ChatGPT vs. NotebookLM: WHICH AI TOOL BETTER GENERATES MCQ LISTENING TEST ITEMS?

Fitriani

Department of English Language Education

STAIN Mandailing Natal

Mandailing Natal, Indonesia

fitriemje@gmail.com

Abstract

This study compares the quality of multiple-choice listening test items generated by ChatGPT and NotebookLM for EFL classroom-based assessment, focusing on two key dimensions: stem clarity and distractor plausibility. Using a small-scale comparative evaluative design, the study analyzed thirty AI-generated listening items, consisting of fifteen items produced by ChatGPT and fifteen items produced by NotebookLM. The items were generated from three listening transcripts representing different listening contexts and objectives through an identical zero-shot prompting procedure. Each item was evaluated using an analytic rubric covering stem clarity and distractor plausibility, supported by qualitative content analysis of the stems, options, answer keys, and alignment with the source transcripts. The findings show that both tools generated highly clear and focused stems, although ChatGPT performed slightly better in producing concise, direct, and learner-friendly question stems. However, a more notable difference was found in distractor plausibility. NotebookLM produced more plausible, competitive, and transcript-grounded distractors, particularly for speaker intention, main idea, and simple inference items, whereas ChatGPT often produced formally consistent distractors that were sometimes weakly connected to the listening input, easily eliminated through contextual guessing, or affected by answer-key patterns. These findings indicate that ChatGPT and NotebookLM offer complementary strengths in AI-assisted listening item development. ChatGPT may be more useful for drafting clear stems, while NotebookLM may provide stronger support for constructing source-based distractors. Nevertheless, neither tool can be used as a standalone solution for developing valid listening assessments. Human expert review remains essential to ensure construct validity, improve distractor function, remove technical clues, and align AI-generated items with sound principles of EFL listening assessment.

Keywords - AI-assisted language assessment, EFL listening assessment, multiple-choice items, stem clarity, distractor plausibility

Introduction

The construction of listening assessments is a complex, labor-intensive, and costly process since it requires specialized expertise in designing several interrelated test components, including script selection, audio production, and item development. Aryadoust et al. (2024) emphasize that listening assessment development demands careful attention to task design, while Buck (2001) notes that listening tests involve distinctive challenges as they must represent how spoken language is processed in real time. In EFL higher education, further, listening tests are expected to provide valid evidence of learners' ability to process spoken input, identify relevant information, infer meaning, and interpret communicative intent. Thus, listening assessment is not simply a matter of writing questions after an audio text, but a systematic process requiring assessment literacy, technical judgment, and careful control of construct-relevant difficulty.

Within this demanding process, multiple-choice questions (MCQs) test type remain common due to their efficiency in administration, objective scoring, and practical value in interpreting students' performance (Butler, 2018). These advantages make MCQs particularly attractive in classroom-based and large-scale EFL assessment, where teachers and lecturers often need to assess learners regularly and consistently. However, the practicality of MCQs does not reduce the need for rigorous item construction. Effective MCQ listening items require clear stems that specify the task unambiguously, answer options that align with the audio input, and plausible distractors to attract learners who misunderstand or only partially understand the

listening text (Anckar, 2011). Poorly written stems may obscure the

intended construct, while weak distractors may fail to distinguish learners who comprehend the audio from those who rely on guessing or test-wise strategies. As Haladyna (2004) argues, effective MCQ development depends on systematic procedures for constructing and validating items; therefore, the quality of MCQ listening items depends not only on linguistic accuracy but also on how well stems and distractors support valid measurement.

In classroom-based EFL assessment, the challanges of constructing high-quality MCQ listening items is intensified by the need to serve both formative and summative purposes (Trinh, 2023). Teachers are often required to design repeated assessments to monitor students' progress, provide feedback, and evaluate achievement across instructional units. This demand places considerable pressure on assessment practice as item writing requires linguistic competence, assessment literacy, knowledge of item-writing principles, and sufficient time for review and revision. Consequently, manual item development remains a demanding professional task, especially in higher education contexts where teachers must balance assessment design with teaching, research, and administrative responsibilities. These practical constraints have encouraged growing interest in tools that can assist educators in developing assessment materials more efficiently without weakening the quality and validity of the resulting test items.

To address these constraints, generative artificial intelligence has increasingly been explored as a support tool for assessment development. As stated by Xu et al.,(2024) that large language models and AI-based educational tools can quickly generate prompts, answer options, and alternative item versions, suggesting potential benefits for reducing teachers' workload and supporting automatic item generation. Circi et al. (2023) describe automatic item generation as an important development in assessment design, particularly as machine learning approaches become more widely adopted, while Chun and Barley (2024) show that AI-generated MCQs can be evaluated in comparison with human-developed items. In language assessment, such tools may help educators produce initial drafts of listening comprehension items for later review, revision, and validation by human experts. Nevertheless, D. Shin et al. (2025) underscores that the value of AI-generated items cannot be judged only by speed, fluency, or grammatical correctness. For AI-generated listening items to be useful in assessment contexts, Wang & Meng (2026) emphesizes that their quality must be examined through item-writing principles, especially in relation to whether they produce clear stems and plausible distractors.

However, current research has predominantly focused on the general affordances of AI rather than on the item-level qualities that determine whether AI-generated MCQs are valid and viable (Alhazmi et al., 2024; May et al., 2025). This gap is particularly critical since stem clarity and distractor plausibility directly affect the validity and interpretability of MCQ listening assessment (Yanagawa & Green, 2008). Ambiguous or overly complex stems may introduce construct-irrelevant difficulty, whereas distractors that are too obvious, unrelated to the listening text, or answerable through world knowledge alone may fail to measure listening comprehension accurately. Haladyna (2004) stresses that distractors must function effectively within MCQ design, and Shin et al. (2019) further highlight the importance of systematic distractor development. Thus, evaluating AI-generated MCQ listening items requires more than judging whether the generated language appears fluent; it requires careful analysis of whether the items function according to accepted principles of assessment quality.

Another limitation in the literature is the lack of comparative evidence across different AI tools. Many discussions treat generative AI as a single category, although different tools operate differently and may therefore produce different outputs. ChatGPT functions mainly as a general-purpose conversational AI tool, whereas NotebookLM is designed to generate responses grounded in user-provided source materials thereby supporting source verification and reducing hallucination information (Education, 2024). This distinction is relevant to listening test development as item quality depends on how accurately a tool uses the listening script, formulates the stem, and constructs distractors that are plausible but not misleading.

However, it remains unclear whether these two tools differ in their capacity to

produce unambiguous stems and viable distractors for listening MCQs within an EFL higher education context. Against this background, the present study compares the quality of MCQ listening test items generated by ChatGPT and NotebookLM, focusing specifically on stem clarity and distractor plausibility. Using a small-scale comparative mixed-method design, the study analyzes the generated items quantitatively and qualitatively to identify patterns of quality, strengths, and limitations across the two tools. By comparing a general-purpose generative AI tool and a source-grounded AI tool, this study contributes to research on AI-assisted language assessment and offers practical insights for EFL educators seeking to use AI critically, efficiently, and responsibly in listening test development.

Methodology

This study employed a comparative evaluative research design integrating descriptive scoring and qualitative content analysis. The design was selected to evaluate and compare the quality of multiple-choice listening comprehension items generated by two artificial intelligence tools, namely ChatGPT and NotebookLM. The descriptive scoring component provided a quantitative basis for comparing item quality, while qualitative content analysis allowed a deeper interpretation of the pedagogical and linguistic characteristics of the generated items that could not be fully explained through numerical scores alone.

The objects of this study were AI-generated multiple-choice listening comprehension test items. The unit of analysis was each individual item, consisting of the question stem, answer options, correct answer, distractors, and its alignment with the source transcript. A total of thirty items were analyzed, comprising fifteen items generated by ChatGPT and fifteen items generated by NotebookLM. The data were generated on March 27, 2026, using the free versions of both platforms, specifically ChatGPT with GPT-4o mini and the latest web-based version of NotebookLM available at the time of data collection.

Data were collected through a controlled prompting procedure. Both AI tools received identical prompts for each source text and were instructed to generate five multiple-choice listening comprehension items based on the transcript/audio. The items were generated using a zero-shot prompting strategy. The prompt was carefully engineered based on established assessment frameworks, including Haladyna and Rodriguez's principles (2004) of multiple-choice item construction, Buck's (2001) standards for listening assessment, and Fulcher's (2024a) criteria for text-based item validity. These frameworks were incorporated to ensure that the constructed items followed relevant pedagogical and assessment constraints.

The research instrument used in this study was an analytic scoring rubric developed to evaluate the quality of the AI-generated items. The rubric consisted of two dimensions: stem clarity and distractor plausibility as presented in Table 1.

Table 1. Analytic Scoring Rubric for AI-Generated MCQ Listening Test Items

Dimensions	Focus of Evaluation	Score
Stem Clarity	Evaluates task clarity, single focus, concise/unambiguous language, and independence from options without providing unintended clues.	1–4
Distractor Plausibility	Assesses relevance to transcript, homogeneity, grammatical parallelism, and absence of technical or verbal clues.	1–4

Each dimension was scored on a four-point scale, ranging from 1 = poor, 2 = needs revision, 3 = good, to 4 = excellent. Since the rubric consisted of three dimensions, the maximum score for each item was 12. Detailed indicators for each dimension were used to guide the scoring process.

Data analysis was carried out through two complementary stages. First, descriptive scoring was conducted by evaluating each item using the two dimensions of the analytic rubric. The total and mean scores were calculated to compare the overall quality of items generated by ChatGPT and NotebookLM, while the mean score for each dimension was examined to identify specific quality patterns across the two AI tools. Second, qualitative content analysis was

employed to provide a deeper interpretation of the scoring results. Each item was analyzed in terms of its stem, answer options, correct answerand distractors. The analysis identified recurring strengths and weaknesses, such as clarity or ambiguity of stems and plausibility of distractors. Representative examples were used to justify high and low scores. All scoring and analysis were conducted solely by the researcher to maintain interpretative consistency, and an intra-rater reliability check was performed after a two-week interval to ensure scoring stability, objectivity, and alignment with the established EFL assessment criteria rather than subjective impression.

To strengthen trustworthiness, all source transcripts, prompts, raw AI outputs, scoring sheets, and analytic notes were documented as an audit trail. The use of identical prompts, identical source texts, verbatim AI outputs, and rubric-based analysis supported procedural transparency and consistency. The qualitative interpretation was also grounded in the rubric to ensure that the findings were evidence-based, dependable, and confirmable through a transparent research process.

Finding and Discussion

Stem Clarity in Listening MCQ Test Items

Quantitative Findings

Stem clarity was analyzed using a four-point analytic rubric covering seven indicators: task clarity, single focus, conciseness, unambiguous wording, independence from options, absence of unintended clues, and A1-level appropriateness. The analysis involved 30 listening MCQ items, 15 items each, generated by ChatGPT and NotebookLM across three listening sources and five test objectives.

Overall, the findings show that both tools produced highly clear stems. The combined score was 116 out of 120, with an overall mean of 3.87, indicating an excellent level of stem clarity. No item received a score of 1 or 2, meaning that none of the stems were categorized as poor or requiring major revision. This suggests that both ChatGPT and NotebookLM were able to formulate stems that were generally understandable, focused, and suitable for beginner-level listening comprehension.

However, ChatGPT showed slightly stronger performance than NotebookLM. ChatGPT obtained 59 out of 60 with a mean score of 3.93, while NotebookLM obtained 57 out of 60 with a mean score of 3.80. ChatGPT also produced more maximum-score items, with 14 out of 15 items receiving a score of 4, compared with 12 out of 15 items for NotebookLM. This indicates that ChatGPT was more consistent in producing concise and learner-friendly stems.

Table 2. Quantitative Summary of Stem Clarity

Aspect	ChatGPT	NotebookLM	Overall
Number of items	15	15	30
Total score	59/60	57/60	116/120
Mean score	3.93	3.80	3.87
Achievement percentage	98.33%	95.00%	96.67%
Items scoring 4	14	12	26
Items scoring 3	1	3	4
Items scoring 1–2	0	0	0

Although both tools achieved excellent levels of stem clarity, the difference between ChatGPT and NotebookLM was relatively small. ChatGPT's mean score was 3.93, while NotebookLM's mean score was 3.80. This difference suggests that ChatGPT was only marginally stronger in producing concise and direct stems, rather than substantially superior.

Across test objectives, Specific Information, Detail Comprehension, and Main Idea/Gist achieved the highest combined mean score of 4.00. This indicates that both tools were highly effective in generating clear stems for factual, detail-based, and gist-based questions. In contrast, Speaker Intention/Function and Simple Inference received lower combined means of 3.67. These objectives were more likely to produce stems with minor issues, such as longer wording, less natural phrasing, or additional inferential demand.

Table 3. Stem Clarity by Test Objective

Test Objective	ChatGPT Mean	Notebook LM Mean	Combined Mean
Specific Information	4.00	4.00	4.00
Detail Comprehension	4.00	4.00	4.00
Speaker Intention / Function	4.00	3.33	3.67
Main Idea / Gist	4.00	4.00	4.00
Simple Inference	3.67	3.67	3.67

Across listening sources, Audio 1 and Audio 3 both reached a combined mean of 3.90, while Audio 2 obtained 3.80. The slightly lower score for Audio 2 suggests that phone-message contexts may require more careful stem construction, especially for inference and speaker-function items.

In other words, the stem clarity of the AI-generated listening MCQ items was generally excellent. ChatGPT performed slightly better, particularly in conciseness and A1-level appropriateness, while NotebookLM remained highly viable but showed minor weaknesses in stem length and wording naturalness. Human review is still needed, especially for Speaker Intention/Function and Simple Inference items.

Qualitative Findings

The qualitative findings complement the quantitative analysis by showing that ChatGPT and NotebookLM achieved stem clarity through different strengths. ChatGPT tended to produce stems that were clearer, shorter, and more direct. Most of its stems explicitly indicated the information students needed to listen for, such as "What is Marina's phone number?", "What time does the class begin?", and "Who is Lindsay Black?". These stems were easy to process because they used simple question forms and avoided unnecessary contextual information. This pattern supports ChatGPT's strength in explicitness of task, brevity and conciseness, and linguistic simplicity.

In terms of single focus, ChatGPT also performed consistently well. Most of its stems asked for only one piece of information, such as the number of books, the required course book, or the time a class begins. This reduces the possibility of confusion and helps students focus on one listening target. However, some ChatGPT stems, especially for main idea/gist items, were rather general. Questions such as "What is the message mainly about?" and "What is the talk mainly about?" were clear, but they provided limited contextual guidance. As a result, students may need to rely more on the answer options to understand the intended scope of the question.

NotebookLM, on the other hand, showed stronger performance in specificity and independence from options. Several stems provided more contextual information, such as "What does Marina want John to send to her email?", "What is the main reason for Marina's call?", and "What is the teacher mainly talking about in this recording?". These stems were more self-contained because students could understand the focus of the question without depending heavily on the answer choices. Nevertheless, this contextual strength sometimes resulted in longer stems. For example, "What is Marina doing when she says 'Could you please call me when you are back in the office?'" is clear but may increase reading load for lower-level EFL learners.

Both tools generally avoided serious ambiguity, although their weaknesses differed. ChatGPT occasionally produced stems that were too broad, while NotebookLM sometimes produced stems that were longer or more syntactically complex. Both tools also showed possible unintended cues in speaker-function items because direct quotations such as "Quiet, please", "Please arrive on time", and "Could you please call me…" could make the correct answer too predictable. The comprehensive qualitative findings are presented in Table 4.

Table 4. Summary of Qualitative Findings on Stem Clarity

Aspect	ChatGPT	NotebookLM
Main strength	Simple, concise, direct stems	Specific and contextual stems
Strongest categories	Explicitness, brevity, linguistic simplicity	Specificity, independence from options

Main weakness	Some gist stems are too general	Some stems are longer and heavier
Overall interpretation	Better for clarity and simplicity	Better for contextual guidance

Distractor Plausibility of Listening MCQ Test Items

Quantitative Findings

The analysis of 30 multiple-choice listening test items indicates that NotebookLM produced stronger distractor plausibility than ChatGPT. The analysis used eight distractor plausibility indicators: relevance to the transcript, plausibility, competitiveness, homogeneity, grammatical parallelism, similar length and form, absence of technical/verbal clues, and not answerable by world knowledge alone. Each indicator was rated on a 1–4 scale, where 4 represents Excellent and 1 represents Poor.

NotebookLM obtained a mean score of 3.25, while ChatGPT obtained a mean score of 2.63. Based on the rubric descriptors, NotebookLM falls into the Good category, whereas ChatGPT is positioned between Needs Revision and Good. This finding suggests that NotebookLM's distractors were generally more relevant to the listening context, more plausible, and more competitive in attracting test-takers with partial comprehension of the audio input. In contrast, ChatGPT's distractors showed several weaknesses, particularly in option competitiveness and technical clues caused by answer-key patterns.

Table 5. Overall Distractor Plausibility Scores

Source	Number of Items	Mean Score	Category
ChatGPT	15	2.63	Needs Revision–Good
NotebookLM	15	3.25	Good

At the indicator level, the most notable differences appeared in no technical/verbal clues, competitiveness, and not answerable by world knowledge alone. ChatGPT received the lowest score, 1.5, for no technical/verbal clues because most of its correct answers were placed in option A. This pattern may function as a test-wise clue and reduce item validity. In contrast, NotebookLM scored 3.0 on the same indicator because its answer keys were more evenly distributed, although several items still contained rather direct verbal cues.

For competitiveness, ChatGPT scored 2.0, while NotebookLM scored 3.0, indicating that ChatGPT's distractors were easier to eliminate because some options were less connected to the listening context. For example, in Audio 1, Item 4, ChatGPT used distractors such as "Meeting friends on campus" and "Registering for a course," which were only loosely related to the library context. This also relates to ChatGPT's lower score on not answerable by world knowledge alone, as some items could be answered through contextual guessing rather than careful listening. In Audio 3, Item 5, for instance, students might infer "the course teacher" as the answer to "Who is Lindsay Black?" from the classroom context and options, even without fully processing the audio. By contrast, NotebookLM more often produced distractors within the same semantic field, such as similar phone numbers (Audio 2, Item 1) and time-based options (Audio 3, Item 1), making them more competitive. However, some NotebookLM distractors, such as "new laptop," "set of headphones," and "Friday party," were still insufficiently plausible. Thus, NotebookLM generally produced more plausible distractors, whereas ChatGPT more often included options that could be eliminated through general knowledge or contextual mismatch.

Table 6. Scores by Distractor Plausibility Indicators

Indicator	ChatGPT	NotebookLM
Relevance to transcript	3.0	3.0
Plausibility	3.0	3.5
Competitiveness	2.0	3.0
Homogeneity	3.0	3.5
Grammatical parallelism	3.5	3.5
Similar length and form	3.0	3.5
No technical/verbal clues	1.5	3.0

Not answerable by world knowledge alone	2.0	3.0
Mean	2.63	3.25

In terms of grammatical parallelism, both sources received the same score, 3.5. This shows that ChatGPT and NotebookLM were generally able to construct grammatically parallel options, including noun phrases, prepositional phrases, and infinitive phrases. Therefore, the main weakness of the two item sets was not the grammatical form of the options, but rather the semantic quality and competitiveness of the distractors.

The analysis by test objectives shows that both sources performed best on Specific Information items, with both achieving a score of 4.00. These items typically involved factual information such as numbers, time, phone numbers, or quantities, making the distractors easier to construct in a homogeneous and competitive form. In contrast, the largest difference appeared in Simple Inference, where ChatGPT scored 2.00 and NotebookLM scored 3.33. This indicates that several ChatGPT inference items did not truly require inference, as the answers could be identified explicitly from the transcript. NotebookLM was more effective in constructing options that required test-takers to infer meaning from the listening context.

Table 7. Scores by Test Objectives

Test Objective	ChatGPT	NotebookLM	Comparative Finding
Specific Information	4.00	4.00	Both excellent
Detail Comprehension	3.33	3.00	ChatGPT slightly stronger
Speaker Intention/Function	3.00	3.67	NotebookLM stronger
Main Idea/Gist	2.33	3.00	NotebookLM stronger
Simple Inference	2.00	3.33	NotebookLM much stronger

To sum up, the quantitative findings indicate that both ChatGPT and NotebookLM demonstrated strong performance, although their strengths varied across test objectives. Both tools performed equally well in identifying specific information, each achieving the highest score of 4.00. ChatGPT showed a slight advantage in detail comprehension, with a score of 3.33 compared to NotebookLM's 3.00. In contrast, NotebookLM outperformed ChatGPT in higher-level comprehension areas, particularly speaker intention/function, main idea/gist, and simple inference. The most notable difference was observed in simple inference, where NotebookLM scored 3.33 while ChatGPT reached only 2.00, suggesting that NotebookLM was more effective in interpreting implied meanings. These results suggest that while ChatGPT was slightly stronger in processing detailed information, NotebookLM demonstrated greater effectiveness in interpretive and inferential listening comprehension tasks.

Qualitative Findings

To further explain the numerical patterns, the qualitative analysis examines specific item examples from both tools and reveals that NotebookLM produced more plausible distractors than ChatGPT. Although both tools generated distractors that reflected basic principles of multiple-choice item construction, their outputs differed in quality and function, with ChatGPT tending to produce formally consistent distractors, whereas NotebookLM more frequently generated distractors that were grounded in the listening text and more competitive for test-takers with partial comprehension.

In the ChatGPT-generated items, the main strength was formal consistency. Many distractors were grammatically parallel, similar in length and form, and belonged to the same semantic category as the correct answer. For example, in the item asking about Marina's phone number, the distractors differed only in the arrangement of digits, making them plausible for students who misheard or misordered the number. Similarly, options such as "Level 1 Teacher's Book," "Level 2 Student's Book," and "Level 2 Teacher's Book" showed good homogeneity because they manipulated related information within the same category. This pattern explains ChatGPT's relatively good performance in

grammatical parallelism, homogeneity, and similarity of length and form.

However, ChatGPT's formal consistency did not always result in effective distractor function. Several distractors were only topically related to the transcript but did not represent likely misunderstanding of the audio. Options such as "Buy new toys," "student café," "meeting friends on campus," and "book writer" were relatively easy to eliminate because they were not sufficiently grounded in the specific listening input. This finding is consistent with the lower quantitative scores for competitiveness and dependence on listening comprehension. In addition, the repeated placement of correct answers in option A created a technical clue that could reduce item validity. Thus, ChatGPT-generated distractors require revision not only in semantic plausibility but also in answer-key distribution.

NotebookLM demonstrated more plausible distractor construction. Its distractors were more often based on specific audio details, especially numbers, locations, class information, and objects. For instance, "Room 13 on the first floor," "Room 13 on the second floor," and "Office 7B on the first floor" were competitive because they manipulated closely related details from the transcript. These options required test-takers to distinguish room number and floor information, making them more dependent on careful listening.

NotebookLM also produced distractors that reflected partial comprehension or possible mishearing. In the item asking what students should bring to the next class, options such as "A photocopy of the student's book," "A copy of the teacher's book," and "A photocopy of the level one book" exploited subtle contrasts in the audio. Such distractors are pedagogically more plausible because they represent listening-based misconception rather than arbitrary incorrect information. This pattern explains NotebookLM's higher scores in plausibility, competitiveness, and dependence on the listening text.

Nevertheless, NotebookLM's distractors were not uniformly plausible. Some options, such as "A new laptop," "A set of headphones," and "To invite students to a Friday party," were contextually distant or easily eliminated through common sense. These weaknesses indicate that NotebookLM still requires expert review despite its better overall performance.

Table 8. Comparison of Distractor Plausibility Generated by ChatGPT and NotebookLM

Aspect	ChatGPT	NotebookLM
Main strength	Formally consistent and grammatically parallel distractors	More plausible, competitive, and audio-grounded distractors
Strongest categories	Grammatical parallelism, homogeneity, similarity of length and form	Plausibility, competitiveness, dependence on listening text, and fewer technical clues
Main weakness	Several distractors are only topically relevant, less competitive, and affected by answer-key patterns	Some distractors are still contextually distant, unnatural, or easy to eliminate
Overall interpretation	Useful for maintaining option form, but requires substantial revision for distractor function	More plausible in supporting listening-based distractor quality, but still requires expert revision

These findings reveal that NotebookLM produced more functional distractors for listening MCQ test items, particularly as its options were more consistent in meeting the criteria of plausibility, competitiveness, and absence of technical clues. Nevertheless, NotebookLM still requires revision in several items where the distractors were loosely connected to the audio input or insufficiently relevant. ChatGPT, although significantly plausible in factual-detail items, requires more substantial revision, especially in answer-key distribution, distractor quality for main idea and inference items, and the removal of options that can be easily eliminated through common sense. Overall, the quantitative findings indicate that NotebookLM generated more plausible distractors across most indicators, with the exception of Detail Comprehension items

where ChatGPT performed slightly better. Thus, both item sets should undergo expert review and revision before being administered as final listening test instruments.

The synthesis of findings suggests that stem clarity and distractor plausibility represent two related but distinct dimensions of listening MCQ quality. In terms of stem clarity, both ChatGPT and NotebookLM produced highly understandable and focused stems, with ChatGPT showing a slight advantage in brevity, directness, and linguistic simplicity. However, clear stems did not automatically lead to equally effective distractors. The distractor analysis showed that NotebookLM performed better in generating options that were more plausible, competitive, and closely grounded in the listening input, particularly in interpretive and inferential items. ChatGPT, by contrast, often produced formally consistent options but some distractors were too easily eliminated because they were weakly connected to the audio context or could be answered through general knowledge. Thus, the integrated findings indicate that ChatGPT is more effective in producing simple and accessible question stems, whereas NotebookLM is stronger in constructing listening-based distractors. These patterns suggest that AI-generated listening MCQ items still require human review, especially to balance stem clarity with distractor quality and to ensure that the items validly assess listening comprehension rather than test-wise guessing.

Discussions

Stem Clarity of ChatGPT and NotebookLM Listening MCQ Test Items

The findings indicate that both ChatGPT and NotebookLM generated listening MCQ items with excellent stem clarity, with an overall mean of 3.87 out of 4.00 and no item receiving a poor or failing score. This result suggests that both tools are capable of producing stems that are understandable, focused, and linguistically appropriate for beginner-level EFL learners, consistent with Aryadoust et al. (2024) and Chun and Barley (2024), who found that AI-generated MCQs demonstrate comparable surface-level quality to human-developed items. This finding also aligns with Runge et al. (2024), who demonstrated that recent advances in generative AI can be successfully applied to the task of automated item generation (AIG) to produce conversational content and listening items, suggesting that LLMs have matured sufficiently to handle basic item construction demands in language assessment.

However, the two tools achieved stem clarity through different strengths. ChatGPT produced stems that were more concise and direct, which is particularly important in listening assessment where cognitive load must be managed carefully. As Buck (2001) emphasizes, listening test stems should not introduce additional reading burden that may interfere with the construct being measured. NotebookLM, on the other hand, tended to produce longer and more contextually embedded stems that were more independent from the answer options, reflecting its design as a source-grounded tool. While this contextual specificity is pedagogically valuable, longer stems may increase processing demand for A1-level learners, as Fulcher (2024b) cautions against stem complexity that introduces construct-irrelevant difficulty.

Both tools showed lower performance on Speaker Intention/Function and Simple Inference items, with a combined mean of 3.67, compared to a perfect score of 4.00 on Specific Information and Detail Comprehension items. This pattern is theoretically expected, as inferential and pragmatic stems require more careful formulation to avoid revealing the intended answer through wording choices. The use of direct quotations in speaker-function stems, such as "Quiet, please" and "Please arrive on time," by both tools introduces unintended verbal cues that may allow test-wise learners to respond correctly without processing the listening input. This finding highlights the need for human expert review, particularly for stems targeting higher-order listening skills. As Jang and Sawaki (2025) caution, the benefits of AI in assessment must be evaluated against concerns related to construct validity, scoring transparency, and fairness, suggesting that AI-generated items targeting pragmatic and inferential skills demand especially

rigorous human scrutiny to preserve construct representativeness.

Distractor Plausibility of ChatGPT and NotebookLM Listening MCQ Test Items

The analysis reveals a more notable difference between the two tools in distractor plausibility, with NotebookLM obtaining a mean of 3.25 compared to ChatGPT's 2.63. This finding suggests that while both tools can produce acceptable stems, their capacity to construct effective distractors differs substantially. Haladyna (2004) argues that distractor quality is central to the validity of MCQ assessment, as weak distractors reduce the item's capacity to differentiate learners who genuinely comprehend the input from those who rely on guessing or test-wise strategies. The present findings confirm that this dimension remains more challenging for general-purpose AI tools.

ChatGPT demonstrated strength in formal consistency, producing distractors that were grammatically parallel, similar in length, and belonging to the same semantic category as the correct answer. However, several distractors were only loosely connected to the specific listening content and could be eliminated without careful audio processing, as reflected in its low score of 1.5 for the absence of technical clues and 2.0 for competitiveness. Shin et al. (2019) stress that effective distractors must be grounded in the likely errors or misconceptions that test-takers experience when processing spoken input, rather than being merely plausible on a general semantic level. ChatGPT's distractors, in many cases, did not meet this standard for inference and main idea items. As noted by Kiyani et al. (2025), ChatGPT's limited contextual awareness produces less plausible distractors that often overlook the specific misconceptions required for nuanced listening assessments.

NotebookLM produced more competitive and audio-grounded distractors, particularly by exploiting subtle contrasts within the transcript, such as variations in room numbers, floor levels, time references, and book titles. These options required test-takers to distinguish closely related audio details, which aligns with the principle that effective listening distractors should reflect partial comprehension or likely mishearing (Anckar, 2011). Nevertheless, NotebookLM's performance was not uniformly strong, as some items still contained distractors that were contextually distant or eliminable through common sense, indicating that source-grounding does not fully guarantee distractor quality.

Conclusions

This study compared the quality of MCQ listening test items generated by ChatGPT and NotebookLM, focusing on stem clarity and distractor plausibility. The findings demonstrate that both tools are capable of producing clear and focused stems at an excellent level, suggesting that modern AI tools have internalized fundamental item-writing conventions for straightforward listening objectives. However, the two tools showed meaningfully different strengths: ChatGPT was more effective in generating concise and direct stems, while NotebookLM proved stronger in constructing plausible, competitive, and audio-grounded distractors, particularly for higher-order items such as speaker intention, main idea, and simple inference. Critically, stem clarity alone did not guarantee overall item quality, as ChatGPT's formally consistent distractors were frequently eliminable through world knowledge or contextual guessing rather than careful listening, and its systematic answer-key patterning introduced test-wise clues that further undermined item validity.

These findings suggest that neither tool is sufficient as a standalone solution for developing valid EFL listening assessments. Rather, their complementary strengths point toward a collaborative workflow in which ChatGPT generates initial stems and NotebookLM produces transcript-grounded distractor sets, with human expert review as an indispensable final stage, particularly for items targeting inferential and pragmatic listening skills. For EFL educators in higher education, this study offers evidence that AI tools can meaningfully reduce the burden of initial item drafting without replacing the assessment literacy that professional item development demands. Future research should examine the validity of AI-generated listening items through empirical item analysis, explore the role of prompt engineering in improving

distractor quality, and investigate how different source text types affect the output quality of source-grounded AI tools.

References

Alhazmi, E., Sheng, Q. Z., Zhang, W. E., Zaib, M., & Alhazmi, A. (2024). Distractor generation in multiple-choice tasks: A survey of methods, datasets, and evaluation. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 14437–14458. https://doi.org/10.18653/v1/2024.emnlp-main.799

Anckar, J. (2011). *Assessing foreign language listening comprehension by means of the multiple-choice format: Processes and products* (Issue 159). University of Jyväskylä.

Aryadoust, V., Zakaria, A., & Jia, Y. (2024). Investigating the affordances of OpenAI's large language model in developing listening assessments. *Computers and Education: Artificial Intelligence*, *6*, 100204.

Buck, G. (2001). *Assessing listening*. Cambridge University Press.

Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition*, *7*(3), 323–331.

Chun, J. Y., & Barley, N. (2024). *A comparative analysis of multiple-choice questions: ChatGPT-generated items vs. human-developed items BT - Exploring artificial intelligence in applied linguistics* (C. A. Chapelle, G. H. Beckett, & J. Ranalli (eds.)).

Circi, R., Hicks, J., & Sikali, E. (2023). Automatic item generation: Foundations and machine learning-based approaches for assessments. *Frontiers in Education*, *8*.

Diyab, A., Frost, R. M., Fedoruk, B. D., & Diyab, A. (2025). Engineered prompts in ChatGPT for educational assessment in software engineering and computer science. *Education Sciences*, *15*(2), 156. https://doi.org/10.3390/educsci15020156

Education, G. for. (2024). *NotebookLM for education: AI-powered research and thinking partner*. Google LLC. https://edu.google.com/intl/ALL_us/ai-notebooklm/

Fulcher, G. (2024a). *Practical Language Testing* (2nd Editio). Routledge.

Fulcher, G. (2024b). *Practical Language Testing* (2nd ed.). Routledge.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Routledge.

Jang, E. E., & Sawaki, Y. (2025). Advancing language assessment for teaching and learning in the era of the artificial intelligence (AI) revolution: Promises and challenges. *Language Testing*, *42*(4), 361–368. https://doi.org/10.1177/02655322251348685

Kiyani, A., Hanif, F., Muhammad, M., Iqbal, S., Zaib, N., Bashir, U., & Ali, K. (2025). Benchmarking ChatGPT-generated multiple-choice questions against faculty-authored items in dental education. *Scientific Reports*, *15*, 44805. https://doi.org/10.1038/s41598-025-28492-7

May, T. A., Fan, Y. K., Stone, G. E., Koskey, K. L. K., Sondergeld, C. J., Folger, T. D., Archer, J. N., Provinzano, K., & Johnson, C. C. (2025). An effectiveness study of generative artificial intelligence tools used to develop multiple-choice test items. *Education Sciences*, *15*(2), 144. https://doi.org/10.3390/educsci15020

144

Runge, A., Attali, Y., LaFlair, G. T., Park, Y., & Church, J. (2024). A generative AI-driven interactive listening assessment task. *Frontiers in Artificial Intelligence*, *7*, 1474019.

Shin, D., Kwon, S. K., & Lee, Y. (2025). Examining the efficacy of generative artificial intelligence in item generation: comparative analysis of human-developed and AI-generated reading tests. *Education and Information Technologies*, *30*(16), 23981–24007.

Shin, J., Guo, Q., & Gierl, M. J. (2019). Multiple-choice item distractor development using topic modeling approaches. *Frontiers in Psychology*, *10*, 825.

Trinh, A. H. (2023). Classroom-based assessment in TEFL as a form of formative assessment: Enquiries into theory and practice. *Working Papers in Language Pedagogy*, *18*.

Wang, Y., & Meng, Y. (2026). Optimizing distractor quality in a locally developed second language listening test: Integrating generative AI and psychometric methods. *Language Testing*, *43*(2), 141–164.

Xu, H., Gan, W., Qi, Z., Wu, J., & Yu, P. S. (2024). Large language models for education: A survey. *ArXiv Preprint ArXiv:2405.13001*.

Yanagawa, K., & Green, A. (2008). To show or not to show: The effects of item stems and answer options on performance on a multiple-choice listening comprehension test. *System*, *36*(1), 107–122.