2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION
Paper ID #51989

Exploring Authentic Assessment of Student Spontaneous Thinking Elicited through Open-Ended Questions: From Writing-to-Learn to Time-Windowed Thinking-Aloud-to-Learn

Prof. Wei Zheng, Jackson State University

Dr. Wei Zheng is a Professor of Civil Engineering at Jackson State University. He received his Ph.D. in Civil Engineering from the University of Wisconsin–Madison in 2001 and has more than ten years of industrial experience before joining academia in 2005. His research mainly focuses on applying probabilistic machine learning to decision-making under uncertainty. His work integrates expertise in civil engineering, AI model development, and educational data analytics to improve engineering decision systems and personalized learning support. His current research in civil engineering focuses on using generative AI technologies to extract knowledge from historical engineering reports and records to support data-driven decision-making. He is also leading two ongoing NSF-funded projects and conducts research in two key areas: (1) Engaging students in open-ended responses and reflective writing to strengthen higher-order thinking and learning skills, and (2) Developing AI-powered automatic assessment tools to evaluate these responses, with the ultimate goal of providing adaptive support with bias mitigation and cultural relevance enhancement through AI-powered education to promote equitable and effective learning.

Gokalp Ayar, Jackson State University

Gokalp Ayar is a senior undergraduate student in Computer Science at Jackson State University and a research assistant in the AI and Learning Systems Lab. His work focuses on designing and developing AI-powered educational platforms, including learning management systems that leverage automatic speech recognition and data analytics to support authentic assessment of student learning. His research interests include machine learning, data science, and educational technology, with an emphasis on building scalable, data-driven systems for improving learning outcomes and assessment practices.

Exploring Authentic Assessment of Student Spontaneous Thinking Elicited through Open-Ended Questions: From Writing-to-Learn to Time-Windowed Thinking-Aloud-to-Learn

Wei Zheng[1] and Gokalp Ayar[2]

[1] Department of Civil and Environmental Engineering, Jackson State University, Jackson, Mississippi, USA

[2] Department of Computer Science, Jackson State University, Jackson, Mississippi, USA

Abstract

Writing-to-Learn (WTL) activities use open-ended prompts to help students articulate understanding, reflect on learning strategies, and make their thinking visible. However, the increasing availability of generative AI tools has created a validity challenge for written WTL artifacts, because students may produce polished written responses without fully engaging in their own reasoning. This study explores a complementary approach based on time-windowed oral responses, in which students verbalize their thinking in response to open-ended WTL prompts. The approach is designed to preserve the reflective purpose of WTL while reducing opportunities for AI-generated substitution.

A cloud-based oral-response pipeline was developed to collect students' time-constrained responses, generate automatic transcripts, and make both audio and text available for analysis. The pipeline was deployed in two undergraduate engineering courses to collect responses associated with learning plans, learning evaluations, and learning reflections. This study focuses on a necessary first step: evaluating whether automatically generated transcripts accurately preserve students' spoken responses and intended meaning.

Transcription fidelity was evaluated using word error rate, character error rate, semantic similarity based on embedding representations, error-type analysis, and student self-reported transcript accuracy ratings. Results from 19 students and 51 rated oral-response submissions showed a word error rate of 7.85%, a character error rate of 3.78%, and an average semantic similarity score of 0.93. Student ratings also indicated generally positive perceptions of transcript accuracy, with 82.4% of rated submissions receiving four or five stars.

These findings suggest that time-windowed oral responses can be captured and transcribed with sufficient fidelity to support further analysis within WTL contexts. The study establishes a foundation for integrating transcribed oral responses into future AI-powered, rubric-based assessment pipelines that can evaluate cognitive, metacognitive, and motivational dimensions of student learning, identify learning strengths and weaknesses, and support adaptive feedback in the era of generative AI.

1. Introduction

The rapid adoption of generative artificial intelligence tools, particularly large language models capable of producing fluent and well-structured explanations, has fundamentally altered the landscape of open-ended assessment in higher education [1]. In engineering education, open-ended responses have long been used to elicit students' reasoning, conceptual understanding, and learning processes. However, when students can readily generate written responses using AI tools, instructors face increasing difficulty in determining whether submitted work reflects students' own thinking or externally generated content. This challenge is especially acute in asynchronous learning environments, where instructors have limited opportunities to observe students' reasoning in real time and where traditional safeguards for assessment authenticity are difficult to enforce consistently.

Writing-to-Learn (WTL) pedagogy has been widely adopted in engineering and STEM contexts because it positions learning as an active process of meaning construction rather than passive content consumption [2]–[4]. Through structured activities such as learning plans, learning evaluations, and learning reflections, WTL encourages students to articulate their understanding, monitor their learning, and reflect on the effectiveness of their strategies. These activities target cognitive, metacognitive, and motivational dimensions of learning and align with broader learning-science frameworks that emphasize constructive engagement, including explanation, self-monitoring, and reflection [5]. Traditionally, WTL artifacts are collected as written responses through learning management systems or online forms. While effective in many instructional contexts, written WTL artifacts are now increasingly vulnerable to AI-mediated text generation, raising concerns about their validity as evidence of students' own thinking.

In response to this challenge, this study explores an alternative yet complementary modality for WTL activities: time-constrained oral responses designed to function as structured, think-aloud–like articulations of student thinking. Think-aloud approaches have a long history in cognitive science and learning research as a means of making learners' cognitive and metacognitive processes observable by capturing verbalized reasoning as it unfolds during task engagement [6]. Unlike polished written responses produced after extended deliberation, time-windowed oral responses emphasize immediacy and spontaneity, reducing opportunities for external mediation while preserving access to students' reasoning processes. Importantly, the approach adopted in this study does not seek to replace writing-to-learn practices. Rather, it integrates structured oral responses with existing WTL activities, treating oral articulation as a complementary mechanism for eliciting authentic evidence of learning and thinking in the era of generative AI.

Despite their pedagogical value, think-aloud and oral-response approaches have not been widely adopted at scale in routine classroom assessment due to practical constraints, including instructor workload, data management challenges, and the difficulty of systematically analyzing large volumes of spoken responses. To address these barriers, we developed a cloud-based, time-constrained oral-response pipeline that automates prompt delivery, audio recording, and transcription of students' oral responses. This pipeline enables instructors to incorporate oral, think-aloud–like responses into WTL activities in asynchronous settings while maintaining consistency and scalability. The resulting transcripts create a bridge between oral articulation and

established text-based analysis practices, making it possible to evaluate whether transcribed oral responses can serve as reliable representations of students' original speech.

The present study focuses on a preliminary but critical step in validating this pipeline: evaluating the accuracy of automatically transcribed text relative to students' original oral responses. At this stage, the study does not evaluate automated grading, diagnosis of learning weaknesses, or feedback generation. Instead, it examines whether automatically generated transcripts preserve students' intended meaning with sufficient fidelity to support future learning analysis. This boundary is important because reliable transcription is a necessary prerequisite before oral responses can be used in AI-powered, rubric-based automatic assessment systems.

In the broader project context, the transcription pipeline is intended to connect with an AI-powered, rubric-based automatic assessment system currently being developed for students' written WTL responses. Once oral responses are accurately transcribed, the resulting text can be analyzed using established cognitive, metacognitive, and motivational rubrics to assess students' learning processes, identify learning strengths and weaknesses, and support targeted adaptive feedback. In this way, the oral-response pipeline creates a bridge between students' spontaneous spoken reasoning and future AI-powered formative assessment.

Accordingly, the study is guided by the following research questions:

RQ1. How accurately can automatic transcription capture students' spontaneous oral reasoning in time-windowed WTL prompts?

RQ2. To what extent do transcription errors affect semantic fidelity?

RQ3. How do students perceive the accuracy and adequacy of system-generated transcripts?

By addressing these questions, this paper contributes an empirical evaluation of a scalable oral-response pipeline designed to support WTL pedagogy under generative AI constraints. The results provide foundational evidence regarding transcription fidelity, which is a necessary prerequisite for future work on AI-powered automatic assessment, diagnostic feedback, and adaptive learning support based on students' oral articulation of their thinking.

2. Background and Related Work

Writing-to-Learn (WTL) pedagogy is grounded in the idea that learning is strengthened when students actively articulate understanding, evaluate learning outcomes, and reflect on their learning processes. Foundational work on WTL emphasizes that writing functions as a cognitive activity through which learners construct meaning and make their thinking visible, rather than merely a means of reporting knowledge [2]. In STEM and engineering education, WTL has been implemented through structured prompts that ask students to explain concepts, justify reasoning, and reflect on learning strategies, with prior studies showing its value for supporting conceptual understanding and reflective learning [3], [4]. These practices align with learning-science frameworks that associate constructive engagement—such as explanation, monitoring, and reflection—with deeper learning [5].

The validity of written WTL artifacts, however, has become increasingly uncertain with the widespread availability of generative AI tools. When students can generate fluent written explanations with minimal cognitive effort, written responses alone no longer reliably represent students' own thinking, particularly in asynchronous contexts. Recent analyses of large language models in education highlight this tension between instructional opportunity and assessment risk, motivating the need for alternative or complementary approaches that preserve authenticity while retaining open-ended pedagogy [1].

Think-aloud approaches provide a relevant methodological foundation for addressing this challenge. Originating in cognitive psychology, think-aloud protocols require learners to verbalize their reasoning while engaging in a task, thereby making cognitive and metacognitive processes observable in real time [6]. Compared with post-hoc written explanations, think-aloud responses can capture immediacy, self-monitoring, and strategy use as they occur. Prior work in engineering education has used think-aloud methods to examine student reasoning during concept-inventory assessments. For example, Davis and Hill analyzed students' understanding of force using the Dynamics Concept Inventory, think-aloud, and confusion matrices, and also examined student think-alouds during a Dynamics Concept Inventory exam [7], [8]. However, these studies also illustrate a key practical challenge: obtaining accurate and usable transcripts can be difficult and can limit the scalability of think-aloud analysis. This limitation is directly relevant to the present study, which evaluates whether automatically generated transcripts can provide a reliable basis for analyzing students' oral responses at classroom scale.

Recent advances in automatic speech recognition (ASR) have made scalable oral-response approaches increasingly feasible. Models such as Whisper have substantially improved the ability to convert spoken responses into text [9]. Educational applications of ASR have also been explored in pronunciation learning and automated speaking assessment [10], [11]. Parallel work in learning analytics has examined computational analysis of think-aloud data to study self-regulated learning, while also reinforcing that transcript quality remains a bottleneck for scaling such approaches [12]. However, most prior educational uses of ASR focus on pronunciation, speaking performance, or general transcript generation. Fewer studies directly examine whether ASR-generated transcripts preserve the meaning of students' reflective, learning-oriented responses well enough to support WTL-related analysis.

Equity considerations further underscore the need to evaluate transcription accuracy in the context where the technology will be used. Prior research has documented systematic disparities in ASR performance across speaker populations, including higher error rates for Black speakers and speakers of African American English [13]. In response to such concerns, Howard University and Google Research have developed Project Elevate Black Voices, a data-focused initiative intended to improve speech recognition for African American English and support more inclusive voice technologies [14], [15]. This body of work highlights that transcription accuracy cannot be assumed a priori, especially in minority-serving educational contexts such as HBCUs. Therefore, validating ASR performance is necessary before transcribed speech is used for learning analytics, rubric-based assessment, or adaptive feedback.

Positioned within this landscape, the present study addresses a focused but necessary gap: evaluating whether automatically transcribed text accurately preserves students' oral, think-aloud–like responses collected as part of WTL activities. Rather than proposing a new ASR model or claiming to validate a complete automated assessment system, this study establishes foundational evidence regarding transcription fidelity for learning plans, learning evaluations, and learning reflections. Such evidence is a prerequisite for future integration of oral-response transcripts into AI-powered, rubric-based assessment pipelines that can evaluate students' cognitive, metacognitive, and motivational learning processes.

3. Pedagogical Framework

This study is grounded in a Writing-to-Learn (WTL) pedagogical framework that emphasizes students' active articulation of understanding, reflection on learning strategies, and regulation of learning processes. Within this framework, student learning is understood as the coordinated development of cognitive, metacognitive, and motivational processes. These three perspectives shape how students understand course content, monitor and adjust their learning strategies, and sustain engagement in demanding engineering tasks. They are treated as analytically distinct but interrelated dimensions of learning that can be made observable through carefully designed WTL activities rather than inferred solely from performance outcomes.

The cognitive perspective focuses on what students understand and how they apply domain knowledge. In this study, cognitive regulation is reflected in students' articulation of key concepts, explanation of conceptual relationships, and justification of problem-solving approaches in response to open-ended prompts. The metacognitive perspective emphasizes how students regulate their learning processes, including planning learning activities, monitoring understanding, evaluating strategy effectiveness, and adapting approaches based on feedback. Metacognitive regulation is reflected in students' preparation plans, identification of learning weaknesses, and reflection on strategy effectiveness. The motivational perspective captures why students engage and persist, including goal orientation, perceived task value, and regulation of effort. Motivational regulation is reflected in students' statements about learning goals, relevance of course content, and strategies for maintaining engagement.

Although these perspectives are closely related, they serve distinct analytical roles within the WTL framework. The cognitive perspective addresses what students understand, the metacognitive perspective addresses how students manage their learning, and the motivational perspective addresses why students sustain effort. Making these distinctions explicit allows student responses to be examined as evidence of different learning processes rather than collapsed into a single aggregate indicator. This distinction is important for future rubric-based assessment because each dimension requires different interpretive criteria and may support different types of feedback.

The WTL framework implemented in this study follows a cyclic instructional structure embedded within regular course instruction. Each instructional cycle consists of five recurring components designed to support different phases of learning: (1) practice problems requiring written rationales, (2) a pre-quiz learning plan, (3) a quiz with written rationales and graded feedback, (4) a post-quiz learning evaluation, and (5) a post-quiz learning reflection. Practice

problems and learning plans are completed prior to quizzes to support preparation and planning, while learning evaluations and reflections are completed after quizzes to support monitoring, evaluation, and adaptation. This structure is repeated across instructional cycles, allowing students multiple opportunities to engage with the same reflective and self-regulatory processes.

Each WTL component is intentionally designed to elicit specific learning perspectives. Practice and quiz problem rationales primarily target the cognitive perspective by prompting students to explain reasoning and justify selected solutions. Learning plans emphasize motivational and metacognitive regulation by requiring students to articulate goals, anticipated difficulties, and preparation strategies. Learning evaluations focus on cognitive calibration by prompting students to judge whether intended learning goals were achieved and why. Learning reflections emphasize metacognitive and motivational regulation by encouraging students to analyze learning strategies, effort, and planned adjustments for future learning.

To address concerns about assessment authenticity in the generative AI era, selected open-ended prompts within the learning plan, learning evaluation, and learning reflection components were administered using a time-constrained oral response format. These oral responses are designed to function as structured, think-aloud–like articulations of student thinking, requiring students to verbalize reasoning within a predetermined time window. Unlike polished written responses produced after extended deliberation, time-constrained oral responses emphasize immediacy and spontaneous articulation, reducing the opportunity for AI-generated substitution while preserving the reflective purpose of WTL.

Importantly, the oral response component is integrated within, rather than replacing, the established WTL framework. Students continue to engage in written WTL activities aligned with the cognitive, metacognitive, and motivational perspectives, while oral responses provide a complementary modality for capturing student thinking under constrained conditions. Audio recordings are automatically transcribed and retained alongside written artifacts, creating a parallel textual representation of students' spoken responses. The present study focuses specifically on whether these automatically generated transcripts accurately preserve students' original oral responses across WTL-related prompts, establishing a necessary foundation for future AI-powered, rubric-based assessment and adaptive feedback grounded in this pedagogical framework.

4. System Design of Time-Constrained Oral Response Transcription Pipeline

The study employs a cloud-deployed transcription pipeline designed to collect time-windowed student oral responses, automatically transcribe recorded audio, and prepare paired transcripts for evaluation of transcription fidelity. The pipeline was developed to support open-ended WTL prompts that elicit students' spontaneous thinking while enforcing temporal constraints intended to reduce delayed composition and external assistance.

The system architecture integrates a web-based frontend, a backend workflow service, an automatic speech recognition component, and a cloud-based data storage layer. The frontend interface, deployed using a cloud hosting service, provides access for both instructors and students and supports audio recording within a browser environment. The backend service

manages request handling, enforces timing constraints, assigns file identifiers, stores metadata, and coordinates transcription requests. Metadata, audio identifiers, transcripts, and student-provided information are stored in a centralized database that supports retrieval for instructional review and research analysis.

Student oral responses are recorded automatically through the web interface. Each response is saved as an individual audio file and assigned a unique identifier upon submission. This identifier is used to track the response throughout the transcription and evaluation process, ensuring consistent pairing among audio files, system-generated transcripts, student ratings, and reference transcripts. Audio recordings are not modified after submission.

Speech-to-text transcription is performed automatically after submission. The backend service forwards each audio file to a cloud-based automatic speech recognition service using the OpenAI Whisper API. Transcription is conducted using the Whisper model through API-based inference [9]. For each submission, the system returns a machine-generated transcript without manual post-editing or correction. These transcripts are used directly in the evaluation reported in this study.

Based on system testing, a typical four-minute student oral response required approximately 30-60 seconds from the end of recording to transcript availability in the online platform. This processing time includes audio upload and saving, Whisper API transcription, database storage, and transcript display. The processing time may vary depending on internet speed, audio-file size, server load, and API response time.

The pipeline enforces timing constraints at two levels. Instructors define an overall availability window during which an assignment may be accessed. In addition, instructors specify a fixed response time limit for each oral-response question. Once a student begins recording, a countdown timer is displayed, and recording terminates automatically when the configured time limit is reached. These constraints are enforced by the backend service and apply uniformly across students, allowing responses to be completed asynchronously while preserving temporal control.

Instructors access the system through a secure web interface. Within the instructor panel, instructors create assignments, select oral response as the response modality, and configure availability windows and response time limits. After configuration, the system generates a unique web link that can be distributed to students in a manner similar to common form-based tools. Following submission, instructors can view response status and access system-generated transcripts for instructional or research use.

Students access the assignment using the provided link and are prompted to enter a de-identified student ID code. After viewing the open-ended prompt, students begin recording their response. The interface displays the remaining time during recording. Once the time limit expires, recording stops automatically and the response is submitted for transcription. Students are then shown the system-generated transcript, allowing them to review the output before final submission. The audio recording and transcript are stored together using the unique identifier assigned at submission.

For evaluation purposes, system-generated transcripts were paired with reference transcripts created independently by manually transcribing the original audio recordings. Reference transcripts were prepared without access to the system-generated transcripts to reduce potential bias. Transcript pairs were matched using their unique identifiers and prepared for quantitative accuracy analysis and semantic similarity evaluation.

The current implementation was deployed as a cloud-based prototype to support rapid development, remote access, and classroom testing. However, because students' oral responses may contain sensitive learning data, future versions of the pipeline will explore institutional or local deployment options. In such a configuration, audio storage, transcription, transcript management, and future assessment modules could be hosted on university-controlled servers or in a private cloud environment, reducing the need to transmit student audio data to external services.

The pipeline is intentionally limited in scope. It supports collection, timing enforcement, transcription, transcript display, and transcript preparation, but it does not currently perform automated grading, learning diagnosis, or feedback generation. These capabilities are outside the scope of the present study. However, the pipeline was designed so that validated transcripts can later be passed to an AI-powered, rubric-based assessment module for evaluating students' cognitive, metacognitive, and motivational learning processes and generating formative feedback.

5. Participants and Data Collection

The study was conducted in two undergraduate engineering courses in which Writing-to-Learn (WTL) activities were already integrated into regular instruction. The courses employed open-ended prompts associated with learning plans, learning evaluations, and learning reflections. These prompts were aligned with course content and instructional goals and were part of routine coursework rather than a separate experimental intervention.

A total of 19 unique students provided oral responses for the data reported in this study. Students accessed the time-windowed oral-response prompts asynchronously using instructor-provided links and completed the responses individually. No student names were collected through the system. Submissions were identified using de-identified student ID codes and course identifiers.

Each participating student was eligible to complete oral-response activities associated with three WTL components: learning plans, learning evaluations, and learning reflections. Because submission patterns varied, not all students completed all three activities. Across the participating students, the dataset included 51 oral-response submissions with student self-reported transcript accuracy ratings.

For each oral-response submission, the collected data included (1) an audio recording of the time-windowed oral response, (2) an automatically generated transcript produced by the speech recognition component, (3) metadata identifying the course and response type, and (4) a student self-reported rating of transcript accuracy. These activity types provided varied contexts for

examining whether automatic transcription preserved students' spoken responses across different learning-related prompts.

Data collection focused exclusively on capturing students' spontaneous verbal responses and evaluating transcription fidelity. No intervention was introduced to alter instructional content, and no additional training was provided to students beyond standard instructions for completing oral responses. The study did not grade or evaluate students based on transcription accuracy. All collected data were used solely for research purposes related to evaluating whether the generated transcripts accurately represented students' oral responses.

6. Evaluation Method

This study employed a descriptive evaluation approach to examine the accuracy of automatically transcribed text generated from students' time-constrained oral responses. The focus of the methods is on how oral response data were collected, transcribed, reviewed, and evaluated for transcription fidelity. The study does not evaluate automated grading, learning diagnosis, or adaptive feedback generation at this stage.

6.1 Oral Response Collection Procedure

A cloud-based oral-response pipeline, described in the previous section, was deployed and made accessible to course instructors. Instructors used the platform to post open-ended questions aligned with Writing-to-Learn (WTL) activities, including learning plans, learning evaluations, and learning reflections. For each question, the instructor generated a unique access link, functionally similar to a Google Form link, and distributed it to students.

When students accessed the link, they were prompted to enter identifying information limited to their de-identified student ID code and course identifier. No student names were collected through the platform. After entering the required information, students were presented with a single open-ended prompt and informed of the fixed four-minute response window. Once the recording process began, the system displayed a countdown timer. Students were instructed to provide their response orally within the time limit. When the time expired, the recording was automatically terminated by the system.

All students received the same general instructions and completed responses under the same four-minute time constraint. This procedure helped standardize the response condition while preserving the spontaneous nature of students' oral articulation. The pipeline recorded each student's complete oral response and securely stored the original audio file. This design ensured a consistent response window across students and eliminated the possibility of extending responses beyond the predefined time limit.

6.2 Automatic Transcription and Student Review

Upon completion of the recording, the system automatically processed the audio file using the integrated speech recognition component to generate a text transcript of the student's oral response. The generated transcript was presented to the student after recording and automatic

processing. Students were given the opportunity to review the transcript while the content was still fresh in memory, but they were not allowed to edit or modify the transcribed text.

To support evaluation of transcription accuracy, students were asked to provide a self-reported rating of how accurately the transcript reflected what they intended to say in their oral response. This rating was collected using a five-point scale, where higher values indicated greater perceived accuracy of the transcript relative to the student's spoken response. The rating served as a complementary data source to the objective transcription accuracy analysis conducted by the research team. After the review and rating step, the student's submission—including the audio recording, system-generated transcript, and self-reported accuracy rating—was finalized and stored in the system.

6.3 Researcher Access and Reference Transcript Preparation

All submitted data were stored in the backend of the pipeline and made accessible to the research team for analysis. For each response, the dataset included the original audio recording, the automatically generated transcript, metadata identifying the course and response type, and the student's self-reported transcript accuracy rating.

For the purposes of this study, the research team focused exclusively on evaluating transcription accuracy. The audio recordings served as the reference representation of students' original oral responses, while the automatically generated transcripts served as the system output. Reference transcripts were prepared by listening to the original audio recordings and producing human-verified text representations of the spoken responses. These reference transcripts were then paired with the system-generated transcripts using unique response identifiers.

No automated scoring, diagnostic classification, or feedback generation was performed in this stage of the study. The methods were intentionally limited to assessing whether the transcription output preserved students' spoken content with sufficient fidelity to support potential downstream analysis in future AI-powered, rubric-based assessment work.

Students completed the recordings using typical recording devices available to them, such as built-in microphones on smartphones, laptops, or desktop computers; no special recording hardware was required. Variations in speaking volume, clarity, pauses, device quality, microphone quality, and environmental noise may have influenced transcription accuracy and were therefore considered potential sources of variability in interpreting the results.

7. Analysis and Measures

The analysis was designed to address the three research questions by evaluating transcription accuracy, semantic fidelity, and students' perceptions of transcript adequacy. Consistent with the preliminary and feasibility-focused nature of the study, the analysis focuses on transcription fidelity rather than on automated scoring of student learning, diagnosis of misconceptions, or feedback generation.

To address RQ1, transcription accuracy was evaluated by comparing system-generated transcripts with human-prepared reference transcripts. The original audio recordings were treated as the source representation of students' spoken responses, and the human-prepared reference transcripts served as the comparison standard for quantitative analysis. Word Error Rate (WER) and Character Error Rate (CER) were used as the primary surface-level accuracy measures. WER captures word-level substitutions, deletions, and insertions relative to the reference transcript, while CER provides a finer-grained character-level measure of transcription differences [16].

To address RQ2, the study examined whether transcription errors affected preservation of students' intended meaning. Semantic similarity was computed by converting both the reference transcript and the system-generated transcript into vector embeddings and then calculating cosine similarity between the two representations [17], [18]. This measure was included because exact word-level matching may not fully capture whether a transcript preserves the meaning of students' open-ended and reflective responses. In addition, transcription errors were categorized into substitutions, deletions, and insertions to determine whether errors primarily reflected surface-level wording differences or potential loss of substantive content.

To address RQ3, students' self-reported transcript accuracy ratings were analyzed descriptively. After reviewing the system-generated transcript, students rated how accurately the transcript reflected what they intended to say in their oral response using a five-point scale. These ratings provided a complementary perspective on transcript adequacy from the student's point of view. They were not used as an independent validation measure or for inferential analysis.

Human transcript preparation and verification required approximately 6–10 minutes per four-minute oral response. This process involved listening to the student recording, preparing a human-verified reference transcript, and checking the transcript for accuracy. The required time varied depending on audio quality, speaking clarity, pauses, and the extent of correction needed.

Because this study is exploratory and no established threshold exists for determining transcript adequacy in WTL-oriented think-aloud responses, the results were interpreted using convergent evidence rather than a single pass/fail cutoff. WER, CER, semantic similarity, error-type distribution, and student ratings were considered together to determine whether the generated transcripts preserved students' spoken responses with sufficient fidelity to support future learning analysis.

The scope of analysis was intentionally limited. No automated scoring of student learning, classification of cognitive or metacognitive levels, fairness analysis across demographic groups, or feedback generation was conducted in this stage of the study. These analyses are reserved for future work after transcription fidelity is established.

8. Evaluation Results

This section reports the results of the transcription fidelity evaluation for students' time-constrained oral responses collected through the proposed pipeline. The results are organized to

correspond directly to the three research questions: transcription accuracy, semantic fidelity, and student-perceived transcript adequacy.

8.1 Dataset Overview

The analysis included oral responses from 19 unique students across two undergraduate engineering courses. Responses were associated with three Writing-to-Learn (WTL) activity types: learning plans, learning evaluations, and learning reflections. Across the transcript pairs analyzed in this study, the reference transcripts included 8,420 words and approximately 72,000 characters. Each oral response was collected within a fixed four-minute response window.

Because students did not all complete the same number of oral-response activities, the number of rated submissions exceeded the number of participating students. Across all participants, 51 oral-response submissions included student self-reported transcript accuracy ratings. Therefore, student-rating results are reported at the submission level rather than the individual-student level. Table 1 summarizes the main dataset characteristics and transcription-evaluation results.

Table 1. Summary of transcription evaluation results

Measure	Result
Unique students	19
Rated oral-response submissions	51
Reference transcript words	8,420
Reference transcript characters	Approximately 72,000
Word Error Rate (WER)	7.85%
Character Error Rate (CER)	3.78%
Average semantic similarity	0.93
Total word-level errors	661
Substitutions	221
Deletions	138
Insertions	302
Ratings of 4 or 5 stars	82.4%
Mean student rating	4.33 / 5

8.2 RQ1: Transcription Accuracy

At the word level, the automatic transcription achieved a Word Error Rate (WER) of 7.85%. This value was computed from 221 substitutions, 138 deletions, and 302 insertions, resulting in 661

total word-level errors relative to the human-prepared reference transcripts. The WER result indicates that most words in students' oral responses were transcribed correctly.

At the character level, the Character Error Rate (CER) was 3.78%. Character-level errors were primarily associated with minor spelling differences, partial words, disfluencies, or small variations in phrasing rather than broad transcription failure. Together, the WER and CER results indicate that the automatic transcription system produced a generally accurate textual representation of students' oral responses in this dataset.

8.3 RQ2: Semantic Fidelity and Error-Type Distribution

Semantic similarity analysis examined whether transcription errors affected preservation of students' intended meaning. Using embedding representations and cosine similarity, the average semantic similarity between the human-prepared reference transcripts and the system-generated transcripts was 0.93. This result suggests strong alignment in meaning between the two transcript versions, even when surface-level wording differences were present.

Error-type analysis further clarified the nature of transcription differences. Insertions accounted for the largest proportion of word-level errors, representing 45.7% of all word-level errors. Substitutions accounted for 33.4%, and deletions accounted for 20.9%. Insertions often corresponded to filler words, repetitions, or additional minor words. Substitutions generally reflected alternative word choices or small recognition differences. Deletions were less frequent, suggesting that the system was less likely to omit large amounts of student content. Overall, the error distribution indicates that transcription errors more often involved surface-level wording differences than loss of substantive reasoning.

8.4 RQ3: Student Self-Reported Transcript Adequacy

Students' perceptions of transcript accuracy were collected immediately after each oral response, after students reviewed the system-generated transcript. Ratings were collected using a five-point star scale, with higher ratings indicating stronger agreement that the transcript reflected what the student intended to say.

Of the 51 rated submissions, 30 received five-star ratings, 12 received four-star ratings, 5 received three-star ratings, and 4 received two-star ratings. No submissions received a one-star rating. Overall, 42 of the 51 rated submissions, or 82.4%, received ratings of four or five stars. The mean rating was 4.33.

These ratings were analyzed descriptively and were not used as an independent validation measure. Instead, they provide a complementary student perspective on transcript adequacy. The rating results indicate that students generally perceived the system-generated transcripts as faithful representations of their spoken responses.

8.5 Summary of Findings

Across word-level accuracy, character-level accuracy, semantic similarity, error-type distribution, and student self-reported ratings, the results provide convergent evidence that the automatic transcription system preserved students' time-constrained oral responses with sufficient fidelity for preliminary WTL-related analysis. These findings directly address the study's research questions by showing that the transcripts captured most spoken content accurately, preserved intended meaning at a high level, and were generally perceived by students as adequate representations of their oral responses.

9. Discussion

The purpose of this study was to determine whether students' time-windowed oral responses could be automatically transcribed with sufficient fidelity to support future analysis within a Writing-to-Learn (WTL) framework. The results provide preliminary evidence that the pipeline can capture students' spoken responses with reasonable accuracy. For RQ1, the WER and CER results indicate that most spoken content was preserved at the word and character levels. For RQ2, the high semantic similarity score and error-type distribution suggest that transcription errors generally affected surface wording more than substantive meaning. For RQ3, student ratings indicate that students generally perceived the generated transcripts as adequate representations of what they intended to say.

These findings are important because generative AI creates a validity challenge for written WTL artifacts. Time-windowed oral responses provide a complementary modality that encourages immediate articulation of thinking and reduces opportunities for AI-generated substitution. At the same time, automatic transcription allows oral responses to be analyzed using text-based methods already familiar in WTL research.

Several practical issues should be considered. Technical terminology may be misrecognized in engineering contexts, and pauses, soft speech, microphone quality, and background noise may affect transcription quality. These issues did not appear to substantially distort semantic interpretation in the present dataset, but they should be addressed through clearer student instructions and improved recording procedures in future deployments.

The present study does not validate automated grading, feedback generation, or learning-outcome assessment. Rather, it establishes transcription fidelity as a necessary first step. If validated with larger and more diverse datasets, the pipeline can support future integration with AI-powered, rubric-based assessment tools that evaluate students' cognitive, metacognitive, and motivational responses, identify learning strengths and weaknesses, and provide adaptive feedback.

10. Limitations and Future Work

This study has several limitations. First, the evaluation is based on a relatively small sample from two undergraduate engineering courses. Larger and more diverse samples are needed to examine variability related to course context, prompt type, speaking style, accent, and dialect.

Second, the study focuses only on transcription accuracy, semantic fidelity, error-type distribution, and student-perceived transcript adequacy. It does not evaluate automated grading, learning outcomes, or adaptive feedback. Therefore, the findings should be interpreted as evidence of transcription feasibility rather than evidence of complete assessment validity.

Third, transcription performance may vary under different recording conditions, devices, microphones, internet connections, ASR models, and acoustic environments. Technical terminology, pauses, soft speech, and background noise may also influence transcript quality. Future work should provide clearer student recording instructions and examine how these factors affect transcription accuracy.

Fourth, human-prepared reference transcripts were used for comparison, but formal inter-rater reliability was not computed in this preliminary study. Future studies should include multiple reviewers and agreement measures.

Future work will extend this pipeline in three directions. First, larger HBCU student samples will be used to examine fairness and dialect sensitivity. Second, validated transcripts will be integrated into an AI-powered, rubric-based assessment pipeline for evaluating cognitive, metacognitive, and motivational responses and generating formative feedback. Third, local or institutionally managed deployment will be explored to better protect student audio and transcript data.

References

[1] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, S. Dementieva, F. Fischer, U. Gasser, and G. Kasneci, “ChatGPT for good? On opportunities and challenges of large language models for education,” Learning and Individual Differences, vol. 103, p. 102274, 2023. https://doi.org/10.1016/j.lindif.2023.102274

[2] J. Emig, “Writing as a mode of learning,” College Composition and Communication, vol. 28, no. 2, pp. 122–128, 1977. https://www.jstor.org/stable/356095

[3] S. A. Finkenstaedt-Quinn, M. Petterson, A. Gere, and G. Shultz, “Praxis of writing-to-learn: A model for the design and propagation of writing-to-learn in STEM,” Journal of Chemical Education, vol. 98, no. 5, pp. 1548–1555, 2021. https://doi.org/10.1021/acs.jchemed.0c01482

[4] L. Marks, H. Lu, T. Chambers, S. A. Finkenstaedt-Quinn, and R. S. Goldman, “Writing-to-learn in introductory materials science and engineering,” MRS Communications, vol. 12, pp. 1–11, 2022. https://doi.org/10.1557/s43579-021-00114-z

[5] M. T. H. Chi and R. Wylie, “The ICAP framework: Linking cognitive engagement to active learning outcomes,” Educational Psychologist, vol. 49, no. 4, pp. 219–243, 2014. https://doi.org/10.1080/00461520.2014.965823

[6] K. A. Ericsson and H. A. Simon, Protocol Analysis: Verbal Reports as Data, rev. ed. Cambridge, MA: MIT Press, 1993.

[7] J. L. Davis and A. J. Hill, “Analysis of student understanding of force using the Dynamics Concept Inventory, think-alouds and confusion matrices,” in Proceedings of the 2024 ASEE Annual Conference & Exposition, Portland, OR, USA, 2024. https://doi.org/10.18260/1-2--46571

[8] J. L. Davis and A. J. Hill, “Work in progress for two questions: Confusion matrix analysis of student think-alouds during a Dynamics Concept Inventory exam,” in Proceedings of the 2023 ASEE Annual Conference & Exposition, Baltimore, MD, USA, 2023. https://doi.org/10.18260/1-2--44110

[9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in Proceedings of the 40th International Conference on Machine Learning, vol. 202, pp. 28492–28518, 2023. https://proceedings.mlr.press/v202/radford23a.html

[10] D. Cai, B. Naismith, M. Kostromitina, Z. Teng, K. P. Yancey, and G. T. LaFlair, “Developing an automatic pronunciation scorer: Aligning speech evaluation models and applied linguistics constructs,” Language Learning, vol. 75, suppl. 1, pp. 170–203, 2025. https://doi.org/10.1111/lang.70000

[11] T.-H. Lo, F.-A. Chao, T.-I. Wu, Y.-T. Sung, and B. Chen, “An effective automated speaking assessment approach to mitigating data scarcity and imbalanced distribution,” in Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 1352–1362. https://doi.org/10.18653/v1/2024.findings-naacl.86

[12] J. Zhang, C. Borchers, V. Aleven, and R. S. Baker, “Using large language models to detect self-regulated learning in think-aloud protocols,” in Proceedings of the 17th International Conference on Educational Data Mining, Atlanta, GA, USA, 2024. https://educationaldatamining.org/edm2024/proceedings/2024.EDM-long-papers.13/

[13] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, “Racial disparities in automated speech recognition,” Proceedings of the National Academy of Sciences, vol. 117, no. 14, pp. 7684–7689, 2020. https://doi.org/10.1073/pnas.1915768117

[14] Howard University, “Howard University and Google Research enhance A.I. speech recognition of African American English,” The Dig, 2025. https://thedig.howard.edu/all-stories/howard-university-and-google-research-enhance-ai-speech-recognition-african-american-english

[15] Google Research, “A partnership with Howard University to improve speech technology for Black voices,” Google Research Blog, 2023. https://blog.google/innovation-and-ai/technology/research/project-elevate-black-voices-google-research/

[16] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. draft. Stanford University, 2023. https://web.stanford.edu/~jurafsky/slp3/

[17] OpenAI, “Text embeddings,” OpenAI Documentation, 2024. https://platform.openai.com/docs/guides/embeddings

[18] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, pp. 3982–3992. https://doi.org/10.18653/v1/D19-simi1410