the authors, whose hard work made this research possible. We extend our deep gratitude to Al-Anwar Islamic College for their unwavering support and valuable reviews. This research was funded by a Research Grant Programme from the Research and Community Service Centre (P3M), Al-Anwar Islamic College, Rembang, Indonesia. Special thanks to the professors and students who contributed their time and expertise to assist in this study.

References

Ahmed, I., Kajol, M., Hasan, U., & Datta, P. P. (2023). *ChatGPT vs. Bard: A Comparative Study* [Preprint]. https://doi.org/10.36227/techrxiv.23536290.v1

Alruqi, T. N., & Alzahrani, S. M. (2023). Evaluation of an Arabic Chatbot Based on Extractive Question-Answering Transfer Learning and Language Transformers. *AI*, *4*(3), 667–691. https://doi.org/10.3390/ai4030035

Anwar, S., Kesuma, G. C., & Koderi. (2023). Development of al-Qawaid an-Nahwiyah Learning Module Based on Qiyasiyah Method for Arabic Language Education Department Students | Pengembangan Modul Pembelajaran al-Qawaid an-Nahwiyah Berbasis Metode Qiyasiyah untuk Mahasiswa Jurusan Pendidikan Bahasa Arab. *Mantiqu Tayr: Journal of Arabic Language*, *3*(1), Article 1. https://doi.org/10.25217/mantiqutayr.v3i1.2830

Balloccu, S., Reiter, E., Li, K. J.-H., Sargsyan, R., Kumar, V., Reforgiato, D., Riboni, D., & Dusek, O. (2024). Ask the experts: Sourcing a high-quality nutrition counseling dataset through Human-AI collaboration. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 11519–11545. https://doi.org/10.18653/v1/2024.findings-emnlp.674

Berkey, J. P. (2014). *The Transmission of Knowledge in Medieval Cairo: A Social History of Islamic Education*. Princeton University Press.

Bilquise, G., Ibrahim, S., & Shaalan, K. (2022). Bilingual AI-Driven Chatbot for Academic Advising. *International Journal of Advanced Computer Science and Applications*, *13*(8). https://doi.org/10.14569/IJACSA.2022.0130808

Carvalho, L., Martinez-Maldonado, R., Tsai, Y.-S., Markauskaite, L., & De Laat, M. (2022). How can we design for learning in an AI world? *Computers and Education: Artificial Intelligence*, *3*, 100053. https://doi.org/10.1016/j.caeai.2022.100053

Chaturvedi, S., Thakur, A., & Srivastava, P. (2024). Refining Language Translator Using In-depth Machine Learning Algorithms. *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 1–6. https://doi.org/10.1109/ICRITO61523.2024.10522202

Chen, Y., Clayton, E. W., Novak, L. L., Anders, S., & Malin, B. (2023). Human-Centered Design to Address Biases in Artificial Intelligence. *Journal of Medical Internet Research*, *25*(1), e43251. https://doi.org/10.2196/43251

Dahia, I., & Belbacha, M. (2024). Machine-Learning-based English Quranic Translation: An Evaluation of ChatGPT. *International Journal of Linguistics, Literature and Translation*, *7*(8), 128–136. https://doi.org/10.32996/ijllt.2024.7.8.17

Denkowski, M., & Lavie, A. (2014). Meteor Universal: Language-Specific Translation Evaluation for Any Target Language. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 376–380. https://doi.org/10.3115/v1/W14-3348

Esfandiari, R., & Allaf-Akbary, O. (2024). Assessing interactional metadiscourse in EFL writing through intelligent data-driven learning: The Microsoft Copilot in the spotlight. *Language Testing in Asia*, *14*(1), 51. https://doi.org/10.1186/s40468-024-00326-9

Farghal, M., & Haider, A. S. (2024). Translating classical Arabic verse: Human translation vs. AI large language models (Gemini and ChatGPT). *Cogent Social Sciences*, *10*(1), 2410998. https://doi.org/10.1080/23311886.2024.2410998

Fodhil, M., & Hanifah, S. (2022). Analysis of The Values of Moral Education in Nadzam Imrithy by Sheikh Syarafuddin Yahya Al-Imrithy. *SCHOOLAR: Social and Literature Study in Education*, *2*(1), Article 1. https://doi.org/10.32764/schoolar.v2i1.1477

Fuad, B. (2010). *Terjemah Alfiyah Ibnu Malik dan Penjelasannya*. Mobile Santri.

Gemini Team, Reid, M., Savinov, N., Teplyashin, D., Dmitry, Lepikhin, Lillicrap, T., Alayrac, J., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., Antonoglou, I., Anil, R., Borgeaud, S., Dai, A., Millican, K., Dyer, E., Glaese, M., … Vinyals, O. (2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context* (arXiv:2403.05530). arXiv. https://doi.org/10.48550/arXiv.2403.05530

Hameed, D. A., Faisal, T. A., Alshaykha, A. M., Hasan, G. T., & Ali, H. A. (2022). *Automatic evaluating of Russian-Arabic machine translation quality using METEOR method*. 040036. https://doi.org/10.1063/5.0067018

Haq, Y. N. (2022). *Manhaj al-Imam Ibn Aqil fi Syarhi Alfiah al-Imam Ibn Malik* [bachelor's thesis, Fakultas Dirasat Islamiah]. https://repository.uinjkt.ac.id/dspace/handle/123456789/61799

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

Hidayatullah, A. S., & Fauji, I. (2023). *Bridging Theory and Practice in Arabic Language Education | Indonesian Journal of Islamic Studies*. https://ijis.umsida.ac.id/index.php/ijis/article/view/1724

Inas, A. (2024). Analysis of Nahwu Content in Alfiyah Ibnu Malik. *ALIT: Arabic Linguistics and Teaching Journal*, *1*(1), Article 1. https://journal.zamronedu.co.id/index.php/alit/article/view/76

Inel, O., Draws, T., & Aroyo, L. (2023). Collect, Measure, Repeat: Reliability Factors for Responsible AI Data Collection. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *11*(1), Article 1. https://doi.org/10.1609/hcomp.v11i1.27547

Johnson, D., Goodman, R., Patrinely, J., Stone, C., Zimmerman, E., Donald, R., Chang, S., Berkowitz, S., Finn, A., & Jahangir, E. (2023). *Assessing the accuracy and reliability of AI-generated medical responses: An evaluation of the Chat-GPT model*. https://doi.org/10.21203/rs.3.rs-2566942/v1

Khoshafah, F. (2023). ChatGPT for Arabic-English Translation: Evaluating the Accuracy. *Ministry of Education, Yemen*. https://doi.org/10.21203/rs.3.rs-2814154/v2

Lavie, A., & Agarwal, A. (2007). Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments. *Proceedings of the Second Workshop on Statistical Machine Translation*, 228–231.

Lozano, M., Winthrop, S., Goldsworthy, C., Leventis, A., & Birkenshaw, A. (2024). *Semantic Depth Redistribution in Large Language Models to Contextual Embedding Preservation*. https://doi.org/10.22541/au.173083529.98863661/v1

Muthiah, A., & Zain, L. (2020). Konsep Ittishal Al-Sanad Sebagai Syarat Kajian Kitab Kuning Dalam Tradisi Pesantren An-Nahdliyyah Cirebon. *Jurnal Studi Hadis Nusantara*, *2*(1). https://jurnal.syekhnurjati.ac.id/index.php/jshn/article/download/6746/3133

Olsher, D. (2014). Semantically-based priors and nuanced knowledge core for Big Data, Social AI, and language understanding. *Neural Networks*, *58*, 131–147. https://doi.org/10.1016/j.neunet.2014.05.022

Perkins, M. (2023). Academic integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice, British University, Vietnam*, *20*(2). https://doi.org/10.53761/1.20.02.07

Raj, H., Gupta, V., Rosati, D., & Majumdar, S. (2023). *Semantic Consistency for Assuring Reliability of Large Language Models* (arXiv:2308.09138). arXiv. https://doi.org/10.48550/arXiv.2308.09138

Ras, G., van Gerven, M., & Haselager, P. (2018). Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges. *arXiv Preprint arXiv:1803.07517*.

Russell, R. G., Novak, L. L., Patel, M., Garvey, K. V., Craig, K. J. T., Jackson, G. P., Moore, D., & Miller, B. M. (2023). Competencies for the Use of Artificial Intelligence–Based Tools by Health Care Professionals. *Academic Medicine*, *98*(3), 348–356. https://doi.org/10.1097/ACM.0000000000004963

Stratton, J. (2024). An Introduction to Microsoft Copilot. In J. Stratton, *Copilot for Microsoft 365* (pp. 19–35). Apress.

Sulaeman, I., Syuhadak, S., & Sulaeman, I. (2023). ChatGPT as a New Frontier in Arabic Education Technology. *Al-Arabi: Jurnal Bahasa Arab Dan Pengajarannya= Al-Arabi: Journal of Teaching Arabic as a Foreign Language*, *7*(1), 83–105. http://dx.doi.org/10.17977/um056v7i1p83-105

Zhang, M. (2024). A Study on the Translation Quality of ChatGPT. *International Journal of Educational Curriculum Management and Research*, *5*(1). https://doi.org/10.38007/IJECMR.2024.050121