# ACA-Bot: Evaluating Pedagogical Dialogue Generation in a Spoken LLM-Based Oral Examiner

**Julie Bauer and Krist Shingjergji and Aki Härmä**
Department of Advanced Computing Sciences (DACS)
Maastricht University
Maastricht, The Netherlands

## Abstract

Oral exams can capture students' conceptual understanding, but they are difficult to scale. To improve on scalability, this paper presents ACA-Bot, a speech-and-LLM-based oral examiner for introductory statistics. ACA-Bot uses a cascaded architecture combining speech recognition, LLM-based dialogue control, and text-to-speech, and is designed around pedagogical dimensions such as diagnostic questioning, adaptivity, neutrality, content alignment, coherence, consistency, and feedback quality. We evaluated the system through a user study, LLM-as-judge scoring, and qualitative conversation analysis. Results suggest that ACA-Bot is promising for formative assessment and exam preparation, particularly when pedagogical quality can be guided by prompts, rubrics, and predefined topic structures. The system performed less consistently on diagnostic and adaptive questioning, sometimes moving on before students had fully demonstrated their understanding. Speech-related problems, including interruptions and recognition errors, also affected interaction quality. These findings show the potential and limitations of spoken LLM-based assessment.

## 1 Introduction

Oral exams are a valuable form of assessment because they allow students to explain their understanding spontaneously and allow examiners to ask follow-up questions when answers are incomplete or unclear (Theobold, 2021). This type of assessment might, therefore, capture a student's understanding more fully than written exams (Theobold, 2021; Sato et al., 2019). However, oral exams are resource-intensive and thus difficult to scale (Bayley et al., 2024; Memon et al., 2010). This limits their use, particularly in large courses.

Conversation-Based Assessment (CBA) tries to address scalability by using one or multiple AI agents to conduct assessment dialogues (Zapata-Rivera et al., 2023). Recent studies have used LLMs to conduct text-based, multi-turn assessments, ask follow-up questions, and assess students' understanding from their responses (Bergerhoff et al., 2025; Hou et al., 2025). However, these studies focus on text-based assessment dialogues. They do not examine what happens when the interaction is spoken, as in an oral exam. They also focus less on how oral assessment principles can be used to design and evaluate automated oral examiners.

In an assessment context, pedagogical considerations become particularly important. An examiner should collect sufficient and representative evidence of a student's understanding (Kane, 2013; Joughin, 1998; Pearce and Chiavaroli, 2020; Memon et al., 2010), and, for formative purposes, provide high-quality feedback after the assessment (Hattie and Timperley, 2007; Shute, 2008). In a spoken LLM-based system, these principles must be reflected in the examiner's questions, follow-up behavior, dialogue control, and feedback.

Therefore, this paper presents ACA-Bot, where ACA stands for Automatic Conversational Assessment, extending prior text-based work on automatic conversational assessment (Bergerhoff et al., 2025). It is a speech-and-LLM-based oral exam system designed around pedagogical principles for oral exams. Introductory statistics was used as the assessment domain. The system uses speech recognition, text-to-speech, LLM-based response generation, diagnostic state tracking, and rule-based dialogue control to conduct structured oral exams.

This study makes three contributions. It defines seven dimensions for designing and evaluating generated assessment dialogue; presents ACA-Bot's structured dialogue architecture and diagnostic state tracking; and evaluates the system using user and simulated conversations, LLM-based ratings, and qualitative conversation analysis. The findings identify strengths and remaining challenges in spo-



*Proceedings of the 19th International Natural Language Generation Conference: Long Papers*, pages 1276–1301
October 17-21, 2026 ©2026 Association for Computational Linguistics

ken CBA.

## 2 Background and Related Work

### 2.1 Pedagogical Perspectives on Oral Exams

An assessment system for oral exams should assess students in a pedagogically aligned way. Oral assessment literature describes validity, fairness, and reliability as important principles. Validity concerns whether an assessment measures the construct it is intended to measure, while content validity concerns whether the questions sufficiently represent the relevant domain (Memon et al., 2010). Questions should therefore remain aligned with the intended learning goals, and the examiner should collect sufficient evidence to judge a student's understanding (Kane, 2013).

Fairness requires students to receive comparable opportunities to demonstrate their knowledge. Since oral exams are interactive, differences in follow-up questions or the amount of support provided may advantage some students over others. Reliability concerns whether students with similar levels of understanding are assessed in a similar way. Variation in examiner behavior can therefore affect both fairness and reliability, but for different reasons (Memon et al., 2010). A pedagogically grounded oral exam should balance structure with flexibility: it should be sufficiently controlled to support comparable assessment, but should still allow an examiner to respond to the student's reasoning (Joughin, 1998).

How an examiner asks questions is also important. Clarifying and probing questions can make students' reasoning more visible, while leading questions may guide students towards an expected answer (Pearce and Chiavaroli, 2020). Examiner prompts should therefore be neutral and avoid hints, expected answers, excessive praise, or discouragement. When an oral exam has a formative purpose, feedback should also be accurate, specific, clear, balanced, and connected to what the student demonstrated. It should identify relevant strengths and weaknesses and provide actionable recommendations for improvement (Hattie and Timperley, 2007; Shute, 2008). These principles suggest that an LLM-based examiner should be guided by explicit pedagogical behaviors rather than conduct an exam freely (Beale, 2025; Sonkar et al., 2024).

### 2.2 Conversation-Based Assessment

Conversation-Based Assessment (CBA) uses dialogue with artificial agents to collect evidence of what a student knows or can do (Zapata-Rivera et al., 2023). It builds on conversational tutoring, but its main purpose is assessment rather than teaching. The agent should therefore probe what the student already knows without giving away the answer. Follow-up questions can provide further opportunities to explain an incomplete response (Zapata-Rivera et al., 2024).

Earlier CBA systems used predefined dialogue paths and response categories, which supported controlled dialogue and scoring but required substantial knowledge engineering (Zapata-Rivera et al., 2023). Zapata-Rivera et al. (2024) translated Evidence-Centered Design principles into an LLM prompt, while Forsyth et al. (2024) found that the LLM produced more flexible conversations but that these conversations could be more difficult to score. Bergerhoff et al. (2025) developed a text-based LLM assessment system that asked follow-up questions and generated grades and feedback, but sometimes continued probing for too long or moved on too early. Hou et al. (2025) used a multi-agent architecture to provide more control over the assessment, although the student-facing agents sometimes used leading questions and their classifications showed only partial agreement with human coding. Together, these studies show that LLMs can support flexible assessment conversations, but also create challenges around probing, dialogue progression, neutrality, and control. None of these studies evaluated these behaviors in a real-time spoken oral exam.

### 2.3 Speech-Based Educational Dialogue Systems

Studies comparing spoken and typed tutoring found no consistent difference in learning outcomes, but showed that students form their answers differently (D'Mello et al., 2011; Litman et al., 2006). Typed answers can be planned and revised, while spoken answers are produced more spontaneously. Speech has been explored in tutoring systems, where users can explain their reasoning through multi-turn dialogue and receive hints or explanations (Forbes-Riley and Litman, 2011; Cole et al., 2018; Geris and Alce, 2026). It has been examined less in Conversation-Based Assessment, where the goal is to collect evidence of what a student can demon-

strate independently.

Although tutoring systems are similar to CBA systems, speech-technology problems may affect them differently. Speech-recognition errors, latency, and incorrect turn-taking can disrupt both settings (Chen et al., 2026; Ji et al., 2024; Castillo-López et al., 2025). In CBA, these problems may also change or limit the evidence collected about a student's understanding. A recognition error may change the meaning of an answer, while an interruption may prevent a student from fully explaining it.

## 3 Methodology

### 3.1 Pedagogical Dimensions

Pedagogical quality was operationalized through seven observable examiner behaviors intended to support valid, fair, and reliable assessment. These behaviors were derived from the oral-assessment principles described in Section 2.1 and used to guide ACA-Bot's design:

1. **Diagnostic questioning.** Whether the examiner asks enough relevant questions and gives users sufficient opportunities to demonstrate their understanding.
2. **Adaptive questioning.** Whether follow-up questions respond to the previous answer and target the specific part that was unclear or incomplete.
3. **Neutrality.** Whether the examiner avoids teaching, hints, leading questions, expected answers, excessive reassurance, praise, or discouragement.
4. **Prompting consistency.** Whether users with comparable answers receive comparable opportunities to clarify or expand their responses.
5. **Content alignment.** Whether questions remain focused on the intended statistical concepts and learning goals.
6. **Assessment coherence.** Whether the questions and topic transitions form a logical and consistent assessment dialogue.
7. **Feedback quality.** Whether the final feedback is accurate, specific, clear, balanced, actionable, and aligned with the user's performance.

### 3.2 System Architecture and Assessment Design

ACA-Bot is a speech-based oral-exam system with separate components for automatic speech recognition (ASR), dialogue management with LLMs, and text-to-speech (TTS), shown in Figure 1. We used a cascaded architecture because it gave us more control over the conversation flow and made it possible to inspect the individual system components (Chen et al., 2026; Liu et al., 2025).

Users interacted with ACA-Bot through a web-based interface adapted from LiveKit's open-source *agent-starter-react* template (LiveKit, n.d.). AssemblyAI's (AssemblyAI, n.d.) *Pro 3* model was used for speech transcription and turn detection, OpenAI's (OpenAI, 2026) *GPT-5 mini* was used for the LLM components, and ElevenLabs' (ElevenLabs, n.d.) *Flash V2* model generated speech. The system logged user and examiner turns, dialogue states, task and subtopic transitions, feedback, and latency.

Dialogue management was done by an examiner LLM, a state LLM, and a rule-based dialogue controller. A feedback LLM generated feedback after the exam. Dialogue management was designed around the pedagogical dimensions defined in Section 3.1, particularly examiner neutrality, diagnostic questioning, adaptivity, and evidence-based feedback. The prompt instructions and dialogue-control rules are summarized in Appendix A.3–A.4.

The **examiner LLM** (labeled "LLM Response Generation" in Figure 1) generated user-facing responses during the exam. It was instructed to act as an examiner rather than a tutor: to ask brief, non-leading questions, avoid teaching or revealing answers, and base follow-up questions on the user's previous response. Depending on the action selected by the dialogue controller, it was prompted to introduce a new subtopic, ask a targeted follow-up question, or move to the next subtopic.

A Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) pipeline retrieved task-relevant content from a manually verified introductory-statistics reference document and provided it to the examiner LLM. The task and subtopics were also provided. This was done to support content alignment.

The **state LLM** (labeled "State Classification" in Figure 1) was instructed to track the diagnostic evidence collected for each subtopic by assigning one

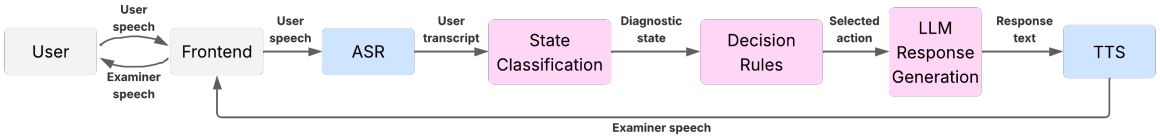



Figure 1: Overview of the system architecture and information flow.

of four states: *diagnostically resolved*, *partially reasoned*, *non-diagnostic*, or *unseen*. These states represented whether a response was judged to provide sufficient evidence of understanding or a misconception, require further clarification, provide no usable evidence, or not yet have been addressed.

The **dialogue controller** (labeled "Decision Rules" in Figure 1) used these classifications to determine dialogue progression. Under the predefined rules, resolved subtopics were completed, partially complete responses triggered a targeted follow-up, non-diagnostic subtopics were closed, and unseen subtopics received an opening question. The system moved to the next task once all subtopics had been completed or closed. Follow-up and turn limits were included to reduce repetition and prevent the dialogue from becoming stuck.

The **feedback LLM** was instructed to generate concise post-exam feedback based on evidence from the conversation. It was prompted to identify demonstrated strengths, gaps, and misconceptions and to provide specific recommendations for improvement without assigning a grade.

The oral exam was designed to make users' understanding and misunderstandings of basic statistical concepts visible through the conversation. The assessment contained three tasks: hypothesis testing, mean and median, and variance. These topics were selected because they are common in introductory statistics courses and allow questions about definitions, interpretation, and reasoning.

Each task was divided into subtopics covering different parts of the statistical concept. These subtopics gave the examiner a clearer questioning structure and provided context used to decide when sufficient diagnostic evidence had been collected. All participants received the same tasks and subtopics, while follow-up questions were adapted to their responses. This was intended to support consistency across participants (Memon et al., 2010).

Within each task, earlier subtopics focused on definitions and basic understanding, while later subtopics required more interpretation, reasoning, or application. This broadly followed the cognitive progression described in Bloom's revised taxonomy (Anderson and Krathwohl, 2001). The complete rubric and subtopics are provided in Appendix A.2, Table 3.

Users received spoken feedback only after completing the oral exam. Feedback was delayed so that it could not influence later answers and affect the assessment. At the same time, it allowed the system to show users what they appeared to understand, where their reasoning was incomplete, and which misconceptions were visible in the conversation.

### 3.3 User Study and Survey Evaluation

A user study was conducted to evaluate how users experienced ACA-Bot. Participants were recruited through the researchers' network and completed a short spoken mock exam followed by a Qualtrics survey covering pedagogical quality, speech interaction, feedback quality, and suitability for learning and assessment. Participants were encouraged to provide honest feedback and were informed that both positive and negative responses could guide system development. Most items used five-point Likert scales and were summarized using frequencies, percentages, medians, and standard deviations; open-ended responses were analyzed for recurring themes. The complete questionnaire appears in Appendix B.

### 3.4 Synthetic Student Simulation

To obtain additional controlled conversational data, we simulated students using an LLM. Each simulated student completed a full text-based oral exam with ACA-Bot, allowing the dialogue policy to be evaluated without interference from speech-recognition or turn-taking errors. Before each simulation, the student was assigned a knowledge profile. For each subtopic, the student's knowledge state was independently sampled as correct, partial, wrong, or no knowledge, with probabilities

.35, .40, .15, and .10, respectively. The profile was included in the prompt because, without explicit knowledge constraints, the LLM tended to provide overly correct answers. Because subtopics were sampled independently, the simulation did not model dependencies between related concepts.

The student LLM was prompted to reveal its reasoning gradually. It could clarify or add information after follow-up questions, but only within the limits of its assigned knowledge profile. The simulations were controlled tests and were not meant to fully represent real students, since LLM responses can be less variable and more fluent than human responses.

### 3.5 LLM-as-Judge Evaluation

The conversations were evaluated using DeepEval (Confident AI, 2026). Custom metrics were created for six of the seven pedagogical dimensions. Prompting consistency was not scored because it requires comparing multiple conversations, instead of evaluating the quality of one conversation. It was therefore only examined descriptively through the qualitative analysis and should be considered under-evaluated in the present study. Each metric defined criteria for rewarding or penalizing specific examiner behaviors and included descriptive score anchors ranging from 0 to 1.

Three LLMs independently evaluated the conversations: GPT-4.1, Mistral Medium 3.5, and Claude Sonnet 4.5. The three judge scores were averaged for each conversation and pedagogical dimension, after which the resulting combined scores were summarized across conversations using the mean (M) and standard deviation (SD). For content-sensitive dimensions, including content alignment and diagnostic questioning, the rubric and reference material were provided as additional evaluation context.

Inter-LLM agreement was assessed using the mean absolute pairwise difference (MAD) between the three judges' scores for each conversation and dimension. We used MAD because several dimensions showed limited between-conversation variance (see Table 1), which makes measures such as ICC and Spearman correlation less informative.

The judges were explicitly instructed to evaluate examiner behavior rather than student performance. Incorrect, incomplete, vague, or uncertain user answers were therefore not penalized themselves, but were used as evidence for determining whether the examiner responded appropriately. The complete criteria and score anchors are provided in Appendix F.

### 3.6 Qualitative Conversation Analysis

The first author performed a structured qualitative analysis of all completed user conversations. Each conversation was examined turn by turn along the pedagogical dimensions defined in Section 3.1. The analysis focused on examples in which ACA-Bot's behavior aligned with or deviated from these dimensions, and representative excerpts were selected to illustrate recurring strengths and weaknesses.

## 4 Results

### 4.1 Data Sets

In this study, we collected two oral exam datasets: the user-study conversations and the simulated student interactions. Of 49 logged conversations, 34 were completed and included in the main analyses. The 15 incomplete conversations were excluded but remain an important system limitation.

During data collection, the speech configuration was adjusted because users experienced problems with interruptions. The initial speech configuration used AssemblyAI's voice activity detection for turn-taking and included 13 completed conversations. In the updated configuration, ACA-Bot waited four seconds after a user's response before moving on; this group included 21 conversations. Since both groups were relatively small and participants were not randomly assigned, comparisons between the configurations are only used to describe patterns and not to make strong statistical claims.

The second dataset contained 40 controlled conversations between simulated students and ACA-Bot. Simulated students produced more regular and shorter answers than users, while examiner response lengths were similar across datasets. Complete conversation descriptives are provided in Appendix D.

### 4.2 Participant Characteristics

Most participants were current higher-education students ($n = 26$, 76.5%), and 32 participants (94.1%) had received at least some formal statistics education. Participants differed in their statistics confidence and in how recently they had studied statistics. Most reported advanced English proficiency ($n = 24$, 70.6%), while 21 participants

(61.8%) had previous experience with oral or spoken assessment. Attitudes toward oral assessment and AI in education were generally positive. Complete participant characteristics are provided in Appendix C.

### 4.3 Pedagogical Quality

Participants generally evaluated ACA-Bot's pedagogical behavior positively. Thirty-three participants agreed or strongly agreed that ACA-Bot assessed rather than taught them, and the same number agreed that it focused on important statistical ideas. Twenty-nine agreed that it avoided giving away the correct answer, 26 found questions clear, and 24 found follow-up questions relevant.

Responses were less positive regarding opportunities to demonstrate reasoning. Fourteen participants agreed that ACA-Bot asked them to explain their reasoning, while 15 disagreed. Sixteen agreed that the exam reflected their actual understanding, compared with 12 who disagreed. Most notably, 20 participants disagreed that they had a fair opportunity to clarify, correct, or expand their answers, while only seven agreed. Thus, although ACA-Bot was generally perceived as relevant and assessment-focused, it did not always allow users to demonstrate their understanding fully.

### 4.4 Feedback Quality

Feedback was evaluated positively across all items. Thirty-three participants agreed or strongly agreed that it was clear and understandable. Thirty-two agreed that it was consistent with the exam, specific about strong and weak answers, and balanced between strengths and weaknesses. Twenty-seven agreed that it accurately identified their understanding and misunderstandings, while 29 found that it provided clear guidance for improvement.

### 4.5 Speech Modality

The speech interaction received mixed evaluations. Twenty-nine participants agreed or strongly agreed that it made the interaction feel like an oral exam, and 27 agreed that ACA-Bot responded quickly enough after they finished speaking.

Turn-taking was the clearest problem. Twenty-six participants indicated that ACA-Bot interrupted them while speaking, and 20 indicated that speech-recognition or audio problems affected their ability to express their answers. Sixteen disagreed that speaking helped them explain their reasoning naturally, and the same number disagreed that speech made demonstrating understanding easier or that the interaction felt like a natural conversation.

Responses under the updated configuration were more positive for speech recognition, interruptions, naturalness, and ease of demonstrating understanding. Nevertheless, the main speech-related difficulties were not fully resolved.

### 4.6 Suitability for Learning and Assessment

Participants primarily considered ACA-Bot useful for practice and formative assessment. Twenty-nine participants agreed or strongly agreed that it would be useful for practising statistics before an exam, and the same number agreed that it could identify strengths and weaknesses. Twenty-four would recommend it as a study tool, while 26 considered it a useful additional assessment tool.

Participants were more hesitant about formal assessment. Seventeen disagreed that they would feel comfortable taking an oral exam conducted by ACA-Bot. Only 11 agreed that it provided a fair indication of their understanding, and 12 would accept the conversation as evidence in a graded assessment. Views on replacing parts of a traditional oral exam were divided: 14 agreed, 12 disagreed, and eight were neutral. Twenty-six disagreed that they would prefer an ACA-Bot-style oral exam over a written exam.

### 4.7 Open-Ended Survey Responses

Thirty-one participants answered at least one open-ended survey question. The most common concern was rushed turn-taking or interruptions ($n = 21$). Participants described pauses, hesitation, or thinking time being interpreted as the end of an answer, which prevented them from fully demonstrating their understanding. Participants also requested more control over the interaction ($n = 12$), including ways to indicate that they had finished, repeat a question, revise an answer, or return to an earlier question. Thirteen participants wanted deeper or more adaptive questioning, particularly through targeted follow-ups, practical examples, and questions that probed their reasoning.

Participants commonly viewed ACA-Bot as useful for practice, exam preparation, and identifying knowledge gaps ($n = 18$), and 13 described the final feedback as clear, useful, or accurate. However, 10 wanted more detailed, example-based, or actionable feedback. Sixteen participants expressed reservations about graded assessment, particularly because interruptions, recognition errors, and limited

opportunities to correct answers could affect fairness and accuracy. The complete thematic breakdown is provided in Appendix E.2.

### 4.8 LLM-as-Judge Results

Table 1 presents the combined LLM-as-judge results for the user and simulated conversations. For the user conversations, content alignment received the highest score ($M = .97$, $SD = .06$), followed by feedback quality ($M = .92$, $SD = .06$) and neutrality ($M = .91$, $SD = .16$). Diagnostic questioning ($M = .73$, $SD = .20$) and adaptive questioning ($M = .68$, $SD = .26$) received lower and more variable scores. ACA-Bot therefore generally remained aligned with the intended content and produced positively evaluated feedback, but behaved less consistently when probing and adapting to users' answers.

The simulated conversations received higher scores on most dimensions. The largest difference was found for adaptive questioning, which increased from .68 for the user conversations to .94 for the simulations. Assessment coherence increased from .78 to .91, while diagnostic questioning increased from .73 to .82. This suggests that ACA-Bot's dialogue behavior was more consistent under controlled text-based conditions. However, the simulated students also produced more regular and constrained answers than actual users.

Across the judges' reasoning, lower scores were mainly linked to missed or generic probing opportunities and failures to repair interrupted or disrupted interactions. Agreement between the LLM judges was highest for content alignment and feedback quality. In contrast, there was more disagreement for diagnostic and adaptive questioning, particularly in the user conversations.

This disagreement was not evenly distributed across the three judges. Claude consistently scored diagnostic questioning, adaptive questioning, and neutrality lower than GPT and Mistral in the user conversations (e.g., diagnostic questioning: .56 for Claude versus .82 and .81 for GPT and Mistral; full breakdown in Appendix F, Table 16). GPT and Mistral scored these dimensions more similarly to each other. This suggests a judge-specific scoring tendency rather than general disagreement across all three judges.

Overall, the judges were more consistent when evaluating relatively well-defined behaviors, such as content alignment. They differed more when judging whether enough diagnostic evidence had been collected and whether follow-up questions were sufficiently adaptive. The combined scores should therefore be interpreted with some caution. Complete model-specific scores, pairwise agreement results, and recurring patterns in the judges' reasoning are provided in Appendix F.

Table 1: Combined LLM-as-judge scores and inter-LLM score agreement.

| | User | | | Simulated | | |
|---|---|---|---|---|---|---|
| **Dimension** | **M** | **SD** | **MAD** | **M** | **SD** | **MAD** |
| Diagnostic questioning | 0.73 | 0.20 | 0.21 | 0.82 | 0.10 | 0.17 |
| Adaptive questioning | 0.68 | 0.26 | 0.26 | 0.94 | 0.08 | 0.08 |
| Neutrality | 0.91 | 0.16 | 0.15 | 0.95 | 0.13 | 0.11 |
| Content alignment | 0.97 | 0.06 | 0.04 | 0.98 | 0.04 | 0.03 |
| Assessment coherence | 0.78 | 0.18 | 0.16 | 0.91 | 0.10 | 0.12 |
| Feedback quality | 0.92 | 0.06 | 0.06 | 0.92 | 0.05 | 0.06 |

*Note.* MAD = mean absolute pairwise difference between the three LLM judges' scores for the same conversation; lower values indicate closer absolute agreement.

### 4.9 Qualitative Conversation Analysis

The qualitative analysis supported the survey and LLM-as-judge findings. Strong examples involved targeted follow-up questions after incomplete answers. Weak examples involved interruptions, missed conversational repair, inconsistent follow-up questioning, and evaluative and leading wording. Feedback was generally aligned with the conversation, although it sometimes credited knowledge partly introduced by the examiner. Representative excerpts are provided in Appendix G.

## 5 Discussion

This study evaluates whether a spoken LLM system can generate assessment dialogue that follows predefined pedagogical principles. ACA-Bot performed most consistently on dimensions that could be controlled more directly through predefined content, prompts, rubrics, and dialogue rules. It generally stayed aligned with the intended statistics concepts, remained in an examiner rather than tutor role, and produced feedback that users and LLM judges evaluated positively.

Diagnostic and adaptive questioning were less consistent. ACA-Bot sometimes asked relevant follow-up questions, but also moved on before users had fully explained their answers. Comparable answers did not always receive similar opportunities for clarification. This suggests that a structured dialogue policy can support content coverage and coherence, but can also make the interaction too rigid when the system needs to respond to a

specific user or situation.

At the same time, diagnostic and adaptive questioning can be interpreted in different ways. Evaluators may disagree about when enough evidence has been collected, whether another follow-up question is needed, and what type of question is appropriate without becoming leading. These dimensions should therefore not be defined by the LLM alone. Teachers, domain experts, and assessment experts should help define what sufficient evidence and appropriate examiner behavior mean within a specific assessment context. These definitions can then guide the prompts, dialogue rules, and evaluation criteria. Representative conversations should subsequently be evaluated by humans to determine whether the LLM performs these behaviors well enough. LLM-as-judge scores can support this evaluation, but should not be treated as sufficient evidence of pedagogical quality on their own. This is consistent with our judge comparison, where one model diverged systematically from the other two on these dimensions in particular.

Another challenge is that feedback sometimes credited students with understanding that had largely been introduced through the examiner's own follow-up question. This points to a broader problem for transcript-based feedback generation: once a follow-up question has supplied part of an answer, it becomes difficult to tell from the transcript alone whether the student already knew it or simply repeated it back. Future versions of the system could address this by having the state LLM flag when a follow-up reveals part of an answer, so that the feedback LLM does not credit that specific piece of knowledge to the student, or by requiring the student to restate or apply the idea independently before it is counted as demonstrated.

Speech interaction was closely connected to the observed weaknesses. User conversations received lower diagnostic, adaptive, and coherence scores than the controlled text-based simulations. This difference cannot be explained only by speech, because simulated students also produced more regular and constrained answers. Nevertheless, interruptions, incomplete turns, and recognition problems sometimes caused ACA-Bot to move on before a user had finished or to treat an interaction problem as an answer. Speech quality therefore affected not only usability, but also the evidence of a user's understanding.

The updated speech configuration, which gave users more time to answer, was associated with more positive interaction ratings and higher scores on several pedagogical dimensions. These patterns should be interpreted cautiously because the groups were small and participants were not randomly assigned. Users also continued to report that they did not always have enough opportunity to clarify, correct, or expand their answers. Future versions should therefore provide better conversational repair and give users more control over when they have finished answering or want to revise a response.

Participants mainly valued ACA-Bot for practising statistics, identifying knowledge gaps, and receiving feedback. The findings therefore support its potential use as a formative assessment or exam-preparation tool. They do not yet provide sufficient support for summative use. Summative use would require evidence that students receive sufficient and comparable opportunities to demonstrate their understanding, comparisons with trained human examiners, and human expert evaluation of whether ACA-Bot's diagnostic and adaptive questioning is pedagogically appropriate. Overall, the findings show that pedagogical dialogue management and speech interaction cannot be evaluated separately in spoken Conversation-Based Assessment.

## 6 Conclusion

This study presented ACA-Bot, a speech-and-LLM-based oral examiner designed around predefined pedagogical dimensions. ACA-Bot received its strongest ratings for content alignment, neutrality, and feedback quality. Assessment coherence was more moderate and was substantially stronger in controlled simulations than in spoken user interactions. However, the findings also show that spoken interaction and pedagogical quality cannot be treated separately. Interruptions and recognition errors affected whether users could clarify or expand their answers and therefore influenced the evidence collected about their understanding. ACA-Bot therefore currently appears most suitable for formative assessment and exam preparation. Future work should focus on more reliable turn-taking and conversational repair so that users have enough opportunity to clarify and expand on their answers.

## Limitations

This study provides an initial evaluation of ACA-Bot with a relatively small convenience sample in an experimental setting. It was not conducted

as part of an actual course or graded assessment, which limits transferability to authentic educational contexts. The sample was also not diverse in ways that likely matter for a speech-based system: most participants reported advanced English proficiency, and the sample did not include a range of accents or non-native speech patterns. Since recognition errors directly affected diagnostic tracking and dialogue progression, ACA-Bot's fairness and reliability for speakers with different accents or proficiency levels remains untested. We focused on introductory statistics, a course common across many academic domains, and used short oral exams to support controlled evaluation. This domain also has clearly defined concepts and a small set of expected misconceptions, which likely made diagnostic tracking easier than it would be in a less structured domain such as ethics or philosophy, where answers do not converge on a single expected understanding. Further research should therefore examine whether the findings extend to more complex or less structured subjects and settings. Also, as the LLM ratings and qualitative analysis only included completed interactions, they might overrepresent successful system use. The qualitative analysis was conducted by one researcher, and the LLM-as-judge scores may reflect model-specific assumptions about pedagogical quality: one judge, for instance, scored diagnostic questioning, adaptive questioning, and neutrality systematically lower than the other two, so the combined scores partly reflect which judges were included. Simulated students enabled controlled testing but do not fully capture real user behavior. Future studies should therefore include larger and more diverse samples, authentic educational settings, multiple human evaluators, and comparisons with trained human examiners.

## Ethical Considerations

Under the applicable procedures of the Department of Advanced Computing Sciences (DACS) at Maastricht University, this study was determined not to require formal ethics-committee review. Participants received written information about the study and its procedures before providing informed consent. Participation was voluntary, and participants could withdraw without negative consequences.

The collected data included audio recordings, conversation transcripts, timestamps, survey responses, and coded administrative information. Names and student numbers were not stored in the chatbot logs, and research data were associated with participant codes or technical identifiers. Participants were informed that voice recordings may be more identifiable than written transcripts. The data are stored securely by Maastricht University for ten years.

Human audio is not publicly released because of privacy and identifiability concerns. The system was evaluated in a mock-examination setting, and the present evidence does not support its use for high-stakes or graded assessment.

## Data Availability

The complete ACA-Bot system is made available in a public GitHub repository [1], and includes the source code, configuration files, assessment materials, simulated conversations and evaluation results. Human audio is not released because of privacy concerns. Anonymous written data, including human conversation transcripts and survey responses, may be made available to researchers upon reasonable request, subject to participant consent and the data-protection requirements of Maastricht University.

## References


Lorin W. Anderson and David R. Krathwohl, editors. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York, NY, USA.

AssemblyAI. n.d. AssemblyAI Documentation. Accessed: May 28, 2026.

Tiffany Bayley, Kyle D. S. Maclean, and Tessa Weidner. 2024. Back to the Future: Implementing Large-Scale Oral Exams. *Management Teaching Review*.

Russell Beale. 2025. Dialogic Pedagogy for Large Language Models: Aligning Conversational AI with Proven Theories of Learning. *arXiv preprint*.

Jan Bergerhoff, Johannes Bendler, Stefan Stefanov, Enrico Cavinato, Leonard Esser, Tommy Tran, and Aki Härmä. 2025. Automatic conversational assessment using large language model technology. In *Proceedings of the 2024 16th International Conference on Education Technology and Computers*, pages 39–45. Association for Computing Machinery.

Galo Castillo-López, Gaël de Chalendar, and Nasredine Semmar. 2025. A survey of recent advances on turn-taking modeling in spoken dialogue systems. In *Proc. 15th Int. Workshop on Spoken Dialogue Systems Technology*, pages 254–271.


[1] https://github.com/Juulba/SPEECH-ACA-BOT

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. 2026. VoiceBench: Benchmarking LLM-based voice assistants. *Transactions of the Association for Computational Linguistics*.

Ronald Cole, Cindy Buchenroth-Martin, Timothy Weston, Liam Devine, Jeannine Myatt, Brandon Helding, Sameer Pradhan, Margaret McKeown, Samantha Messier, Jennifer Borum, and Wayne Ward. 2018. One-on-One and Small Group Conversations with an Intelligent Virtual Science Tutor. *Computer Speech & Language*, 50:157–174.

Confident AI. 2026. DeepEval: The LLM Evaluation Framework. DeepEval Documentation, 2026. Accessed: May 31, 2026.

Sidney K. D'Mello, Nia Dowell, and Arthur C. Graesser. 2011. Does It Really Matter Whether Students' Contributions Are Spoken Versus Typed in an Intelligent Tutoring System with Natural Language? *Journal of Experimental Psychology: Applied*, 17(1):1–17.

ElevenLabs. n.d. ElevenLabs Documentation. Accessed: May 28, 2026.

Kate Forbes-Riley and Diane Litman. 2011. Designing and Evaluating a Wizarded Uncertainty-Adaptive Spoken Dialogue Tutoring System. *Computer Speech & Language*, 25(1):105–126.

Carol M. Forsyth, Diego Zapata-Rivera, Edith Aurora Graf, and Yang Jiang. 2024. Complex Conversations: LLM vs. Knowledge Engineering Conversation-Based Assessment. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 868–871.

Ali Geris and Günter Alce. 2026. Real-Time Voice-Based LLM Integration for XR Tutoring: A Prototype Implementation. In *Proc. IEEE Int. Conf. Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*, pages 285–289, Osaka, Japan.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of Educational Research*, 77(1):81–112.

Xinying Hou, Carol M. Forsyth, Jessica Andrews-Todd, James Rice, Zhiqiang Cai, Yang Jiang, Diego Zapata-Rivera, and Arthur C. Graesser. 2025. An LLM-Enhanced Multi-Agent Architecture for Conversation-Based Assessment. In *Proc. 26th Int. Conf. Artificial Intelligence in Education (AIED)*, pages 119–134, Palermo, Italy.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, Xiaoda Yang, Zehan Wang, Qian Yang, Jian Li, Yidi Jiang, Jingzhen He, Yunfei Chu, Jin Xu, and Zhou Zhao. 2024. WavChat: A Survey of Spoken Dialogue Models. *arXiv preprint*.

Gordon Joughin. 1998. Dimensions of oral assessment. *Assessment & Evaluation in Higher Education*, 23(4):367–378.

Michael T. Kane. 2013. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1):1–73.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Diane J. Litman, Carolyn P. Rosé, Kate Forbes-Riley, Kurt VanLehn, Dumisizwe Bhembe, and Scott Silliman. 2006. Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence in Education*, 16(2):145–170.

Zhanxun Liu, Yifan Duan, Mengmeng Wang, Pengchao Feng, Haotian Zhang, Xiaoyu Xing, Yijia Shan, Haina Zhu, Yuhang Dai, Chaochao Lu, Xipeng Qiu, Lei Xie, Lan Wang, Nan Yan, Zilong Zheng, Ziyang Ma, Kai Yu, and Xie Chen. 2025. X-Talk: On the Underestimated Potential of Modular Speech-to-Speech Dialogue System. *arXiv preprint*.

LiveKit. n.d. agent-starter-react. GitHub repository. Accessed: May 28, 2026.

Muhammed Ashraf Memon, Gordon Rowland Joughin, and Breda Memon. 2010. Oral assessment and postgraduate medical examinations: Establishing conditions for validity, reliability and fairness. *Advances in Health Sciences Education*, 15:277–289.

OpenAI. 2026. OpenAI API model documentation. OpenAI Platform Documentation. Accessed: June 4, 2026.

Jacob Pearce and Neville Chiavaroli. 2020. Prompting candidates in oral assessment contexts: A taxonomy and guiding principles. *Journal of Medical Education and Curricular Development*, 7:2382120520948881.

Brian K. Sato, Cynthia F. C. Hill, and Stanley M. Lo. 2019. Testing the test: Are exams measuring understanding? *Biochemistry and Molecular Biology Education*, 47(3):296–302.

Valerie J. Shute. 2008. Focus on formative feedback. *Review of Educational Research*, 78(1):153–189.

Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard G. Baraniuk. 2024. Pedagogical Alignment of Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13641–13650.

Allison S. Theobold. 2021. Oral Exams: A More Meaningful Assessment of Students' Understanding. *Journal of Statistics and Data Science Education*, 29(2):156–159.

Diego Zapata-Rivera, Carol M. Forsyth, Edith Aurora Graf, and Yang Jiang. 2024. Designing and Evaluating Evidence-Centered Design Based Conversations for Assessment with LLMs. In *Joint Proceedings of the Human-Centric eXplainable AI in Education and the Leveraging Large Language Models for Next Generation Educational Technologies Workshops*, volume 3840 of *CEUR Workshop Proceedings*, Atlanta, Georgia, USA.

Diego Zapata-Rivera, Jesse R. Sparks, Carol M. Forsyth, and Blair A. Lehman. 2023. Conversation-Based Assessment: Current Findings and Future Work. In Robert J. Tierney, Fazal Rizvi, and Kadriye Ercikan, editors, *International Encyclopedia of Education*, 4 edition, pages 504–518. Elsevier, Amsterdam, Netherlands.

# A Detailed Methodological Materials

This appendix provides the methodological details that were shortened in the conference paper. It focuses on the parts that are needed to understand how ACA-Bot was designed, how the dialogue progressed, and how the study and evaluations were carried out. The complete executable implementation, including the exact prompts and configuration files, is present in the code repository. The prompt descriptions below present the instructions and information that shaped the experiments in a readable form.

## A.1 Pedagogical Dimensions and System-Design Requirements

The pedagogical dimensions introduced in the main paper were translated into concrete system-design requirements. Table 2 summarizes how each dimension informed the dialogue structure, prompting constraints, and feedback generation in ACA-Bot.

## A.2 Assessment Tasks and Subtopics

The assessment contained three introductory-statistics tasks. Each task was divided into subtopics because the task labels alone were too broad to determine what should be questioned and when sufficient evidence had been collected. The same subtopics were used for every participant, while follow-up questions could vary based on the user's response. Table 3 contains the statistical tasks and subtopics used in the oral exam.

## A.3 LLM Components and Prompt Instructions

ACA-Bot used separate LLM components for examiner responses, diagnostic-state tracking, and final feedback. Separating these roles made it possible to constrain dialogue progression outside the user-facing response generator.

### A.3.1 Examiner LLM

The examiner LLM generated all user-facing responses during the exam. Its prompt assigned it the role of an oral examiner rather than a tutor and included the following instructions:

- Assess the user's own statistical understanding without teaching, correcting, hinting, or revealing the expected answer.
- Avoid leading questions and avoid praise, reassurance, discouragement, or explicit judgments of whether an answer is correct.
- Use a brief neutral acknowledgment only when this helps the conversational flow.
- Keep responses short and natural for a spoken interaction and ask one question at a time.
- Avoid bullet points, numbering, checklist-style questioning, unnecessary repetition, and unnecessary restatement of the user's answer.
- When a follow-up is requested, build it on the user's preceding answer and target the part that remains incomplete or unclear.
- Follow the action selected by the dialogue controller: introduce a task and opening subtopic, ask a targeted follow-up, move to a new subtopic, or provide a transition.

At each turn, the examiner received the current task, current subtopic, the preceding conversation for that task, the action selected by the controller, and task-relevant reference material. The examiner did not independently determine whether a subtopic was complete.

### A.3.2 State LLM

The state LLM operated in the background and classified the amount of diagnostic evidence collected for every subtopic in the current task. It received the current task and target subtopic, the full list of subtopics for that task, and the conversation history for the task. It was instructed to place every subtopic in exactly one of the four states in Table 4 and to provide a short reason for each classification.

All subtopics in the current task were classified at the same time. This allowed the state LLM to recognize when a user addressed more than one subtopic in a single answer and was intended to reduce unnecessary repetition later in the conversation.

### A.3.3 Feedback LLM

The feedback LLM received the complete examiner-user transcript after the assessment. It was instructed to first make clear that the exam had ended and then address the user directly. The feedback instructions required the model to:

- base all statements on evidence in the transcript;
- identify exact concepts the user demonstrated correctly, but only when these strengths were clearly supported;
- identify important gaps, weaknesses, or misconceptions only when they were visible in the user's own answers;

Table 2: Pedagogical dimensions and how they were translated into ACA-Bot's design.

| Pedagogical dimension | Translation into ACA-Bot's design |
|---|---|
| Diagnostic questioning | The dialogue policy tracked whether a user's answer provided enough evidence to judge understanding or identify a misconception. An incomplete or unclear answer could therefore trigger a follow-up question before the system moved on. |
| Adaptive questioning | Follow-up questions were generated from the user's preceding response and were instructed to target the specific part that remained unclear or incomplete. |
| Neutrality | The examiner was prohibited from teaching, correcting, hinting, asking leading questions, revealing expected-answer content, or giving praise, reassurance, or discouragement during the exam. |
| Prompting consistency | All users received the same tasks and subtopics. The same diagnostic states, follow-up limits, and progression rules were used to determine whether the system asked another question or continued. |
| Content alignment | Questions were restricted to predefined statistics tasks and subtopics. Task-relevant reference material was supplied to the examiner to support alignment with the intended content. |
| Assessment coherence | The interaction followed a controlled sequence consisting of an introduction, task and subtopic progression, possible follow-up questions, transitions, and feedback after the exam. |
| Feedback quality | The feedback component received the complete transcript and was instructed to identify only transcript-supported strengths, gaps, misconceptions, and concrete suggestions for improvement. |

Table 3: Statistical tasks and subtopics used to structure the oral exam.

| Task | Subtopics |
|---|---|
| Hypothesis testing | • Explain what the null hypothesis ($H_0$) and alternative hypothesis ($H_1$) represent. |
| | • Interpret what a p-value means in the context of hypothesis testing. |
| | • Explain how the significance threshold is used to draw a conclusion about $H_0$. |
| | • Explain how certain we can be about the conclusion of a hypothesis test. |
| Mean and median | • Explain how the mean is computed. |
| | • Explain how the median is found, including the difference between odd and even numbers of observations. |
| | • Reason about how outliers affect the mean. |
| | • Reason about how outliers affect the median. |
| Variance | • Explain what variance is without using a formula. |
| | • Interpret what high versus low variance means for a dataset. |
| | • Explain the relationship between variance and standard deviation. |

Table 4: Diagnostic states used by the state LLM.

| State | Meaning |
|---|---|
| Diagnostically resolved | The response provided enough direct evidence to identify the user's understanding or a specific misconception. A resolved answer did not need to be correct. |
| Partially reasoned | The user meaningfully addressed the subtopic, but the answer remained incomplete or unclear. A targeted follow-up could still provide useful diagnostic evidence. |
| Non-diagnostic | The subtopic was explicitly targeted, but the response provided no usable evidence, for example because the user said that they did not know, gave no meaningful attempt, or responded irrelevantly. |
| Unseen | The subtopic had not yet been explicitly attempted or meaningfully addressed. |

- avoid praising an entire topic when only one part was demonstrated;
- avoid criticizing content that was not asked or attempted;
- provide a concrete recommendation for each important weakness or misconception;
- avoid vague advice such as "study more" unless it specifies what should be reviewed or practised;
- give no grade and keep the response to two or three short paragraphs.

### A.4 Dialogue Control and Progression

The state classification was not used as a direct instruction to the examiner. A rule-based controller first validated and updated the state and then selected the next conversational action. The main

progression rules were as follows:

1. A diagnostically resolved subtopic was marked as completed and was not questioned again.
2. A partially reasoned subtopic triggered a targeted follow-up question, provided that the follow-up limit had not been reached.
3. When the follow-up limit was reached without sufficient evidence, the subtopic was closed and the system continued.
4. A non-diagnostic subtopic was closed because it had been attempted without producing usable evidence.
5. An unseen target subtopic received an opening question.
6. After a subtopic was completed or closed, the controller selected the next unresolved subtopic in the same task.
7. When all subtopics in a task were completed or closed, the system moved to the next task.
8. When no tasks remained, the exam ended and the feedback LLM generated the final response.

The controller also checked that each subtopic appeared in one valid state only. Progress was preserved across turns: a resolved subtopic could not later be downgraded, and a partially reasoned subtopic could be upgraded to resolved but not downgraded to non-diagnostic. Spoken-interaction safeguards were added as well. The first silence on a subtopic triggered a request to repeat; a second silence for the same subtopic was treated as non-diagnostic. A maximum-turn limit for each task prevented the system from becoming stuck in repeated follow-ups or unresolved exchanges.

### A.5 Reference Material and RAG

ACA-Bot was grounded in an introductory-statistics reference document covering hypothesis testing, mean and median, and variance. The document was composed with LLM assistance and manually checked for accuracy. Its text was extracted, divided into smaller chunks, embedded, and stored in a vector database. For each task, the most relevant chunks were retrieved and supplied to the examiner together with the task and subtopic information. This was intended to reduce reliance on the model's unrestricted general knowledge and to keep the questions aligned with the assessment content.

### A.6 Logging and Recorded Variables

Each exam received a unique conversation identifier. The logs contained user and examiner messages, final feedback, and relevant system messages. For each message, the system recorded the speaker role, timestamp, response time, current task, current subtopic, and diagnostic-state information. For user turns, the diagnostic log also contained the state classification, next target subtopic, and next action selected by the controller. User audio was stored separately from the text logs but was not analyzed in the present study.

### A.7 Speech Configurations

The speech configuration was changed during data collection after participants experienced interruptions. The groups were not randomized and were too small for strong causal comparisons; the distinction was therefore used descriptively.

### A.8 User Interface and Study Procedure

Participants were recruited through the researcher's personal and academic network. Before taking part, they received study information and a consent form explaining the procedure, data collection and storage, and their right to withdraw. They were asked to complete the study in a quiet environment, use a functioning microphone, answer as they would in a real oral exam, and complete the exam in one sitting. Participants were explicitly encouraged to provide honest feedback and were informed that both positive and negative experiences could guide further development.

The study consisted of a short spoken mock exam followed by the Qualtrics questionnaire. Through the web interface (see Figure 2), participants heard ACA-Bot's questions and could read the corresponding transcript. They answered verbally, received possible follow-up questions, and were given feedback after the final task.

### A.9 Synthetic Student Simulation

Forty text-based conversations were generated with synthetic students. Speech recognition, speech synthesis, and turn detection were excluded so that the dialogue policy could be examined without speech-technology errors. Before each simulated exam, one fixed knowledge state was sampled independently for every assessment subtopic. The implementation used the four states and probabilities shown in Table 5. Because states were sampled

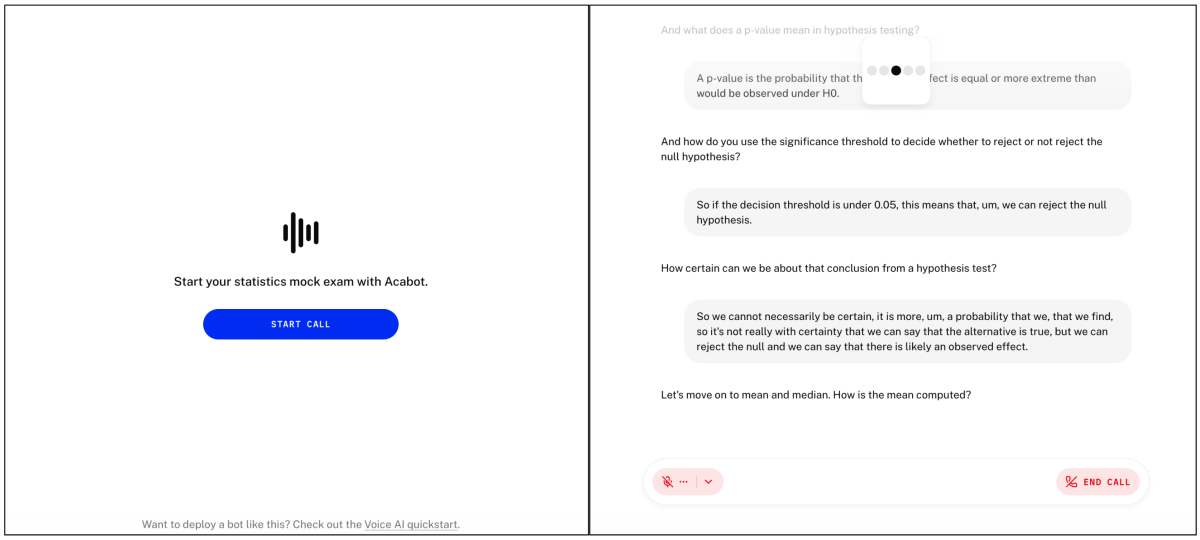



Figure 2: Adapted LiveKit interface used during the user study.

Table 5: Knowledge states used for the synthetic students.

| State | Probability | Intended behavior |
|---|---|---|
| Correct | 0.35 | The student possessed the relevant knowledge but revealed it gradually across the opening question and possible follow-ups. |
| Partial | 0.40 | The student had limited or incomplete knowledge and did not become fully precise after follow-up questions. |
| Wrong | 0.15 | The student followed a predefined misconception consistently rather than switching to a correct answer. |
| No knowledge | 0.10 | The student stated uncertainty or gave a very vague answer and did not provide knowledge from related subtopics. |

Table 6: Illustrative synthetic knowledge profiles for the p-value subtopic.

| State | Knowledge and response constraint |
|---|---|
| Correct | The student first knows that a p-value concerns how surprising the observed result is, then adds that this is evaluated under $H_0$, and only after further probing gives the more precise probability interpretation. The student also knows that the p-value is not the probability that $H_0$ is true. |
| Partial | The student knows that a p-value concerns how surprising a result is and vaguely connects this to $H_0$, but cannot provide the full probability definition. |
| Wrong | The student believes that the p-value is the probability that $H_0$ is true and, after follow-up, interprets $p = .03$ as a 3% probability that $H_0$ is correct. |
| No knowledge | The student says that they are not sure and does not answer adjacent hypothesis-testing subtopics. |

independently, the simulation did not model dependencies between related concepts.

The synthetic-student prompt instructed the model to answer naturally as a student in an oral exam, remain within the current subtopic, and reveal only one reasoning step per answer. On the first question, it used only the first available knowledge step; each follow-up could add at most one next step. A partial or incorrect student remained within the limits of the sampled state, and a student with no knowledge did not suddenly produce a correct answer. This stepwise structure created opportunities for ACA-Bot to ask diagnostic follow-ups instead of receiving complete answers immediately.

Table 6 gives one example of how the knowledge profiles were specified. Equivalent profiles were defined for all eleven subtopics in the implementation.

## B Questionnaire

The following questionnaire was filled out through Qualtrics after participants completed the ACA-Bot oral exam. The questionnaire consisted of demographic questions, Likert-scale items, and open-ended questions. Unless otherwise stated, Likert-scale items were answered on a five-point scale ranging from Strongly disagree to Strongly agree.

### B.1 Demographics

Participants were first asked to provide background information.

1. **Which best describes you?**
   - Undergraduate student
   - Master's student
   - PhD student
   - Other type of student
   - Recent graduate
   - Not a student and not a recent graduate
   - Prefer not to say
2. **What is your main field of study or academic background?**
   - Computer Science, Data Science, AI
   - Health or Medical Sciences
   - Economics or Business
   - Engineering
   - Education
   - Psychology
   - Political or Social Sciences
   - Physics, Chemistry, Mathematics, Biology, or related natural sciences
   - Arts, History, Literature, Languages, or other humanities
   - Earth, Environmental, or Geographical Sciences
   - Other
   - Prefer not to say
3. **What is the highest level of statistics education you have completed?**
   - No formal statistics education
   - Introductory statistics course
   - Intermediate statistics course
   - Advanced statistics course
   - Statistics or data science is/was a major part of my degree/work
   - Other
   - Prefer not to say
4. **When did you last study statistics formally?**
   - Currently studying statistics
   - Less than 6 months ago
   - 6–12 months ago
   - 1–2 years ago
   - More than 2 years ago
   - I have never studied statistics formally
   - Prefer not to say
5. **How confident do you feel about your basic statistics knowledge?**
   - Very unconfident
   - Unconfident
   - Somewhat confident
   - Moderately confident
   - Very confident
   - Prefer not to say
6. **What is your overall level of English?**
   - Basic: "I can understand and communicate simple ideas in English."
   - Intermediate: "I can understand and discuss academic topics in English, though sometimes with effort."
   - Advanced: "I can explain academic ideas clearly and confidently in English."
   - Prefer not to say
7. **Have you previously taken an oral examination, viva, or spoken assessment?**
   - Yes
   - No
   - Not sure
   - Prefer not to say
8. **What is your general attitude toward oral examinations or spoken assessments?**
   - Very negative
   - Negative
   - Neutral
   - Positive
   - Very positive
   - Prefer not to say
9. **What is your general attitude toward AI applications, particularly in education?**
   - Very negative

- Negative
- Neutral
- Positive
- Very positive
- Prefer not to say

### B.2 Pedagogical Quality

Participants were asked to rate their agreement with the following statements about their experience with ACA-Bot.

Scale: Strongly disagree, Disagree, Neutral, Agree, Strongly agree.

1. The chatbot assessed my understanding of statistics concepts rather than only my ability to recall definitions or formulas.
2. The chatbot asked me to explain the reasoning behind my answers.
3. The chatbot avoided repeating similar follow-up questions unnecessarily.
4. The chatbot's follow-up questions were relevant to what I had previously said.
5. The chatbot's follow-up questions helped clarify what I understood or misunderstood.
6. The chatbot asked follow-up questions without giving away the correct answer.
7. The chatbot gave me a fair opportunity to clarify, correct, or expand on my answers.
8. The chatbot's questions were clear enough for me to understand what was being asked.
9. The interaction felt like a structured oral examination.
10. The difficulty of the chatbot's questions was appropriate for my level of statistics knowledge.
11. The chatbot covered the statistics topic broadly enough to judge my understanding.
12. The chatbot focused on important statistical ideas rather than minor details.
13. During the assessment, the chatbot mainly asked questions to assess my understanding rather than teaching me the material.
14. My performance in the chatbot exam reflected my actual understanding of the statistics topics.

**Open-ended question:**

15. What could ACA-Bot do differently to better assess your understanding of statistics?

### B.3 Feedback Quality

Participants were asked to rate their agreement with the following statements about the feedback provided after the exam.

*Scale: Strongly disagree, Disagree, Neutral, Agree, Strongly agree.*

1. The feedback was consistent with what happened during the exam.
2. The feedback accurately identified what I understood and misunderstood during the exam.
3. The feedback was specific enough for me to understand which parts of my answers were strong or weak.
4. The feedback gave me clear guidance on how I could improve my understanding or answers.
5. The feedback gave a balanced view of my performance, including both strengths and areas for improvement.
6. The feedback was clear and understandable.

**Open-ended question:**

7. In what ways could the feedback provided after the exam be improved to make it more useful, clear, or helpful for your learning?

### B.4 Speech Modality

Participants were asked to rate their agreement with the following statements about the speech-based interaction with ACA-Bot.

Scale: Strongly disagree, Disagree, Neutral, Agree, Strongly agree.

1. Speaking my answers helped me explain my reasoning more naturally than typing would have.
2. The speech-based format made the interaction feel more like an oral examination.
3. The speech-based format made it easier to show my understanding of statistics.
4. The speech-based format made me more nervous than a text-based interaction would have.
5. Speech recognition or audio issues made it harder to express my answers.
6. I would prefer typing my answers over speaking them.
7. ACA-Bot interrupted me while I was still speaking.

8. Speaking to ACA-Bot felt like a natural conversation.
9. ACA-Bot responded quickly enough after I finished speaking.

### B.5 Perceived Suitability for Assessment Use

Participants were asked to rate their agreement with the following statements about the possible use of ACA-Bot for oral assessment and statistics learning.
Scale: Strongly disagree, Disagree, Neutral, Agree, Strongly agree.

1. I would feel comfortable taking an oral exam conducted by ACA-Bot.
2. ACA-Bot could be useful for practicing statistics before an exam.
3. I would recommend ACA-Bot to a friend or fellow student as a tool to support studying statistics.
4. ACA-Bot could help identify strengths and weaknesses in a student's statistics understanding to support further learning.
5. ACA-Bot could be a useful additional tool to support statistics assessment.
6. ACA-Bot could replace parts of a traditional oral exam.
7. I would trust ACA-Bot to give a fair indication of my statistics understanding.
8. I would accept that my conversation with ACA-Bot could be used as evidence in a graded statistics assessment.
9. I would prefer an ACA-Bot-style oral exam over a written exam for a statistics course.

**Open-ended question:**

10. What do you see as the main opportunities and limitations of using ACA-Bot for statistics learning or assessment?

### B.6 Survey Analysis and Item-Level Summary Statistics

Closed-ended responses were summarized using frequencies, percentages, medians, and standard deviations. The median was used as the main measure of central tendency because the Likert responses were ordinal. The open-ended responses were coded for recurring themes. The tables below preserve the item-level median and standard-deviation results that were omitted from the shorter paper. Negatively phrased speech items were reverse-coded for these summary statistics so that higher values consistently indicate a more positive evaluation.

## C Participant Characteristics

Table 7: Participant characteristics by speech configuration group.

| Characteristic | Init. | Upd. | All |
|---|---|---|---|
| **Participant role** | | | |
| Undergraduate student | 4 (30.8) | 3 (14.3) | 7 (20.6) |
| Master's student | 4 (30.8) | 13 (61.9) | 17 (50.0) |
| PhD student | 2 (15.4) | 0 (0.0) | 2 (5.9) |
| Recent graduate | 1 (7.7) | 4 (19.0) | 5 (14.7) |
| Not a student and not a recent graduate | 2 (15.4) | 1 (4.8) | 3 (8.8) |
| **Academic background** | | | |
| Computer science/data science/AI | 3 (23.1) | 11 (52.4) | 14 (41.2) |
| Economics or business | 2 (15.4) | 4 (19.0) | 6 (17.6) |
| Natural sciences/mathematics | 3 (23.1) | 0 (0.0) | 3 (8.8) |
| Health or medical sciences | 0 (0.0) | 3 (14.3) | 3 (8.8) |
| Political or social sciences | 1 (7.7) | 1 (4.8) | 2 (5.9) |
| Earth/environmental/geographical sciences | 2 (15.4) | 0 (0.0) | 2 (5.9) |
| Engineering | 1 (7.7) | 0 (0.0) | 1 (2.9) |
| Humanities | 1 (7.7) | 0 (0.0) | 1 (2.9) |
| Education | 0 (0.0) | 1 (4.8) | 1 (2.9) |
| Other | 0 (0.0) | 1 (4.8) | 1 (2.9) |
| **Highest statistics education** | | | |
| No formal statistics education | 0 (0.0) | 2 (9.5) | 2 (5.9) |
| Introductory statistics course | 4 (30.8) | 4 (19.0) | 8 (23.5) |
| Intermediate statistics course | 5 (38.5) | 9 (42.9) | 14 (41.2) |
| Advanced statistics course | 1 (7.7) | 1 (4.8) | 2 (5.9) |
| Statistics/data science major part of degree/work | 3 (23.1) | 5 (23.8) | 8 (23.5) |
| **Last formal statistics study** | | | |
| Currently studying statistics | 2 (15.4) | 1 (4.8) | 3 (8.8) |
| Less than 6 months ago | 2 (15.4) | 1 (4.8) | 3 (8.8) |
| 6–12 months ago | 2 (15.4) | 5 (23.8) | 7 (20.6) |
| 1–2 years ago | 1 (7.7) | 4 (19.0) | 5 (14.7) |
| More than 2 years ago | 6 (46.2) | 9 (42.9) | 15 (44.1) |
| I have never studied statistics formally | 0 (0.0) | 1 (4.8) | 1 (2.9) |
| **Confidence in basic statistics** | | | |
| Very unconfident | 1 (7.7) | 2 (9.5) | 3 (8.8) |
| Unconfident | 4 (30.8) | 6 (28.6) | 10 (29.4) |
| Somewhat confident | 4 (30.8) | 3 (14.3) | 7 (20.6) |
| Moderately confident | 3 (23.1) | 9 (42.9) | 12 (35.3) |
| Very confident | 1 (7.7) | 1 (4.8) | 2 (5.9) |
| **English level** | | | |
| Advanced | 11 (84.6) | 13 (61.9) | 24 (70.6) |
| Intermediate | 2 (15.4) | 8 (38.1) | 10 (29.4) |
| **Prior oral/spoken assessment** | | | |
| Yes | 8 (61.5) | 13 (61.9) | 21 (61.8) |
| No | 4 (30.8) | 8 (38.1) | 12 (35.3) |
| Not sure | 1 (7.7) | 0 (0.0) | 1 (2.9) |
| **Attitude toward oral assessments** | | | |
| Very negative | 0 (0.0) | 1 (4.8) | 1 (2.9) |
| Negative | 1 (7.7) | 4 (19.0) | 5 (14.7) |
| Neutral | 2 (15.4) | 6 (28.6) | 8 (23.5) |
| Positive | 7 (53.8) | 7 (33.3) | 14 (41.2) |
| Very positive | 3 (23.1) | 3 (14.3) | 6 (17.6) |
| **Attitude toward AI in education** | | | |
| Negative | 3 (23.1) | 0 (0.0) | 3 (8.8) |
| Neutral | 4 (30.8) | 6 (28.6) | 10 (29.4) |
| Positive | 5 (38.5) | 9 (42.9) | 14 (41.2) |

*Continued on next page*

*Table 7 continued from the previous page.*

| Characteristic | Init. | Upd. | All |
|---|---|---|---|
| Very positive | 1 (7.7) | 6 (28.6) | 7 (20.6) |

*Note.* Values are $n$ (%). Percentages are calculated within each column. The groups represent the two speech configurations: the first 13 complete responses and the subsequent 21 responses.

# D Dataset and Conversation Characteristics

## D.1 Conversation Descriptives by Speech Configuration

Table 8: Descriptive metrics for user conversations by speech configuration group.

| | Init. | | Upd. | | All | |
|---|---|---|---|---|---|---|
| **Metric** | **M** | **SD** | **M** | **SD** | **M** | **SD** |
| Number of turns | 35.23 | 6.14 | 34.76 | 7.47 | 34.94 | 6.90 |
| User response length (words) | 18.87 | 9.79 | 19.58 | 10.97 | 19.31 | 10.39 |
| Examiner response length (words) | 12.82 | 0.87 | 12.83 | 0.80 | 12.83 | 0.82 |
| Feedback length (words) | 187.92 | 23.40 | 176.48 | 20.95 | 180.85 | 22.30 |
| Conversation duration (minutes) | 5.62 | 1.19 | 7.48 | 2.34 | 6.77 | 2.16 |

## D.2 User and Simulated Conversations

Table 9: Descriptive metrics for user and simulated conversation data.

| | User | | Simulated | |
|---|---|---|---|---|
| **Metric** | **M** | **SD** | **M** | **SD** |
| Number of turns | 34.94 | 6.90 | 41.20 | 5.36 |
| User response length (words) | 19.31 | 10.39 | 13.30 | 2.02 |
| Examiner response length (words) | 12.83 | 0.82 | 12.00 | 0.53 |
| Feedback length (words) | 180.85 | 22.30 | 198.35 | 22.04 |
| Conversation duration (minutes) | 6.77 | 2.16 | – | – |

# E Survey Results

## E.1 Likert-Scale Survey Items

Figure 3 shows the distribution of responses to the Likert-scale items across the four evaluation dimensions: pedagogical quality, feedback quality, speech modality, and perceived suitability for learning and assessment. Each stacked bar reports the number of participants selecting each response option. Tables 11–14 provide the corresponding medians and standard deviations for the full sample and by speech-configuration group.

## E.2 Open-Ended Survey Themes

Thirty-one of the 34 participants answered at least one of the three open-ended survey questions. Themes were not mutually exclusive, and a participant was counted at most once within each theme.

Table 10: Themes identified in the open-ended survey responses.

| Theme | $n$ |
|---|---|
| Rushed interaction, interruptions, or turn-taking problems | 21 |
| Useful for practice, learning, exam preparation, or identifying gaps | 18 |
| Caution about graded assessment, fairness, or accuracy | 16 |
| Need for deeper, more practical, or more adaptive questioning | 13 |
| Feedback was clear, useful, accurate, or good | 13 |
| Need for more user control during the interaction | 12 |
| Feedback could be more detailed, example-based, or actionable | 10 |
| Technical or speech-recognition issues | 6 |
| Possible assessment benefits | 6 |
| Stress, discomfort, nervousness, or unnatural interaction | 6 |
| Need for clearer structure, context, or exam flow | 5 |

# F LLM-as-Judge Evaluation Details

## F.1 Evaluation Setup and Shared Judge Instructions

The user and simulated conversations were evaluated independently by GPT-4.1, Mistral Medium 3.5, and Claude Sonnet 4.5. The final score for each dimension was the mean of the three judge scores. Diagnostic questioning, adaptive questioning, neutrality, content alignment, and assessment coherence were evaluated from the full dialogue. Feedback quality was evaluated by comparing the final feedback message with the preceding conversation. Prompting consistency was not scored because it requires comparisons across matched conversations rather than evaluation of one conversation at a time.

All judge prompts included the following shared rules:

1. Evaluate only the examiner's behavior.
2. Do not lower the score because the user gives an incorrect, incomplete, vague, or uncertain answer.
3. Use user answers only as evidence for whether the examiner responded appropriately.
4. Score what the examiner did, rather than how well the user performed.
5. Do not give a high score only because the examiner sounded fluent.
6. Use the supplied score anchors and do not invent a different scoring standard.
7. Do not invent a weakness when no evidence for it is present.

The reason accompanying each score was instructed to explain the score interpretation, cite concrete examiner behavior, identify weaknesses that lowered the score, and justify why the selected anchor fit better than a higher or lower score. Every metric used a threshold of 0.70 in the evaluation framework.

For descriptive interpretation in the thesis and paper, scores from 0.85 to 1.00 were described as strong, 0.70 to 0.84 as sufficient to good, 0.50 to 0.69 as mixed or partially sufficient, 0.25 to 0.49 as weak, and scores below 0.25 as very weak.

## F.2 Metric-Specific Criteria

### F.2.1 Diagnostic Questioning

This metric evaluated whether the examiner collected enough relevant evidence to judge the user's understanding of each assessment target. Useful probing included clarifying a vague, short, incomplete, or correct-sounding response; asking for explanation when terminology alone was insufficient; and targeting a missing part of the subtopic. Further probing was not required when the answer was already complete, the user clearly lacked knowledge, the response provided no useful basis for another question, or an additional question would mainly teach or lead the user. High scores therefore reflected sufficient and targeted evidence collection without over-probing. Low scores reflected moving on while the user's understanding was still unclear.

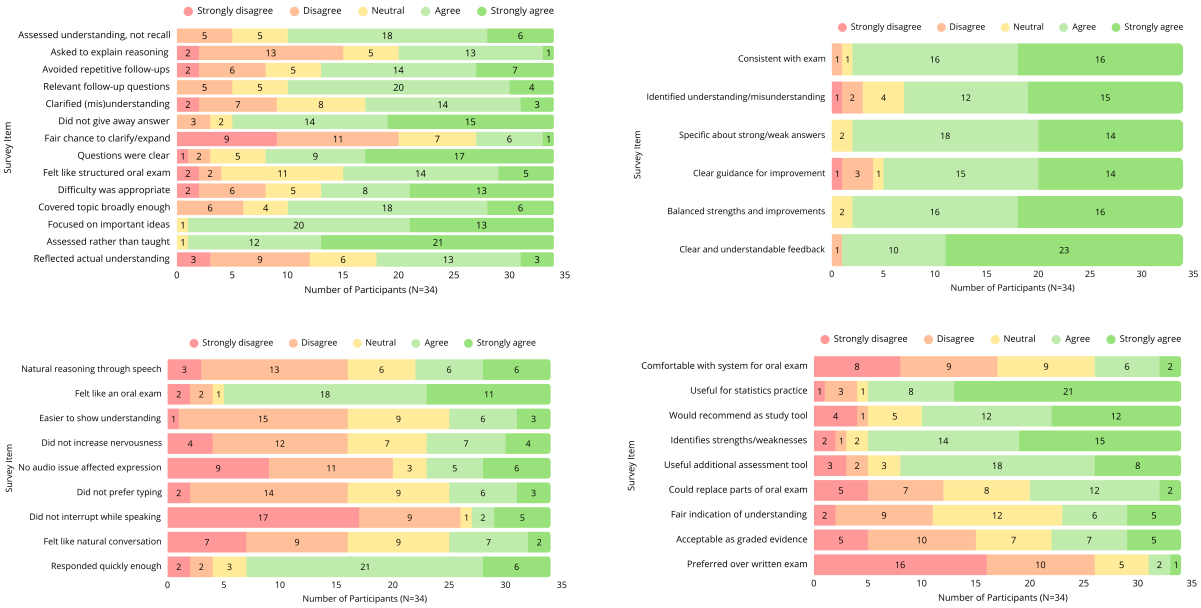



Figure 3: Distribution of Likert-scale responses across the evaluation dimensions. Each stacked bar shows the number of respondents who selected each response option. The total number of respondents per item is 34.

Table 11: Pedagogical-quality item results by speech configuration group.

| | Init. | | Upd. | | All | |
|---|---|---|---|---|---|---|
| **Item** | **Mdn** | **SD** | **Mdn** | **SD** | **Mdn** | **SD** |
| Assessed understanding, not recall | 4 | 1.05 | 4 | 0.85 | 4 | 0.93 |
| Asked to explain reasoning | 2 | 1.13 | 3 | 0.94 | 3 | 1.07 |
| Avoided repetitive follow-ups | 4 | 0.83 | 4 | 1.36 | 4 | 1.19 |
| Relevant follow-up questions | 4 | 0.97 | 4 | 0.83 | 4 | 0.88 |
| Clarified (mis)understanding | 3 | 1.00 | 4 | 1.12 | 3.5 | 1.08 |
| Did not give away answer | 5 | 0.52 | 4 | 1.05 | 4 | 0.91 |
| Fair chance to clarify or expand | 2 | 1.41 | 2 | 0.98 | 2 | 1.16 |
| Questions were clear | 4 | 1.21 | 5 | 0.97 | 4.5 | 1.08 |
| Felt like a structured oral exam | 3 | 1.07 | 4 | 0.74 | 4 | 1.02 |
| Difficulty was appropriate | 4 | 1.50 | 4 | 1.22 | 4 | 1.31 |
| Covered topic broadly enough | 3 | 1.25 | 4 | 0.67 | 4 | 0.97 |
| Focused on important ideas | 4 | 0.63 | 4 | 0.50 | 4 | 0.54 |
| Assessed rather than taught | 5 | 0.66 | 5 | 0.48 | 5 | 0.56 |
| Reflected actual understanding | 2 | 1.34 | 3 | 1.06 | 3 | 1.17 |

Table 12: Feedback-quality item results by speech configuration group.

| | Init. | | Upd. | | All | |
|---|---|---|---|---|---|---|
| **Item** | **Mdn** | **SD** | **Mdn** | **SD** | **Mdn** | **SD** |
| Consistent with exam | 4 | 0.80 | 5 | 0.60 | 4 | 0.70 |
| Identified (mis)understanding | 4 | 1.14 | 5 | 0.96 | 4 | 1.04 |
| Specific about strong and weak answers | 4 | 0.51 | 4 | 0.66 | 4 | 0.60 |
| Clear guidance for improvement | 4 | 0.76 | 5 | 1.20 | 4 | 1.04 |
| Balanced strengths and improvements | 4 | 0.64 | 5 | 0.50 | 4 | 0.61 |
| Clear and understandable feedback | 5 | 0.87 | 5 | 0.44 | 5 | 0.65 |

### F.2.2 Adaptive Questioning

This metric evaluated the quality of follow-up questions when they occurred. A strong follow-up responded to what the user had actually said, targeted the specific missing or unclear part, and avoided generic wording or unnecessary repetition. The metric did not reward the mere presence of follow-up questions and did not penalize the examiner for remaining neutral or for not correcting misconceptions.

Table 13: Speech-modality item results by speech configuration group.

| | Init. | | Upd. | | All | |
|---|---|---|---|---|---|---|
| **Item** | **Mdn** | **SD** | **Mdn** | **SD** | **Mdn** | **SD** |
| Natural reasoning through speech | 3 | 1.15 | 3 | 1.40 | 3 | 1.29 |
| Felt like an oral exam | 4 | 1.26 | 4 | 0.89 | 4 | 1.07 |
| Easier to show understanding | 2 | 1.04 | 3 | 1.05 | 3 | 1.05 |
| Did not increase nervousness | 3 | 1.38 | 3 | 1.17 | 3 | 1.23 |
| No ASR or audio issues affected expression | 1 | 0.96 | 3 | 1.38 | 2 | 1.47 |
| Did not prefer typing | 3 | 0.95 | 2 | 1.18 | 3 | 1.09 |
| Did not interrupt while speaking | 1 | 0.38 | 2 | 1.59 | 1.5 | 1.46 |
| Felt like a natural conversation | 2 | 0.75 | 3 | 1.04 | 3 | 1.20 |
| Responded quickly enough | 4 | 1.01 | 4 | 1.03 | 4 | 1.01 |

Table 14: Perceived-suitability item results by speech configuration group.

| | Init. | | Upd. | | All | |
|---|---|---|---|---|---|---|
| **Item** | **Mdn** | **SD** | **Mdn** | **SD** | **Mdn** | **SD** |
| Comfortable with system for oral exam | 2 | 1.14 | 3 | 1.21 | 2.5 | 1.21 |
| Useful for statistics practice | 5 | 1.44 | 5 | 0.75 | 5 | 1.09 |
| Would recommend as study tool | 4 | 1.60 | 4 | 1.00 | 4 | 1.30 |
| Identify strengths and weaknesses | 4 | 1.41 | 4 | 0.80 | 4 | 1.08 |
| Useful additional assessment tool | 4 | 1.51 | 4 | 0.86 | 4 | 1.16 |
| Could replace parts of oral exam | 2 | 1.04 | 4 | 1.15 | 3 | 1.19 |
| Fair indication of understanding | 3 | 1.12 | 3 | 1.07 | 3 | 1.14 |
| Acceptable as graded evidence | 2 | 1.56 | 3 | 1.11 | 3 | 1.31 |
| Preferred over written exam | 1 | 0.78 | 2 | 1.15 | 2 | 1.07 |

Table 15: Shared interpretation of the score anchors.

| Score | Interpretation |
|---|---|
| 1.00 | The required behavior was consistently strong throughout the interaction. |
| 0.75 | The behavior was mostly strong, with only minor problems or missed opportunities. |
| 0.50 | The behavior was mixed or only partially sufficient. |
| 0.25 | The behavior was mostly weak and important problems occurred repeatedly. |
| 0.00 | The behavior was absent or repeatedly violated. |

#### F.2.3 Neutrality and Non-Leading Prompting

This metric evaluated whether the examiner assessed the user's independent understanding without supplying expected-answer content or influencing the response. Hints, answer leakage, leading questions, correction, praise, reassurance, discouragement, and explicit evaluation reduced the score. Remaining in the examiner role and allowing misconceptions to stand during the assessment were treated as appropriate behavior.

#### F.2.4 Content Alignment

This metric evaluated whether the questions remained within the three intended tasks and their subtopics and avoided unrelated topics or overly minor details. The assessment goals were included in the metric definition. The rubric and reference document were supplied as additional context for this metric.

#### F.2.5 Assessment Coherence

This metric evaluated the conversation as a complete assessment sequence. Strong behavior included a logical order within tasks, understandable announcements when moving to a new topic, and responses that remained connected to what the user said. Abrupt topic changes, disjointed progression, or failure to respond to interaction problems reduced the score.

#### F.2.6 Feedback Quality

This metric evaluated whether the final feedback accurately represented what the user did and did not demonstrate. Strong feedback was specific, clear, balanced, transcript-grounded, and useful for improvement. Generic, vague, inaccurate, or weakly grounded feedback received lower scores. Particular attention was given to whether the feedback distinguished the user's own demonstrated knowledge from information introduced by the examiner.

### F.3 Results by Model for User Conversations

Table 16 reports the mean scores by judge model for the user conversations. Table 17 does so for standard deviations.

Table 16: Mean LLM-as-judge scores by judge model and speech configuration group.

| | GPT | | | Mistral | | | Claude | | | Combined | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Dimension** | **In.** | **Upd.** | **All** | **In.** | **Upd.** | **All** | **In.** | **Upd.** | **All** | **In.** | **Upd.** | **All** |
| Diagnostic questioning | .78 | .85 | .82 | .80 | .82 | .81 | .46 | .61 | .56 | .68 | .76 | .73 |
| Adaptive questioning | .71 | .78 | .75 | .68 | .87 | .79 | .48 | .52 | .50 | .62 | .72 | .68 |
| Neutrality | .93 | .92 | .92 | 1.00 | .97 | .98 | .78 | .86 | .83 | .90 | .92 | .91 |
| Content alignment | .96 | .95 | .95 | .98 | 1.00 | .99 | .98 | .98 | .98 | .97 | .98 | .97 |
| Assessment coherence | .80 | .84 | .83 | .64 | .77 | .72 | .66 | .89 | .80 | .70 | .83 | .78 |
| Feedback quality | .91 | .93 | .92 | .95 | .93 | .94 | .91 | .88 | .89 | .92 | .91 | .92 |

Table 17: Standard deviations of LLM-as-judge scores by judge model and speech configuration group.

| | GPT | | | Mistral | | | Claude | | | Combined | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Dimension** | **In.** | **Upd.** | **All** | **In.** | **Upd.** | **All** | **In.** | **Upd.** | **All** | **In.** | **Upd.** | **All** |
| Diagnostic questioning | .18 | .11 | .14 | .21 | .13 | .16 | .19 | .15 | .18 | .25 | .17 | .20 |
| Adaptive questioning | .21 | .18 | .19 | .32 | .21 | .27 | .26 | .19 | .22 | .28 | .24 | .26 |
| Neutrality | .15 | .18 | .17 | .00 | .13 | .10 | .23 | .14 | .18 | .18 | .16 | .16 |
| Content alignment | .04 | .09 | .08 | .04 | .02 | .03 | .06 | .05 | .05 | .05 | .06 | .06 |
| Assessment coherence | .16 | .11 | .13 | .13 | .16 | .16 | .29 | .12 | .23 | .21 | .14 | .18 |
| Feedback quality | .01 | .03 | .03 | .05 | .06 | .05 | .08 | .10 | .09 | .06 | .07 | .06 |

### F.4 Results by Model for Simulated Conversations

Table 18 reports the scores from each judge model for the simulated conversations.

### F.5 LLM Agreement for User and Simulated Conversations

Table 19 reports pairwise absolute differences between the three LLM judges for the user and simulated conversations. Differences were calculated between judge scores for the same conversation and evaluation dimension on the 0–1 scale, with lower values indicating closer absolute agreement.

For the user conversations, GPT and Mistral showed the closest agreement overall (mean absolute difference = .092), while differences were larger for GPT–Claude (.173) and Claude–Mistral (.180). Agreement varied across dimensions: scores were particularly close for content alignment and feedback quality, whereas larger differences occurred for diagnostic and adaptive questioning. For the simulated conversations, pairwise differences were generally smaller, with overall differences of .104 for GPT–Claude, .109 for Claude–Mistral, and .063 for GPT–Mistral. This indicates that the judges agreed more closely in their absolute scores for the simulated conversations than for the user conversations, particularly for adaptive questioning.

### F.6 Recurring Judge-Reasoning Patterns

Table 20 summarizes the recurring patterns identified in the judges' reasoning across the evaluated conversations.

## G Representative Conversation Excerpts

The excerpts below illustrate specific behaviors observed in the qualitative conversation analysis. They are examples rather than frequency estimates. In the excerpts, **E** denotes ACA-Bot and **U** denotes the user.

### G.1 Targeted Follow-Up after an Incomplete Answer

**E:** "And what does a p-value mean in the context of hypothesis testing?"
**U:** "P-value stands for, uh, statistical significance."
**E:** "What does it tell you about the data if the p-value is very small?"

The response was relevant but incomplete. ACA-Bot's follow-up targeted the missing interpretation and provided another opportunity for the user to demonstrate their reasoning. This example relates to diagnostic and adaptive questioning.

### G.2 Failure to Repair User Confusion

**E:** "What do the null hypothesis and the alternative hypothesis represent?"
**U:** "Hello, hello."
**E:** "Let's move on: what does a p-value mean?"
**U:** "Hallo, ist das hier das Mikrofon?" ["Hello, is

Table 18: LLM-as-judge scores for the simulated student conversations ($N = 40$).

| | GPT | | Mistral | | Claude | | Combined | |
|---|---|---|---|---|---|---|---|---|
| **Dimension** | **M** | **SD** | **M** | **SD** | **M** | **SD** | **M** | **SD** |
| Diagnostic questioning | .90 | .07 | .87 | .09 | .70 | .15 | .82 | .10 |
| Adaptive questioning | .91 | .07 | .96 | .07 | .96 | .10 | .94 | .08 |
| Neutrality | .98 | .07 | .99 | .09 | .87 | .23 | .95 | .13 |
| Content alignment | .97 | .05 | .99 | .04 | .99 | .04 | .98 | .04 |
| Assessment coherence | .92 | .06 | .86 | .10 | .94 | .13 | .91 | .10 |
| Feedback quality | .92 | .02 | .95 | .05 | .89 | .07 | .92 | .05 |

Table 19: Pairwise absolute differences between LLM judge scores for user and simulated conversations.

| | User | | | Simulated | | |
|---|---|---|---|---|---|---|
| **Dimension** | **GPT–Claude** | **Claude–Mistral** | **GPT–Mistral** | **GPT–Claude** | **Claude–Mistral** | **GPT–Mistral** |
| Diagnostic questioning | .276 | .268 | .096 | .208 | .205 | .088 |
| Adaptive questioning | .298 | .350 | .145 | .093 | .056 | .087 |
| Neutrality | .176 | .179 | .091 | .138 | .149 | .039 |
| Content alignment | .057 | .029 | .045 | .030 | .019 | .026 |
| Assessment coherence | .160 | .171 | .138 | .097 | .155 | .093 |
| Feedback quality | .069 | .085 | .036 | .055 | .073 | .042 |
| **Overall** | **.173** | **.180** | **.092** | **.104** | **.109** | **.063** |

*Note.* Values represent mean absolute differences between pairs of LLM judges when scoring the same conversation on the 0–1 scale; lower values indicate closer absolute agreement. User results are based on 34 completed conversations, with files marked as `_short` excluded. Simulated results are based on 40 conversations. For simulated content alignment, one missing GPT evaluation resulted in 39 matched conversations.

Table 20: Recurring patterns in the judge reasoning.

| **Dimension** | **Reasons for higher scores** | **Reasons for lower scores** |
|---|---|---|
| Diagnostic questioning | Targeted probing and sufficient opportunities to explain. | Missed probing opportunities or moving on before understanding became visible. |
| Adaptive questioning | Follow-ups that directly built on the preceding answer. | Generic follow-ups or failure to repair confusion, interruptions, or incomplete turns. |
| Neutrality | Staying in the examiner role without correction or answer leakage. | Confirmation, praise, evaluation, hints, or leading wording. |
| Content alignment | Questions remained within the intended tasks and subtopics. | Rare minor deviations or questions beyond the intended scope. |
| Assessment coherence | Logical progression and understandable topic transitions. | Abrupt progression, disrupted turns, or failure to respond to interaction problems. |
| Feedback quality | Specific, balanced, and transcript-grounded feedback. | Imprecise descriptions of misconceptions or credit for knowledge introduced by the examiner. |

this the microphone?"]
**E:** "Let's continue: [new question]"

The user appeared to be testing the microphone rather than answering the statistics question. ACA-Bot continued with the planned sequence instead of repairing the interaction, affecting adaptive questioning and assessment coherence.

### G.3 Moving On before the User Finished

**U:** "Um, the null hypothesis—"
**E:** "And what does a p-value mean in hypothesis testing?"

The user was interrupted before completing the answer. This reduced the evidence available about their understanding and affected diagnostic questioning and coherence.

### G.4 Inconsistent Follow-Up Opportunities

**E:** "And how is standard deviation related to variance?"
**U:** "That is like how the general variance is."
**E:** "Can you say how standard deviation is calculated from variance?"

*Compared with:*
**U:** "The standard deviation is calculated based on the variance."
**E:** Ends the exam and gives feedback.

The first vague answer received a clarification opportunity, whereas a similarly limited answer did not. This example illustrates why prompting

consistency requires comparison across similar responses.

### G.5 Evaluation during the Assessment

> **U:** "I think the null hypothesis states there is no effect or no difference, while the alternative hypothesis states there is an effect or difference. Is that correct?"
> **E:** "Yes. Now, what does a p-value mean in the context of hypothesis testing?"

The confirmation revealed whether the answer was correct while the assessment was still in progress, weakening neutrality.

### G.6 Leading Follow-Up and Feedback Alignment

> **E:** "What do the null hypothesis and the alternative hypothesis represent?"
> **U:** "The null hypothesis represents, um,"
> **E:** "What does it usually mean in terms of no effect or no difference?"

The follow-up supplied part of the expected answer rather than only asking for clarification. A related feedback-alignment issue occurred in the following exchange:

> **E:** "What does a very small p-value say about how likely the data are if $H_0$ were true?"
> **U:** "It means that we, um, have really contrasting results."
> *Feedback:* "You showed some partial understanding of hypothesis testing, especially that [...] a very small p-value means the data are unlikely if $H_0$ were true."

The feedback credited an interpretation that had largely been introduced in the examiner's question. This made it difficult to separate independently demonstrated knowledge from examiner-provided content.