Department of Digital Business
Journal of Artificial Intelligence and Digital Business (RIGGS)
Homepage: https://journal.ilmudata.co.id/index.php/RIGGS
Vol. 5 No. 1 (2026) pp: 2153- 2160
P-ISSN: 2963-9298, e-ISSN: 2963-914X

Evaluating AI Tutor Interaction with Ambiguous Junior High Mathematics Questions Using Black-Box

Khoirul Islam[1], Nisfu Laili Saidah[2], Saifudin Yahya[3], Muhammad Miftakhul Syaikhuddin[4]
[1,3]Department of Informatics, Faculty of Engineering, Universitas Negeri Surabaya
[2]Department of Mathematic Education, Faculty of Education, Universitas Hasyim Asy'ari
[4]Department of Information System, Faculty of Science and Technology, Universitas Pesantren Tinggi Darul 'Ulum
khoirulislam@unesa.ac.id

Abstract

The growth of generative artificial intelligence as a mathematics education tutoring tool presents new possibilities of assisting student learning. Nevertheless, successful learning interactions involve AI tutors to address ambiguous questions posed by students in a proper manner, particularly in junior high school mathematics, where questions are frequently incomplete, have ambiguous concepts, or lack reference points to the surrounding context. This paper compares the quality of interaction of two popular AI tutors, ChatGPT and Gemini, in answering ambiguous questions in mathematics at the junior high school level. An approach that was used was black-box testing where only observable input-output behavior was tested without accessing internal model mechanisms. A sample of 50 ambiguous mathematics situations was randomly designed around five types of ambiguity, namely incomplete information, conceptual ambiguity, output format ambiguity, missing context, and contradictory information. Both AI tutors were tested on each scenario once giving a total of 100 dialog interactions. Two independent raters evaluated all interactions based on a Human–AI Interaction rubric which includes ambiguity detection, relevance of clarification, transparency of assumptions, quality of interaction, and quality of solution. The findings show that both systems can identify ambiguity, but Gemini shows higher rates of clarification and more pedagogically suitable interaction patterns than ChatGPT, especially in situations related to the lack of information and contextual ambiguity. The results demonstrate the significance of clarification behavior and interaction design in AI-based tutoring systems and offer a viable way of how AI tutors can be responsibly used in mathematics education at junior high schools.

Keywords: Human-AI Interaction, Generative AI Tutor, Ambiguity, Junior High School Mathematics, Black-Box.

1. Introduction

The development of Large Language Models (LLMs) and artificial intelligence-based tutoring services has opened up new opportunities to support mathematics learning, such as adaptive feedback and quick step-by-step explanations. However, systematic studies show that generative models do not always work well in educational contexts. Models tend to perform well for simple structured questions, but do not perform well for questions that require complex reasoning, graphical representations, or understanding of specific learning contexts[1].

Handling ambiguity in students' questions is one of the significant challenges of interaction. In junior high school mathematics learning, students often ask unclear questions (lacking information), use ambiguous terms, or refer to the context of previous conversations. In such a case, AI tutors should be capable of detecting ambiguity and posing relevant clarifying questions. Such misunderstandings and lack of trust in the user are enhanced in case the system fails to clarify or make any assumptions without being transparent[2]. The application of AI in mathematics to higher-order thinking problems (HOTP) indicates that the effectiveness of support highly depends on the capacity of the system to give contextually relevant feedback and ask about the information that is not provided[3].

Regarding the Human-AI Interaction research, the quality of dialogue, such as the capacity to request clarification, make assumptions, and mend conversations, is one of the main factors influencing the usability and pedagogical safety of AI tutors. Studies in clarification modeling and uncertainty-sensitive questioning indicate that proactive dialogues that request required information can enhance service accuracy and user satisfaction, yet commercial generative systems continue to fail to generate relevant and minimal clarifications[2]. The recent research on NLP also highlights the significance of the creation of clarification questions and uncertainty-conscious mechanisms to

maintain the discussions fruitful and prevent the misleading users[4]. Early studies in the education sector indicate that generative AI can be used to facilitate learning with the addition of features that promote reflection, transparency in reasoning, and formative feedback, rather than end-of-answer features[5]. Moreover, mathematical literature reviews also note that AI-based learning tools are most useful when they facilitate metacognitive reasoning, as opposed to giving final answers[6].

Also, commercial LLM services are black-box, which presents researchers with methodological difficulties: since researchers have little access to internal models (token probabilities, training data, or parameters), the evaluation of dialogue behavior must be performed in an input-output fashion (black-box testing). The black-box method has been applied to quantifying the functional behavior of models in practice without having to access internal architecture, and thus is especially appropriate to comparative studies of AI tutors as commercial services (API/web)[2].

In the pedagogical domain, the recent literature reviews point out that the incorporation of generative AI in mathematics instruction should take into account not only the correctness of the answers, but also such pedagogical factors as the clarity of the reasoning, the feedback that promotes the understanding, and the clarification mechanisms that are age-appropriate. These suggestions apply to junior high school environments, in which different forms of ambiguity frequently occur in story problems and homework tasks[7].

In accordance with this gap, i.e., the necessity to find empirical evidence of how generative AI tutors act when presented with ambiguous questions in the setting of junior high school mathematics, and a lack of internal access to commercial systems, this study performed a comparative analysis based on black-box testing techniques. This study aims at quantifying the clarification, transparency of assumptions, and quality of problem solving of two popular AI tutors in situations involving 50 ambiguous questions that are representative of common forms of ambiguity in learning mathematics in junior high schools. The findings will make an empirical contribution to the body of literature on Human-AI Interaction and offer practical suggestions on how AI tutors can be used in the classroom.

2. Research Methods

2.1. Research Design

The experimental approach adopted in this study is a comparative one where the interaction behavior of generative AI tutors is tested using black-box testing methods to assess how they interact with ambiguous mathematical questions. The black-box method was selected due to the fact that the AI system under testing is a closed commercial service, and the assessment is made on the correlation between the input and output of the system, but does not access the internal mechanisms of the model[8].

The primary area of the evaluation is the Human-AI Interaction area, which is the capability of the AI tutor to identify ambiguity, pose clarifying questions, express assumptions clearly, and solve problems pedagogically based on the context of learning mathematics at junior high school..

2.2. Subjects and Systems Tested

The ChatGPT and Gemini two generative AI tutors, which were used in the study, were chosen because they are popular and used in numerous educational settings. Table 1 displays the specifications of the AI tutors tested in detail.

Table 1. Specifications of the AI Tutor

Aspects	ChatGPT	Gemini
System Provider	OpenAI	Google
System Type	Generative AI Tutor	Generative AI Tutor
Account Access	Premium / Paid	Premium / Paid
Testing Mode	ChatGPT 5.2 Instant	Gemini 3 Instant
Evaluation Approach	Black-box testing	Black-box testing
Interaction Language	Indonesian	Indonesian
Conversation Type	New chat for each scenario	New chat for each scenario
Conversation History	Not used	Not used
System Personalization	Disabled	Disabled

DOI: https://doi.org/10.31004/riggs.v5i1.6363

Number of Scenarios Tested	50 scenarios	50 scenarios
Number of Dialogues	50 dialogues	50 dialogues
Role in Research	System A	System B

The specifications of the AI tutors that were tested in this study are presented in Table 1. The settings of both systems were adjusted to similar testing conditions, such as using premium accounts, instant mode, and new conversation settings in both cases. Such settings were intended to reduce the effects of other variables and make sure that the differences in responses were based on the interaction behavior of the systems to the ambiguous questions.

It is also significant to use similar configurations to reduce performance bias based on the differences in computing access or service restrictions, as suggested in the research on service-based AI systems[9].

All the tests were performed under the following conditions:

a. The scenarios were performed in a fresh conversation (no prior history).
b. User memory or personalization was disabled.
c. The language of interaction was the Indonesian.
d. Each of the scenarios was executed once on each system.

The goal of this setting is to make sure that the differences in responses that are observed are due to the interaction behavior of the system with ambiguity, and not because of the user history or personal settings. To explain the testing process and the location of each element in the testing.

2.3. Research Dataset

The research dataset consists of 50 junior high school-level math question scenarios compiled based on student question characteristics and curriculum material. Each scenario is designed to represent one of the five types of ambiguity as shown in Table 2.

Table 2. Five Types of Ambiguity

Code	Type of Ambiguity	Number of scenarios
D1	Ambiguity Detection	10 questions
D2	Clarification Relevance	10 questions
D3	Assumption Transparency	10 questions
D4	Interaction Quality	10 questions
D5	Resolution Quality	10 questions

This classification of ambiguity refers to linguistic and dialogue studies that group ambiguities based on the source of information uncertainty and context. A controlled scenario-based approach is commonly used in system dialogue evaluation to ensure comparability between systems [10].

Each type of ambiguity is represented by 10 scenarios, resulting in a balanced dataset distribution. Each scenario is arranged in a dialogue format, consisting of:

a. An ambiguous initial question from the student,
b. An initial response from the AI tutor,
c. (If any) a clarification question from the AI,
d. A clarification answer from the student that has been determined by the researcher, and
e. A final response from the AI tutor.

This method enables it to be evaluated at the single-turn and multi-turn interaction levels.

2.4. Testing Procedures

Figure 1 shows the general testing process..

DOI: https://doi.org/10.31004/riggs.v5i1.6363

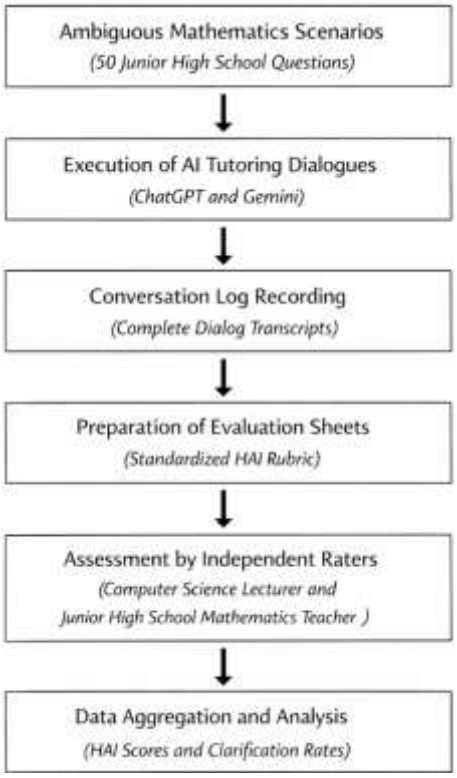

Figure 2. Black-box Testing Procedure for AI Tutors

The figure shows the chronological steps of the experimental process, starting with the construction of the scenarios and the performance of the dialogues to the assessment of the experts and the analysis of the data, without having to learn the internal processes of the AI models. Evaluation of conversation logs eliminates the differences due to typing styles or irregular chat resets, and comparisons across systems are more legitimate[11]. This was done to make sure that all testers evaluated the same data, which is the principle of black-box testing.

2.5. Testers and Assessment Process

Two independent assessors were used in the evaluation, but with different backgrounds, which included a computer science lecturer and a junior high school mathematics teacher. This mixture was intended to represent the technical view of AI interaction and the pedagogical view of learning mathematics. The assessors did pre-calibration before the actual assessment where a small section of the scenarios was used to standardize their interpretation of the assessment rubric.

2.6. Assessment Tools

The assessment tool is a Human and AI Interaction rubric that has a Likert scale of 1-5 and the dimensions are as follows:

Table 3. Human AI Interaction Rubric

Code	Dimensions	Description
D1	Ambiguity Detection	The AI capability to identify ambiguity.
D2	Clarification Relevance	The precision of the clarification questions posed.
D3	Assumption Transparency	The intelligibility of the AI in the formulation of assumptions or alternative interpretations.
D4	Interaction Quality	The readability, comprehensibility and suitability of the language to junior high school students.
D5	Resolution Quality	The final solution after clarification in terms of accuracy and clarity.

DOI: https://doi.org/10.31004/riggs.v5i1.6363

This part was created based on the recommendations to assess human-AI interactions and pedagogical research associated with learning feedback[12]. Moreover, it is also mentioned whether the AI poses the corresponding clarification questions (yes/no) in each case.

2.7. Data Analysis Techniques

Nonparametric statistics that were suitable to analyze ordinal data were used to analyze the data descriptively and inferentially. The analysis was aimed at comparing the scores of AI tutors and determining the level of clarification. This practice is consistent with the method of analyzing dialogue evaluation data and AI-based education systems[10].

3. Results and Discussions

3.1. Overview of Assessment Results

This section will give a summary of the findings of AI tutor interaction assessment using the Human AI Interaction rubric. The analysis was performed on 100 conversations based on 50 ambiguous question situations, which were run on ChatGPT and Gemini. The quantitative findings are summarized in terms of average scores, the variability of scores, the level of clarification, and distribution of scores in terms of the type of ambiguity.

Table 4. Mean Scores of Human–AI Interaction Evaluation

Code	Dimensions	ChatGPT	Gemini
D1	Ambiguity Detection	3.72	4.21
D2	Clarification Relevance	3.15	4.05
D3	Assumption Transparency	3.44	4.02
D4	Interaction Quality	3.58	4.18
D5	Resolution Quality	3.91	4.12

Table 4 indicates that Gemini scored better than ChatGPT in all the assessment dimensions. The most significant differences were observed in the clarification relevance (D2) and interaction quality (D4) dimensions, which are directly connected to the fact that the system should be able to handle a two-way dialogue. These results suggest that the primary difference in the performance of the two AI tutors does not consist of the quality of the final answers, but in the way the system reacts and addresses ambiguities in the questions of the students.

Table 5. Standard Deviation of HAI Scores

Code	Dimensions	ChatGPT	Gemini
D1	Ambiguity Detection	0.64	0.52
D2	Clarification Relevance	0.71	0.48
D3	Assumption Transparency	0.68	0.55
D4	Interaction Quality	0.59	0.46
D5	Resolution Quality	0.61	0.49

The difference in scores provided in Table 5 indicates that Gemini is not only better in its average score, but also demonstrates more consistency in responding to questions than ChatGPT in nearly every dimension. The reduced standard deviation of Gemini, especially in the dimensions D2 and D4, implies that the clarification behavior and the quality of interaction of the system is relatively constant in different ambiguity situations.

Table 6. Clarification Rates Across AI Tutors

AI Tutor	Clarification Rate (%)
ChatGPT	46%
Gemini	72%

Table 6 indicates that there is a big difference in the degree of clarification between the two AI tutors. In the majority of ambiguous situations, Gemini poses clarifying questions, whereas ChatGPT is likely to give an answer without clarification in nearly half of the cases. This trend is significant in relation to the learning of mathematics in the junior high school setting since clarification is significant in avoiding implicit assumptions that may ultimately result in misconceptions.

DOI: https://doi.org/10.31004/riggs.v5i1.6363

Table 7. Average HAI Scores by Ambiguity Type

Code	Dimensions	ChatGPT	Gemini
T1	Incomplete Information	3.42	4.10
T2	Ambiguous Meaning	3.55	4.08
T3	Answer Format	3.62	4.00
T4	Missing Context	3.21	4.15
T5	Contradictory Information	3.48	4.05

According to Table 7, the greatest discrepancies between the two AI tutors are in T4 (Missing Context) and T1 (Incomplete Information). ChatGPT is more likely to make guesses in these kinds of ambiguity or proceed with the calculation with implicit assumptions, whereas Gemini is more likely to seek clarification before resolving the problem. On the other hand, the difference in scores in T3 (Answer Format) is not as large, which means that both systems are quite able to deal with ambiguities concerning the output representation.

Overall, Gemini performed better than ChatGPT on nearly every assessment dimension. The most significant differences were observed in the clarification relevance (D2) and interaction quality (D4) dimensions, which are directly connected to the fact that the system should be able to handle a two-way dialogue. Such results suggest that the quality of interaction, and not the simple accuracy of answers, is the distinguishing factor of the situation with educational AI tutors, which is consistent with the recent research on Human-AI Interaction, which highlights the significance of adaptive dialogue and proactive clarification[13].

3.2. Comparison of Ambiguity Detection and Clarification

Regarding ambiguity detection (D1), both AI tutors were able to identify ambiguity in questions by students. Nevertheless, there were variations in their follow-up on that ambiguity. Gemini would more frequently ask clarification questions that are relevant, whereas ChatGPT in some instances would tend to make a leap to conclusions based on some assumptions and not seek further confirmation. This behavior increases the risk of overconfidence and conceptual errors—phenomena documented in previous educational AI studies[14].

This outcome is shown in the clarification rate, in which Gemini has a higher percentage of clarification questions compared to ChatGPT. The significance of this pattern is that HAI literature indicates that the inability of a system to seek clarification during ambiguous situations may lead to the escalation of the risk of overconfidence and conceptual errors[[15]. In the context of junior high school mathematics learning, assumptions that are not explicitly stated have the potential to cause misconceptions, especially among students who are still building basic conceptual understanding.

3.3. Analysis Based on Type of Ambiguity

The additional analysis using five categories of ambiguity reveals that AI tutor behavior achieved in the analysis varies in Table 8.

Table 8. Analysis based on Type of Ambiguity

Code	Type of Ambiguity	Description
T1	Incomplete Information	Gemini always wants to obtain more related information to fill in a question, whereas ChatGPT is more likely to give one answer with unspoken assumptions. This trend can be aligned with the results that generative models tend to fail to differentiate between underdetermined problems and those that are directly solvable.
T2	Ambiguous Meaning	Both AI tutors are relatively capable of identifying ambiguity, but Gemini more often offers alternative interpretations or requests confirmation of user preferences. This method is deemed closer to the principle of pedagogical alignment, which puts the significance of searching meaning prior to giving an ultimate answer.
T3	Answer Format	Both AI tutors are relatively capable of identifying ambiguity, but Gemini more often offers alternative interpretations or requests confirmation of user preferences. This method is deemed closer to the principle of pedagogical alignment, which puts the significance of searching meaning prior to giving an ultimate answer.

DOI: https://doi.org/10.31004/riggs.v5i1.6363

T4	Missing Context	The most apparent distinction is that ChatGPT in certain instances attempts to infer the context that the student wants to convey whereas Gemini more frequently requests the user to restate the question word-to-word. This context guessing behavior is not safe in the HAI perspective since it may steer the conversation in the wrong direction.
T5	Contradictory Information	Both systems can identify inconsistency in data, although Gemini is more reliable in pointing out the contradictions prior to proceeding with calculations. The results are in line with earlier research that suggests that error checking abilities are still a problem to the LLM-based AI tutor.

3.4. Interaction Quality and Problem Solving

The dimension of interaction quality (D4) reveals that Gemini is inclined to use more structured and suitable language to the level of understanding of the junior high school students. Meanwhile, in other instances, ChatGPT will offer mathematically correct explanations and less attention to readability and logic sequence.

The difference in the scores is not as high as that of the interaction dimension in terms of solution quality (D5). This demonstrates that the primary distinction between the two AI tutors is not their calculating capacity, but their manner of interacting with and directing users. These results support the thesis that the assessment of educational AI tutors must not only be based on the quality of answers, but also focus on the quality of dialogue and cognitive assistance[13].

3.5. Implications for Junior High School Mathematics Learning

The findings of this paper have significant implications in the use of AI tutors in mathematics learning in junior high school. AI tutors, who can deal with ambiguity effectively, can decrease misconceptions and boost student confidence. On the other hand, systems that are too fast to arrive at conclusions without clarification may serve to strengthen misunderstandings.

These results suggest that in the framework of formal education, teachers should teach students to be critical when using AI tutors, such as by encouraging students to verify assumptions and seek alternative explanations. Moreover, the outcomes can be used to feed AI system developers with the input to enhance clarification mechanisms and transparency of assumptions in learning conversations.

4. Conclusion

This paper will analyze how AI tutors can interact in solving ambiguous math problems in the junior high school level through the black-box testing approach. The analysis is based on the elements of human-AI interaction, namely the possibility to identify ambiguity, the importance of clarification, the clarity of assumptions, the quality of interaction, and the quality of problem solving. The findings reveal that despite the fact that both ChatGPT and Gemini have the capacity to identify ambiguity in the questions of students, there are always differences in the way they pursue the ambiguity. Gemini is more common and more precise in enquiring about clarifying questions that are relevant and show a more organized nature of interaction which is befitting the junior high school mathematics learning environment. Conversely, ChatGPT in several instances simply gives answers with implicit assumptions without sufficient explanations. These findings confirm that the effectiveness of AI tutors in education is not only determined by the accuracy of answers, but also by the quality of dialogic interactions that support student understanding. Pedagogical and explicit capacity to deal with ambiguity is one of the main aspects of avoiding misconceptions and making AI tutors more acceptable to users. Practically, the results of this study have implications for educators and AI system developers. Teachers need to guide students to use AI tutors critically, while developers are expected to strengthen clarification mechanisms and transparency of assumptions in learning dialogues. For further research, it is recommended to involve direct interaction with students and evaluate the long-term impact of using AI tutors on mathematical concept understanding.

DOI: https://doi.org/10.31004/riggs.v5i1.6363

Reference

[1] B. Pepin, N. Buchholtz, and U. Salinas-Hernández, "A Scoping Survey of ChatGPT in Mathematics Education," *Digital Experiences in Mathematics Education*, vol. 11, no. 1, pp. 9–41, Apr. 2025, doi: 10.1007/s40751-025-00172-1.

[2] A. Testoni and R. Fernández, "Asking the Right Question at the Right Time: Human and Model Uncertainty Guidance to Ask Clarification Questions," Long Papers. [Online]. Available: https://chat.openai.com/

[3] M. Nafis Mumtaz, I. Ramadani, I. Sunan, and K. Yogyakarta, "Analysis of the Utilization of AI Chat GPT in the Academic Life of Prospective Mathematics Teacher Students Universitas Negri."

[4] R. Deng, M. Jiang, X. Yu, Y. Lu, and S. Liu, "Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies," *Comput. Educ.*, vol. 227, Apr. 2025, doi: 10.1016/j.compedu.2024.105224.

[5] W. Holmes, M. Bialik, and C. Fadel, "Artificial Intelligence In Education Promises and Implications for Teaching and Learning," 2019. [Online]. Available: http://bit.ly/AIED-

[6] L. Busuttil and J. Calleja, "Teachers' Beliefs and Practices About the Potential of ChatGPT in Teaching Mathematics in Secondary Schools," *Digital Experiences in Mathematics Education*, vol. 11, no. 1, pp. 140–166, Apr. 2025, doi: 10.1007/s40751-024-00168-3.

[7] C. Walkington, "The implications of generative artificial intelligence for mathematics education," *Sch. Sci. Math.*, 2025, doi: 10.1111/ssm.18356.

[8] A. B. Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Dec. 2019, [Online]. Available: http://arxiv.org/abs/1910.10045

[9] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, Inc, Mar. 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[10] N. U. J. *et al.*, "A Systematic Review of Generative AI in Education," *Journal of Computer Sciences and Applications*, vol. 12, no. 1, pp. 25–30, Sep. 2024, doi: 10.12691/jcsa-12-1-4.

[11] Y. Tsuta, N. Yoshinaga, and M. Toyoda, "Uncertainty-aware Automatic Evaluation Method for Open-domain Dialogue Systems," 2023.

[12] S. Amershi *et al.*, "Guidelines for human-AI interaction," in *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery, May 2019. doi: 10.1145/3290605.3300233.

[13] E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," Apr. 01, 2023, *Elsevier Ltd*. doi: 10.1016/j.lindif.2023.102274.

[14] Y. Wardat, M. A. Tashtoush, R. AlAli, and A. M. Jarrah, "ChatGPT: A revolutionary tool for teaching and learning mathematics," *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 19, no. 7, 2023, doi: 10.29333/ejmste/13272.

[15] M. J. Q. Zhang and E. Choi, "Findings of the Association for Computational Clarify When Necessary: Resolving Ambiguity Through Interaction with LMs," Apr. 2025.

DOI: https://doi.org/10.31004/riggs.v5i1.6363