

Human-AI Collaboration for Qualitative Analysis in Participatory Design: Refining the Writing Analytics Tool

Andrew Potter
Arizona State University
Tempe, United States
ahpotter@asu.edu

Zeinab Serhan
Arizona State University
Tempe, United States
zserhan@asu.edu

Nishad A. Patne
Arizona State University
Tempe, United States
npatne@asu.edu

Püren Öncel
University of Minnesota
Minneapolis, United States
oncel001@umn.edu

Ishrat Ahmed
Arizona State University
Tempe, United States
Ishrat.Ahmed.1@asu.edu

Tracy Arner
Arizona State University
Tempe, United States
tarner@asu.edu

Rezwana Islam
Arizona State University
Tempe, United States
rislam11@asu.edu

Rod D. Roscoe
Arizona State University
Tempe, United States
rod.roscoe@asu.edu

Laura A. Allen
University of Minnesota
Minneapolis, United States
lallen@umn.edu

Scott A. Crossley
Vanderbilt University
Nashville, United States
scott.crossley@vanderbilt.edu

Danielle S. McNamara
Arizona State University
Tempe, United States
dsmcnamara@asu.edu

This study introduces a hybrid human-AI workflow to qualitative data analysis within the participatory design of the Writing Analytics Toolkit (WAT), an open-source platform that provides formative feedback on student writing using natural language processing. The toolkit includes a classroom-facing implementation (WAT Classroom; WAT-C), designed to support instruction, and a researcher-facing implementation (WAT Researcher; WAT-R), designed to support analytic and validation workflows. Nine experienced college writing instructors (with 97 cumulative years of teaching) participated in focus group sessions to evaluate an early

prototype of the classroom version of WAT (WAT-C), offering formative input on usability, instructional alignment, and feedback clarity. To analyze the resulting qualitative data, we employed a novel AI-augmented analytic process: GPT-4o, integrated within a secure, retrieval-augmented system, to generate inductive codes and preliminary themes from transcripts. These AI-generated outputs were iteratively reviewed, critiqued, refined, and synthesized by researchers, supporting both analytical scalability and interpretive rigor. This human-AI partnership enabled efficient thematic exploration while preserving methodological transparency and researcher judgment. Findings from both qualitative and complementary survey data identified four key design priorities: (1) clearer, more concise feedback, (2) increased instructor customization, (3) reduced administrative burden, and (4) a simplified user interface. These insights directly informed subsequent revisions to WAT-C, including a redesigned feedback interface, customizable metric targets, learning management system integration, and a more intuitive layout. This work illustrates how large language models (LLMs) can support inductive qualitative analysis within participatory design workflows. Moreover, results demonstrate how this workflow can inform iterative educational technology development. Implications include the need to ensure ethical oversight, researcher-led interpretation, and alignment with instructional priorities when incorporating AI into the design of educational technologies.

Keywords: writing analytics, participatory design, generative AI, qualitative data analysis, natural language processing, educational technology

1. INTRODUCTION

Automated writing evaluation (AWE) and writing analytics tools offer scalable support for both teaching and learning of writing by providing students with immediate, formative feedback to guide revision (Potter et al., 2025; Strobl et al., 2019; Wilson & MacArthur, 2024). However, developing effective AWE tools requires user-centered design to ensure alignment with instructional needs (Tuhkala, 2021). These design processes are labor-intensive and difficult to scale (Abrams et al., 2004), but recent advances in large language models (LLMs) present new opportunities to reduce the time and effort involved in analyzing qualitative user feedback. We posit that mindful integration of LLMs may potentially streamline critical participatory design processes in AWE tools (McNamara & Potter, 2024).

This paper focuses on how LLMs can be used to support qualitative data analysis during participatory design (Muller & Kuhn, 1993; Ten Holter, 2022; Wacnik et al., 2025). Our study is situated within the broader development of the Writing Analytics Toolkit (WAT), an open, research-driven platform that provides writing analytics for students, teachers, and researchers, designed in collaboration with writing instructors and researchers. WAT comprises two complementary implementations: WAT Classroom (WAT-C), which supports instructional use by students and teachers, and WAT Researcher (WAT-R), which supports post hoc linguistic analysis of writing corpora. In this paper, we focus on WAT Classroom (WAT-C), the classroom-facing implementation of the toolkit.

The paper makes two interrelated contributions. The primary contribution is methodological; it introduces a hybrid AI-human workflow for inductive thematic analysis that leverages LLMs to assist in qualitative coding and theme generation. The second, applied contribution demonstrates how this workflow can be used to inform the iterative redesign of WAT-C through analysis of instructor feedback. To contextualize this work, the current phase builds on a prior participatory design study that established WAT-C's core interface, functionality, and descriptive writing metrics (Li et al., 2022). In this phase, instructors participated in two focus group

sessions to evaluate and refine these metrics, interpret their instructional meaning, and improve how feedback is presented and integrated into instruction.

To analyze instructor insights, we employed a mixed-methods design that combined LLM-assisted qualitative analysis with human coding and review. Focus group transcripts were uploaded to a secure analysis platform, where the ChatGPT-4o model (OpenAI, 2024) was guided through a structured prompt framework to identify preliminary themes. These outputs were then verified, refined, and supplemented by human researchers to enhance the trustworthiness and interpretive accuracy of the findings. This hybrid AI-human workflow enabled a systematic and efficient analysis of instructor feedback, informing the redesign of WAT-C to better meet instructional needs. In doing so, the study demonstrates how LLM-supported workflows can enhance participatory design processes while maintaining methodological rigor and improving the efficiency of qualitative analysis. While much of the AI research in education has centered on detection (e.g., identifying student characteristics, predicting outcomes, or distinguishing human- from AI-generated text), this study illustrates a different use of AI that supports qualitative interpretation and collaborative design. By employing LLMs to assist researchers and educators in making sense of complex feedback data, the work extends AI's role in education beyond detection toward participatory, interpretive, and human-centered applications.

2. RELATED WORK

AWE and writing analytics tools are systems designed to provide students and teachers with immediate, formative feedback on their writing, offering scalable opportunities for revision and practice (McNamara & Kendeou, 2022; Potter et al., 2025; Shermis & Wilson, 2024). A number of AWE systems have been developed to support writing improvement, offering students formative feedback on structure, cohesion, and other analytical traits (Butterfuss et al., 2022; Correnti et al., 2022; Link et al., 2014; Knight et al., 2017, 2020). These tools have shown promise for improving writing outcomes and writing self-efficacy, particularly when paired with classroom instruction (Dikli & Bleyle, 2014; Fleckenstein et al., 2023; Liu et al., 2017; Palermo & Thomson, 2018; Wilson & Roscoe, 2020).

Despite their growing presence in writing instruction, AWE systems vary considerably in the extent to which they are adopted and sustained in educational contexts (Potter & Wilson, 2022). Prior work indicates that students' perceptions of AWE significantly predict their future intentions to continue using or recommending the tool, even if such perceptions have minimal impact on their immediate revision behaviors (Roscoe et al., 2017). Students have also expressed dissatisfaction with the personalization, fairness, or depth of the feedback provided by AWE systems, which can affect their engagement and, in turn, the tool's instructional effectiveness (Wilson, Delgado, et al., 2024). Similarly, teacher perceptions and contextual factors, such as curriculum alignment, time constraints, and availability of training, can serve as barriers to implementation fidelity and instructional effectiveness (Wilson, Zhang, et al., 2024). Writing instructors have raised concerns about the pedagogical misalignment of systems that rely on predictive scoring instead of providing descriptive, actionable feedback (Li et al., 2022). In addition to these instructional and implementation concerns, AWE systems are susceptible to design biases that can reinforce inequities and standard language ideologies (Goldshtein, Alhashim, & Roscoe, 2024; Goldshtein, Ocumpaugh, et al., 2024). The proprietary nature of most AWE systems also limits transparency, adaptability, and access for educators and researchers (McNamara & Potter, 2024; Strobl et al., 2019).

A range of studies have applied participatory design methods, a form of user-centered design that emphasizes collaboration with end users throughout the development process (Muller & Kuhn, 1993). For example, such methods have been used to identify meaningful writing process indicators for feedback with different stakeholders (Conijn et al., 2022), integrate teacher-centered design into intelligent tutoring systems (Stone et al., 2018), and co-develop feedback platforms that combine automated and peer review to support English language learners (Liaquat et al., 2021). Collectively, these studies underscore the importance of developing writing analytics tools in close collaboration with end users to ensure the tools are usable, interpretable, and aligned with authentic educational needs.

Although user-centered design is critical for developing effective and pedagogically aligned writing technologies, it is also time-consuming and labor-intensive, particularly when it comes to analyzing qualitative user data (Abrams et al., 2004; Roscoe et al., 2018). Focus group transcripts, open-ended surveys, and interviews provide rich insight into user needs and instructional contexts, but interpreting these data requires significant human effort and expertise. In addition to the volume of qualitative data, the process of codesign introduces additional complexities, as designers must navigate the creative collaboration of diverse stakeholders while balancing the tension between understanding existing practices and envisioning future ones (Steen, 2011). As a potential approach to these challenges, LLMs may help reduce the labor demands of user-centered development, and in this study, we examine their use specifically for qualitative data analysis, a promising application alongside others such as crowdsourced feedback and iterative prototyping (McNamara & Potter, 2024).

GenAI tools, particularly LLMs, have recently been evaluated as a tool to support qualitative data analysis (Khalid & Witmer, 2025; Zambrano et al., 2023). LLMs are neural networks trained on massive text corpora for language processing and generation, and they form the foundation of many generative artificial intelligence (GenAI) tools (see Minaee et al., 2024, for a review). Initial applications of LLMs for qualitative data analysis have primarily focused on deductive coding, in which the model applies predefined codebooks to structured textual data (Chew et al., 2023; Kirsten et al., 2024; Xiao et al., 2023). However, a growing body of work has begun to explore their use in inductive coding, where codes are generated directly from the data (Bijker et al., 2024; Chen et al., 2024; Khalid & Witmer, 2025; Katz et al., 2024; Prescott et al., 2024; Theelen et al., 2024; Turobov et al., 2024; Zhang, Wu, Xie, Kim, & Carroll, 2023; Zhang, Wu, Xie, Lyu, et al., 2023; Zhang et al. 2024; Zhao et al., 2024). Inductive analysis differs from deductive approaches in that researchers generate codes and themes directly from the data, allowing patterns, meanings, and concepts to emerge organically through iterative interpretation (see Saldaña, 2014; Thomas, 2006). As a result, inductive analysis is typically more complex and interpretive than deductive coding. Interestingly, some studies suggest that LLMs may appear more effective in inductive tasks, not because they outperform human coders, but because the lack of a predefined codebook makes it harder to identify clear errors, particularly when models generate plausible yet superficial themes (Bijker et al., 2024; Chen et al., 2024).

Nevertheless, studies have shown that when guided by structured prompts, LLMs can support inductive coding processes that yield themes aligned with those generated by human researchers, and that they can also reduce the time required for analysis (Bijker et al., 2024; Theelen et al., 2024; Prescott et al., 2024; Zhao et al., 2024). Their effectiveness, however, depends on several factors, including the choice of tools, prompt design, and the degree of human involvement throughout the workflow. Prior studies indicate better performance in inductive coding when LLMs are compared to machine learning and natural language processing approaches (Chen et al., 2024), and that newer models are more reliable with human coders than earlier models (Kirsten et al., 2024).

Beyond model selection, prompt engineering has emerged as a critical factor in performance. Across multiple studies, prompts were found to perform best when they met three key criteria: (1) specifying the model's role (e.g., "You are a qualitative research expert"), (2) defining the expected structure and format of the input and output (e.g., datasets, themes, justification), and (3) incorporating reasoning-based strategies such as rationale generation, exemplars, or chain-of-thought prompting to mirror human interpretive processes (Wei et al., 2022; Zhang, Wu, Xie, Kim, & Carroll, 2023; Zhang, Wu, Xie, Lyu, et al., 2023; Zhang et al. 2024). For instance, researchers have improved model outputs in qualitative data analysis by presenting multiple rounds of example annotations, guiding the model to revise or refine its coding decisions in response to human feedback, or prompting it to generate justifications alongside each label (Chew et al., 2023; Dunivin, 2024; Zhao et al., 2024). These prompt engineering practices enhance transparency, interpretability, and researcher control over analytic content and meaning-making, making LLM outputs easier to audit, refine, and integrate into the development of qualitative findings. Informed by these insights, our study adopts a structured prompt framework that implements several of these strategies (see Method).

In addition to prompt design, recent studies have also examined methodological enhancements to improve LLM performance through integrated, human-in-the-loop workflows. For instance, during codebook co-development, models are prompted iteratively with human-reviewed examples to refine category definitions and clarify boundary cases (Katz et al., 2024; Khalid & Witmer, 2025; Zhao et al., 2024). Notably, the strongest performance in terms of model accuracy and reliability in inductive analyses has been observed in hybrid workflows where human researchers guide early iterations, review LLM-generated codes, and incorporate contextual knowledge to calibrate and revise model outputs (Chen et al., 2024; Theelen et al., 2024; Zhang, Wu, Xie, Lyu, et al., 2023). These findings underscore that LLMs are most effective when embedded within structured, iterative analytic processes that leverage human expertise at key points in the workflow. Even when assisted by AI, it is the researchers who ultimately shape, interpret, and take responsibility for the analysis (Hayes, 2025).

Despite these promising developments, recent findings have also underscored important limitations in LLM-assisted inductive coding. Although LLMs can generate outputs that resemble and are reliable with researcher-led qualitative data analysis (e.g., crafting themes, summaries, and codes), they do not think, interpret, or reason as human experts do (e.g., Hayes, 2025). Their outputs are shaped by patterns in their training data, which embed the biases of those data (e.g., Warr & Heath, 2025). These challenges are especially salient in qualitative research, where the analysis centers on understanding participants' social contexts and meaning-making processes, and where researcher reflexivity and positionality play a critical role in shaping interpretation (e.g., Aspers & Corte, 2019; Fossey et al., 2002; Corlett & Mavin, 2018). In inductive qualitative data analysis, LLM outputs may overlook subtle meanings, produce overly broad or redundant codes, or generate hallucinated interpretations that misrepresent or fabricate textual data (Chen et al., 2024; Turobov et al., 2024; Zhang et al., 2024). While structured prompts and hybrid workflows can help reduce these risks, the judgment of trained researchers remains critical to guiding analysis, interpreting ambiguity, and ensuring that codes align with the goals of a study. Even when LLMs perform well, they lack expertise and lived experience to interpret complex social meaning (Imundo et al., 2024). Qualitative researchers are not neutral instruments; their identity, perspective, and relationship to the research context shape the insights that emerge from the data (Bourke, 2014). Consequently, there is growing consensus that LLMs should not be viewed as replacements for human qualitative researchers, but rather as tools that can augment human insight and support scalable, rigorous analysis under researcher supervision (Bijker

et al., 2024; Hayes, 2025; Katz et al., 2024). Thus, maintaining a human-centered orientation is essential for using LLMs in qualitative research.

3. PRESENT STUDY

The present study responds to calls for best practices in LLM-assisted qualitative analysis by pursuing a dual aim: (1) to test a hybrid AI-human workflow for qualitative data analysis and to refine the alpha version of WAT-C through participatory design with writing instructors, and (2) to examine how a hybrid AI-human workflow can support, rather than replace, qualitative data analysis within the design process. Using a mixed methods approach, qualitative analysis of focus group transcripts was integrated with quantitative survey results to synthesize writing instructors' feedback and design recommendations for the next iteration of WAT-C. In this context, the study explores how LLMs can support qualitative analysis without replacing the interpretive expertise of human researchers. Specifically, the analysis examined how writing instructors interpreted the usability and instructional relevance of WAT-C metrics and affordances, and identified their design contributions and priorities for improving the tool. The study also evaluates how LLMs can be integrated into a participatory design workflow to support, rather than replace, human-led qualitative analysis.

4. METHOD

4.1. RESEARCH DESIGN

This study employed a participatory design research framework. Participatory design emphasizes the active involvement of users in the development and evaluation of tools and systems (Muller & Kuhn &, 1993; Wacnik et al., 2025). This approach has been leveraged in educational contexts to engage teachers in developing educational technologies, including writing analytics tools (Conijn et al., 2022; Cumbo & Selwyn, 2022; Tuhkala, 2021). As Wacnik et al. (2025) notes in a recent systematic review on participatory design, the method encompasses a broad range of user-centered design practices and is not defined by a singular methodology. Instead, participatory design varies across projects in its timing, techniques, and the degree of stakeholder involvement. Following Wacnik et al.'s (2025) framework, this study engaged in a participatory design process characterized by several key leverage points that influence design equity. First, the process fell between emergent and predetermined: while the interface, writing metrics, and feedback types were initially designed by the research team, instructors provided feedback to revise and reshape how these elements were defined and presented. Second, stakeholder participation was direct, with instructors actively contributing to design discussions (e.g., input on interface changes and metric names and their interpretations). Third, the timing of participation occurred at a mid-stage in development, building on an earlier phase in which different writing teacher participants contributed to the creation of WAT-C's initial interface and metric framework (Li et al., 2022), but preceding classroom implementation. Fourth, instructor engagement consisted of two focus group sessions within a single design phase without re-engagement in later stages, constituting a one-time participatory process rather than iterative involvement across multiple phases (Wacnik et al., 2025). Although participation in this phase was limited due to time constraints, it contributed to a broader iterative process spanning multiple stages of design (Li et al., 2022). Finally, we used multiple participatory techniques, including structured focus groups and prototype evaluation, to capture instructor insights and guide tool refinement.

To analyze data collected through this participatory design process, we employed a convergent mixed-methods design for data analysis, in which qualitative and quantitative data were collected concurrently, analyzed independently, and integrated for interpretation (Creswell & Clark, 2017). This design prioritized qualitative data (QUAL + quan) to center participants' voices, and quantitative and qualitative survey data were used to triangulate qualitative themes (Fielding, 2012; Toledo & Shannon-Baker, 2023).

Following initial open coding and theme development from focus group transcripts, quantitative survey results and open-ended responses were analyzed in parallel to examine patterns in instructors' perceptions of WAT-C's usefulness, usability, interpretability, and behavioral intention. Integration of the two strands occurred during the interpretation and reporting phases, consistent with convergent mixed-methods procedures (Creswell & Clark, 2017). Specifically, qualitative themes were compared with and contextualized by quantitative and open-ended comments, and then combined evidence was synthesized in a joint display to illustrate areas of convergence and divergence across data sources.

4.2. PARTICIPANTS

Writing instructors ($n = 9$) were recruited at a large public university in the southwestern United States to represent the potential user group (see Table 1). Each participant taught entry-level English writing courses. The participants' teaching experience varied, ranging from 1 to over 20 years, with an average of 9.4 years of college-level writing instruction. All participants provided informed consent to participate following IRB-approved guidelines and were awarded a stipend upon completion of their participation.

Table 1: Participant background information

Name	Title	Courses Taught	Years Teaching College	Typical Class Size	Technology Used for Writing Feedback
Ava	Instructor	Introductory composition courses	24	24	None
Anna	Assistant Teaching Professor	Introductory and upper-level composition courses	20	24	None
Marcus	Graduate Teaching Assistant	Introductory composition	13	20	None
Tara	Graduate Teaching Assistant	Introductory composition	12	20	TurnItIn
Seth	Instructor	Introductory and intermediate composition courses	12	24	ChatGPT (student-guided use)
Jason	Instructor	Introductory composition	7	~20	TurnItIn
Gavin	Instructor of record	Composition for multilingual students	6	20–25	Grammarly, ChatGPT, Gemini, WordTune

Au- tumn	Graduate Teaching Assistant	Introductory composi- tion	2	25	None
Kayla	Graduate Teaching Assistant	Introductory composi- tion	1	20–24	ChatGPT (used in all rounds of revi- sion)

Note. Participant names are pseudonyms.

4.3. MATERIALS

4.3.1. Writing Analytics Tool

Participants engaged with the alpha version of WAT-C, a web-based platform intended to provide students with formative feedback and support instructors in managing and evaluating writing assignments. Key features were evaluated in terms of usability (how easy the system is to navigate), usefulness (how helpful the features are for instruction), interpretability (how clearly the feedback is presented), and pedagogical alignment (how well the system supports instructors' instructional goals). Developed through prior participatory design research (see Li et al., 2022), the version of WAT-C used in this study was revised to shift emphasis away from predictive scoring and toward descriptive feedback aligned with classroom practice. Instructors in an earlier design cycle expressed a preference for maintaining their role as the primary evaluator of writing quality and favored analytics that could support revision and pedagogical decision-making.

Recent work in AWE and writing analytics highlights the need for AI-based writing tools to move beyond detection and scoring toward feedback that is pedagogically meaningful and adaptable to classroom contexts. For example, *eRevise* (Correnti et al., 2024; Wang et al., 2020) is an AWE system that exemplifies this shift through the inclusion of revision instruction, making feedback more formative and student-centered. Other studies stress aligning automated feedback with learning design and disciplinary writing practices through instructor co-design (e.g., Knight et al., 2020). Research on implementation further shows that AWE effectiveness depends on teacher and student engagement as much as technical performance (Huang et al., 2025). Building on this pedagogical direction, WAT-C prioritizes teacher agency, customization, and descriptive feedback to enhance interpretability and instructional integration.

As such, the system generates analytics in three categories for persuasive and source-based essays: *basic analytics* (e.g., word count, sentence length), a *holistic overall score* (1–6 scale), and *advanced analytics*. WAT-C also produces a *holistic quality score* (1–4 scale) for summaries. The advanced analytics were developed using natural language processing (NLP) tools and principal component analysis on a large dataset of student essays. These metrics reflect dimensions of writing such as lexical sophistication, sentence cohesion, and idea development. Table 2 presents the set of advanced metrics included in the alpha version of WAT-C, along with brief descriptions used in the interface.

Table 2: WAT-C advanced metrics for persuasive and source-based essays

Advanced Metric Name	Brief Definition
Sophisticated Wording ^{1,2}	The extent to which more advanced, or less commonly found words are used.

Development of Ideas ^{1,2}	The extent to which ideas in an essay are developed and elaborated throughout an essay.
Word Variety ¹	The extent to which the text contains a variety of different words.
Information Density ¹	The extent to which an essay contains sentences with dense and precise information, which emphasize complex noun phrases and sophisticated and specific function words.
Word Concreteness ^{1,2}	The degree to which the text contains words that are tangible and refer to concepts that can be experienced by the senses, on average.
Conventional Language ^{1,2}	The extent to which the text contains words and phrases using grammar constructions mirrored in published texts.
Sentence Cohesion ^{1,2}	The degree to which there is overlap between words across sentences.
Academic Focus ¹	The extent to which the text addresses a prompt and uses academic wording.
Conversational Writing Style ^{1,2}	The extent to which the text is written in an informal conversational style, emphasizing personal opinions, anecdotes, and complex sentences.
Language Variety ²	The extent to which the text contains a variety of different words and sentence structures.
Academic Writing ²	The extent to which sophisticated vocabulary and complex sentence structures are used.
Source Integration ²	The extent to which the text contains overlap of words from the source text(s).
Varied Sentence Structure ²	The degree to which an essay includes a diversity of different sentence structures.

Note. ¹ Indicates metric is used for persuasive essays, and ² Indicates metric is used for source-based essays

Each advanced metric is presented to students on a five-level continuum. Students can explore each metric through clickable features, including pop-up explanations, revision suggestions, and example use cases (see Figure 1). These features are designed to help students interpret feedback concerning their communicative goals and revise accordingly. Annotation A highlights the left end of the continuum, which students can click to view guidance on how to revise when their writing falls at a lower level of the metric (in this case, more consistent wording). Annotation B shows the right end of the continuum, which provides guidance for revising toward a higher level (in this case, more diverse). Annotation C displays the "Click here to learn" link, which opens a pop-up window (see Figure 2) that explains when different levels of the metric may be most appropriate, including examples of writing tasks that benefit from consistent language (e.g., technical reports) versus those that benefit from more diverse or expressive language (e.g., personal narratives). This pop-up also offers revision suggestions tailored to communicative purpose and audience. Finally, Annotation D identifies the information icon, which students can hover over to access a brief description of the selected metric level.

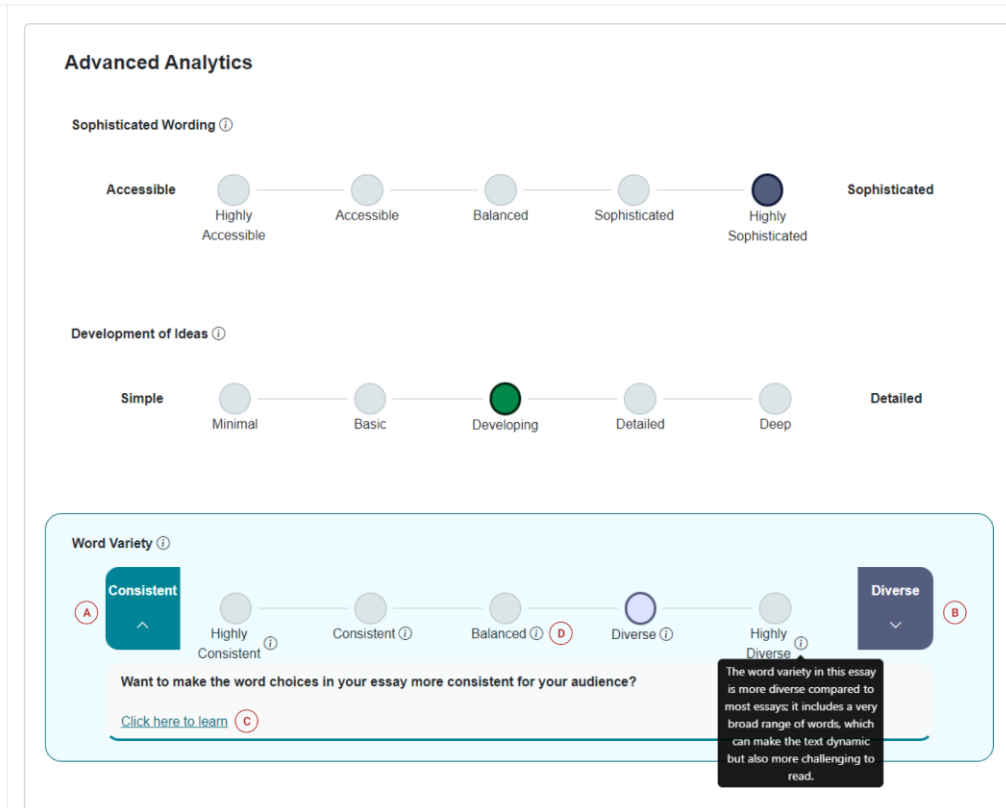


Figure 1: WAT-C Alpha advanced metric and feedback display in student interface.

Each metric also includes written feedback that appears when students click on the “Click here to learn” (Annotation C in Figure 1) assigned to their essay (see Figure 2). Effective feedback is most impactful when it includes detailed information about both the task and the strategies students can use to improve their work (Hattie & Timperley, 2007; Wisniewski et al., 2020) and when it supports students in evaluating their writing to inform meaningful revisions (MacArthur, 2016). To operationalize these principles in a format that is interpretable and actionable for students, we presented writing metrics and written feedback in a Why–When–How feedback framework. The Why section explains the rationale for the feedback and why it would make sense for the writer to revise their essay to increase or decrease alignment with the metric (e.g., making word choice more diverse can make the text more dynamic but challenging to comprehend; see Figure 2). The When section describes situations or genres where the writing behavior is most appropriate, helping students understand how context influences effective writing. The How section provides concrete revision strategies students can apply to improve their work. For example, students viewing feedback for “Word Variety” might read why lexical diversity enhances reader engagement, when to prioritize consistency over variation, and how to revise by replacing repeated words with synonyms or expanding on ideas. These structured feedback elements are designed to make advanced analytics interpretable, actionable, and instructionally relevant.

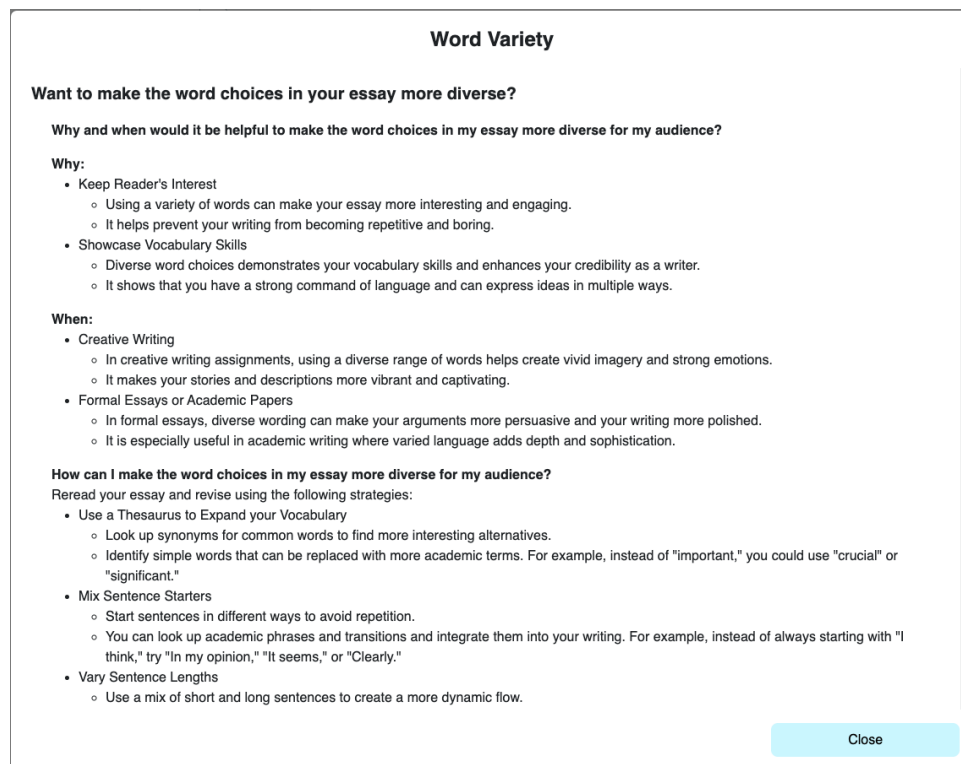


Figure 2: Example written feedback comments associated with the Word Variety metric.

The teacher interface allowed participants to assign writing tasks, select which metrics to include in feedback, upload rubrics or mentor texts, and track student submissions. The student interface included a dashboard for managing assignments, reviewing feedback, and resubmitting revised drafts. During the focus group sessions, participants interacted with these core features of the system and provided detailed feedback on its usability, pedagogical alignment, and feedback design.

4.3.2. Focus Group Protocol

The focus group discussions were guided by a semi-structured protocol developed by the research team. This phase followed the initial design cycle reported in Li et al. (2022), where writing instructors co-developed WAT-C's core interface and decided to use descriptive rather than predictive metrics. This study represents a second, refinement-oriented phase of participatory design research. According to Wacnik et al.'s (2025) framework, the process incorporated direct, mid-stage, and feedback-driven participation that balanced predetermined tool features (e.g., the initial metrics created by research team members with expertise in applied linguistics and NLP) with emergent instructor input on how those features should be presented, interpreted, and applied within WAT-C.

The protocol included structured activities and open-ended prompts designed to elicit feedback and design-oriented reflection. Two focus group sessions with instructor participants ($n = 9$) were conducted over consecutive days via Zoom, each lasting approximately two hours and including two breakout rooms for small-group discussion. In the first session, instructors completed a "scavenger hunt" activity to explore key interface components, including using the teacher interface to locate a student essay, review feedback for a specific metric, and submit a feedback comment to a hypothetical student. These tasks were followed by facilitated

discussions in breakout groups, where instructors shared usability observations and raised interface design concerns. The session continued with a metric-by-metric walkthrough of first and second student drafts using example persuasive and source-based essays that were written by students. For each of WAT-C's 13 advanced metrics, instructors reviewed the metric definition, the level assigned to the essay, and the associated feedback for the example essays. They then evaluated the accuracy and pedagogical value of the feedback and discussed how each metric could be revised to better reflect their instructional goals. These discussions often involved critiques of clarity, tone, contextual relevance, and suggestions for alternate phrasing or scaffolding, which were documented and used to inform future design decisions.

The second session continued this design-feedback cycle, focusing on the remaining metrics and culminating in a collaborative discussion of which aspects of WAT-C should be kept, revised, or removed. Instructors also responded to focus group prompts designed to elicit concrete implementation recommendations, including how the tool could be used in classroom settings and where additional support might be needed. Throughout both sessions, instructors were asked to evaluate existing features and explicitly encouraged to identify pain points, propose modifications, and reflect on how the tool could better align with their pedagogical practices.

4.3.3. Survey Instrument

The post-session survey, completed by the same nine instructors who participated in the focus groups, was developed to evaluate participants' experiences with WAT-C. Survey constructs were grounded in the Technology Acceptance Model (TAM; Davis, 1989; Venkatesh & Davis, 2000), which identifies perceived usefulness, perceived ease of use, and behavioral intention to use as central predictors of technology adoption. The survey included items addressing additional implementation-related concerns, such as perceived feasibility and instructional alignment. The instrument included both closed- and open-ended items and was structured into five domains: (a) perceived usefulness of advanced metrics and predictive scores, (b) perceived usefulness and structure of written feedback comments associated with advanced metrics, (c) interpretability and clarity of visual feedback displays, (d) usefulness of instructional customization and feedback features, and (e) behavioral intention to adopt and perceived feasibility for classroom integration.

All closed-ended items were rated on a 6-point Likert-type scale. Items assessing perceived usefulness, clarity, and agreement ranged from 1 (Not useful at all / Strongly disagree) to 6 (Extremely useful / Strongly agree). Items assessing behavioral intention and perceived instructional impact ranged from 1 (Very unlikely) to 6 (Very likely). One item assessing anticipated time burden was reverse-coded, with responses ranging from 1 (Will take a great deal more time) to 6 (Will save a great deal of time).

In the first domain, participants rated the perceived usefulness of WAT-C's advanced linguistic metrics and predictive quality scores for supporting student writing. Metrics included items such as Academic Focus, Word Variety, Word Concreteness, and Overall Quality Score. Participants rated each item using a 6-point Likert scale from 1 (Not useful at all) to 6 (Extremely useful).

The second domain focused on WAT-C's written feedback comments associated with the advanced metrics. Participants rated the usefulness and clarity of the Why–When–How feedback structure, including the extent to which each component helped students understand, apply, and revise their writing. Items in this section were rated on the same 1 to 6 usefulness/agreement scale, with an additional reverse-coded item evaluating whether added metric definitions would be beneficial or overwhelming.

The third domain assessed the interpretability and visual clarity of WAT-C's on-screen feedback displays. Items addressed the perceived usefulness and accuracy of visual elements, such as the color-coded spectrum, "i" icon definitions, and the "Click Here to Learn" link. Participants responded using a 1 (Strongly disagree) to 6 (Strongly agree) scale.

In the fourth domain, participants evaluated the perceived usefulness of instructional features that supported customization and feedback delivery. These included the ability to upload personalized rubrics, assign multiple drafts, control student-facing analytics, and communicate through comment tools. All items were rated on a 1 to 6 scale, with higher values indicating greater perceived usefulness for classroom instruction.

The fifth and final domain addressed participants' behavioral intention to use WAT-C and perceptions of feasibility for classroom integration. Items in this section asked instructors to rate their likelihood of adopting the tool, perceived usefulness for improving student outcomes, and any concerns related to time burden or integration into existing teaching practices. These items were rated on a 1 (Very unlikely) to 6 (Very likely) scale, with the exception of the reverse-coded time item.

In addition to the closed-ended items, participants completed 15 open-ended questions designed to elicit elaborated feedback on WAT-C's metrics, feedback structure, user interface, and instructional alignment. These qualitative responses were used to contextualize and triangulate the closed-ended survey findings (see Section 6.5). The complete survey instrument is available in the project's Open Science Framework (OSF) repository: https://osf.io/8j9y6/?view_only=3a21ed900475418a8c258ea595cbe53b.

4.3.4. AI Analysis Platform (ASU CreateAI Chatbot)

To integrate LLMs into qualitative data analysis, we used GPT-4o (OpenAI, 2024) to support initial theme identification. While LLMs can accelerate early coding processes, they also introduce concerns related to hallucinated output (Xu et al., 2024) and data privacy (Golda et al., 2024). To address these risks, we conducted all analyses using Arizona State University's CreateAI Chat platform, a secure, closed-system environment that complies with institutional data protections and supports retrieval-augmented generation (RAG) (Ahmed et al., 2025; Arizona State University, 2024). The version of CreateAI Chat used for this study did not permit the creation of a system-level prompt and operated with a fixed temperature setting of 0.5, which could not be modified by users.

The CreateAI platform's RAG architecture allowed the model to retrieve relevant text segments from preapproved sources, specifically, the focus group transcripts and Braun and Clarke's (2006) article on thematic analysis, before generating responses. This retrieval step constrained model output to contextually grounded information, thereby reducing (though not eliminating) the potential for hallucinated or unsupported content (Bécharde & Ayala, 2024; Golda et al., 2024; Xu et al., 2024). To guide the model's approach to qualitative coding, we uploaded Braun and Clarke's (2006) thematic analysis article, which had been used in prior research for this purpose (e.g., Yang et al., 2024). Both the segmented transcripts and the Braun and Clarke (2006) article were included as retrieval sources to support consistent, context-aware responses.

4.4. PROCEDURES

The study consisted of two focus group sessions conducted over two consecutive days via Zoom led by members of the research team. Each session lasted two hours and included two breakout

rooms to facilitate smaller group discussions. Data was collected through audio recordings and Zoom chat transcripts from each room.

In the first session, participants were introduced to WAT-C through a guided walkthrough of its primary features, including the login process, classroom dashboard, and writing task management interface. Participants then completed a scaffolded scavenger hunt to explore the tool's functionality. Specifically, they were asked to (1) locate the feedback page for a specific student essay, (2) find and review the *Language Variety* metric for that essay, and (3) examine the remaining advanced metrics before submitting a feedback comment to the student. After completing the tasks, participants discussed their user experience navigating the interface and interpreting the feedback in small groups. Following the scavenger hunt, participants were guided through an analysis of WAT-C's 13 advanced metrics using example student essays. For each metric, participants reviewed the metric definition, the metric level assigned to the essay (e.g., if the sophisticated wording in an essay was rated as balanced, accessible, highly sophisticated), and the associated feedback comments (i.e., the written comments associated with increasing or decreasing the level of each metric). Instructors evaluated the clarity and perceived accuracy of the metric definitions and values and discussed whether the feedback provided would be interpretable and actionable for students. The sample texts included first drafts written by students and second drafts revised by researchers. Participants compared both drafts and considered how metric feedback aligned with actual revisions. These activities spanned multiple genres, including persuasive essays, source-based essays, and summaries.

In the second session, participants continued evaluating the remaining metrics and discussed their accuracy, clarity, and pedagogical relevance. The session concluded with a discussion focused on which aspects of WAT-C should be retained, revised, or removed, as well as the feasibility of using WAT-C in their instruction.

Throughout both sessions, instructors were invited to reflect on how WAT-C's feedback aligned with their instructional goals and to propose revisions to the feedback language, metric definitions, and interface features. Although these discussions did not include real-time design iteration, instructors contributed direct feedback and solution-oriented recommendations, consistent with Wacnik et al.'s (2025) framework, which recognizes focus groups as a participatory technique that may be used to evaluate prototypes and may be implemented at various stages of design. Instructors participated in two connected sessions within a single phase of the design process. While their engagement was limited to this phase, it was direct and functioned within a partially predetermined design context, allowing instructors to influence the interpretation, presentation, and instructional alignment of existing WAT-C features. Although participants did not engage in direct iteration of the metrics or interface during this phase, the study builds on a prior participatory cycle (Li et al., 2022) and contributed to future design decisions through instructor evaluation and targeted discussions focused on improving the tool's alignment with classroom needs.

4.5. DATA SOURCES AND PREPARATION

The primary data source for this study was qualitative, derived from two recorded Zoom focus group sessions. Audio from the main and breakout room discussions was auto-transcribed by Zoom. To ensure that all contributions were preserved in context, the time-stamped chat logs were systematically integrated into the auto-generated transcripts at the point in the conversation when they occurred. Trained research assistants reviewed the transcripts for accuracy, corrected errors, and removed identifying information. Pseudonyms were assigned to all participants.

LLM token limits can lead to models that ignore parts of the datasets and thus produce incomplete responses (Hayes, 2025). This issue may be addressed in part by segmenting files and employing RAG (Dunivin, 2024). Thus, to prepare transcripts for analysis, the full dataset was segmented into 17 topic-based files corresponding to specific WAT-C metrics and structured focus group activities (e.g., discussion of summaries, feedback comments, user experience and interface features). Each file ranged from approximately 324 to 5,758 words to preserve semantic coherence and remain within the GPT-4o token limit (approximately 10,000 tokens per prompt). This segmentation strategy allowed for alignment between the structure of the transcript data and the analytic focus of the study, which may improve the accuracy of LLM outputs (Dunivin, 2024).

To complement the focus group data, a post-session survey was administered using Qualtrics immediately following the second session. The survey included quantitative Likert-type items, ranking tasks, and qualitative open-ended responses.

4.6. DATA ANALYSIS

4.6.1. Overview of Human-AI Workflow

Qualitative analysis followed a collaborative workflow combining GPT-4o with expert human review. The goal of this workflow was to increase efficiency and support systematic theme development while maintaining rigor and interpretive accuracy. A step-by-step summary of the workflow for qualitative, quantitative, and mixed-methods integration is presented in Table 3. This process used a general inductive approach (Thomas, 2006), with structured coding and synthesis strategies (e.g., Saldaña, 2014), and integration of qualitative and quantitative data (Creswell & Clark, 2017).

Table 3: Human-AI workflow for data analysis in qualitative and mixed-methods research

Step	Action	Agent / Tool	Purpose	Output
1	Clean and segment transcripts	Human researchers	Prepare manuscripts for analysis to fit transcript segments within LLM token limits	Transcript segments ($\leq 4,000$ tokens)
2	Upload anonymized transcripts	MyAI Builder (secure ASU platform)	Ensure IRB compliance and secure access	Transcripts available for AI processing
3	Develop structured prompts	3 core researchers	Align prompts with research questions using best-practices for prompt engineering	Chain-of-thought prompt templates
4	Generate preliminary codes and themes	3 core researchers using GPT-4o (via ASU CreateAI platform)	Independently apply structured prompts in GPT-4o to assist with identifying and editing preliminary codes, definitions, quotations, and rationales	Draft analytic output (human-AI assisted)

5	Review and re-fine GPT-human output	3 core researchers	Compare and reconcile individual code lists, validate quotations, revise theme labels, and elaborate definitions	Refined code-book draft
6	Independent transcript coding	2 independent researchers	Provide triangulation and external validation of themes	Annotated coding and analytic memos
7	Final synthesis and theme integration	Full research team	Resolve discrepancies and consolidate overarching categories	Finalized themes
8	Generate thematic responses to RQs	GPT-4o + researchers	Translate themes into structured answers to research questions	Aligned themes and design recommendations summary
9	Analyze survey responses	Human researchers	Describe participant perceptions and instructional priorities	Descriptive statistics, ranked metrics
10	Compare qualitative and quantitative results	GPT-4o + researchers	Identify convergence and divergence across data sources	Integrated mixed-methods findings

Note. This workflow reflects a general inductive analysis approach (Thomas, 2006), with structured coding and iterative synthesis informed by Saldaña (2014). Integration procedures follow convergent validation strategies in mixed-methods research (Fielding, 2012; Creswell & Clark, 2017).

4.6.2. Prompt Engineering and LLM Analysis

Prompt development was conducted iteratively by three researchers and guided by best practices that emphasize prompts that are concise, logical, explicit, adaptive, and reflective (Chen et al., 2024; Lo, 2023). Informed by insights from the literature, the team collaboratively designed prompts that met three core criteria: (1) clearly specifying the model's role, (2) defining input and output formats, and (3) incorporating reasoning-based strategies (Theelen et al., 2024; Zhang, Wu, Xie, Kim, & Carroll, 2023; Zhang, Wu, Xie, Lyu, et al., 2023, Zhang et al., 2024).

Prompts were constructed to distinguish among open, axial, and selective coding phases and included structured input/output templates, role-based framing (i.e., instructing the model to act as a qualitative researcher), and chain-of-thought prompting to elicit the rationale for each analytic step (Theelen et al., 2024; Zhang, Wu, Xie, Kim, & Carroll, 2023; Zhang, Wu, Xie, Lyu, et al., 2023, Zhang et al., 2024). Additional inspiration came from studies proposing systematic prompt frameworks for inductive coding and codebook generation, which emphasize the importance of consistency in labeling, justification logic, and thematically organized outputs (Katz et al., 2024; Zhao et al., 2024).

Because CreateAI Chat does not support a persistent system-level prompt or allow modification of temperature settings (fixed at 0.5), all role instructions and analytic specifications were delivered through the initial user prompt at the beginning of each session. This opening prompt instructed the model to conduct open coding by identifying salient concepts, naming and

defining each code, selecting representative quotations, and providing a brief justification for each coding decision.

Follow-up prompts were used to refine outputs by requesting clearer code definitions, additional examples, and attribution of quotes (Theelen et al., 2024). This approach reflects a form of chain-of-thought prompting (Wei et al., 2022), where the model is guided to articulate intermediate reasoning steps, consistent with prior applications of prompting in qualitative data analysis (Zhang, Wu, Xie, Lyu, et al., 2023; Zhang et al., 2024).

Illustrative prompts and corresponding outputs are summarized in Table 4. The model's output included preliminary themes related to usability, clarity, and interpretation of WAT-C's metrics, which were subsequently reviewed and refined by the research team (see Section 5.5.4).

Table 4: Illustrative prompts and corresponding LLM outputs

Prompt Step	Purpose of Prompt	Illustrative Prompt Text	Illustrative Model Response
Step 1: Role Assignment	Signal the model's role and calibrate expectations to generate structured, expert-like responses.	"You are an expert qualitative researcher with extensive experience in thematic analysis. I will provide you with qualitative data, and I want you to analyze it step-by-step, explaining your thought process at each stage. Please confirm you understand this role."	"Yes, I understand the role. I will guide you through the step-by-step process of conducting a thematic analysis on the qualitative data you provide, explaining my thought process at each stage."
Step 2: Contextual Grounding	Provide the model with necessary background knowledge to guide accurate and context-aware analysis.	"The data is from a focus group study with college-level writing instructors evaluating an early version of a feedback tool called WAT. Confirm you understand the focus group research study."	"Yes, I understand the context of the focus group research study. The study involves college-level writing instructors evaluating WAT's metrics for student feedback."
Step 3: Anchor to Qualitative Framework	Ground the model in an established qualitative method (Braun & Clarke, 2006) and align with prior research.	"Use the file 'thematic_analysis_revised' as a guide for how to conduct qualitative analysis. Confirm you understand thematic coding."	"Yes, I understand thematic coding. Thematic analysis involves identifying, analyzing, and reporting patterns (themes) within data. The process typically includes familiarization, initial coding, theme generation, review, naming, and reporting."
Step 4: Conduct Open Coding	Guide the model to generate rich, multi-participant code structures that capture contextual detail and attribution.	"Here is the text from one transcript. Conduct initial open coding on this data. Identify key concepts, phrases, or ideas that appear significant. For each code, provide the code name, definition, example quotes from multiple participants, and a	Code Name: Customization of Metrics Definition: The need for instructors to customize metrics based on individual student needs and assignment types. Representative Quote: "I would love complete customization for what we do, decide to include or not include." Rationale: This code highlights the desire for flexibility in the tool to cater to

		rationale. Present your results in a table."	diverse teaching styles and student needs.
Step 5: Revise for Rigor	Prompt the model to refine and substantiate its analysis by providing multiple supporting auditability, participant voice, and qualitative rigor.	"Can you revise so that you attribute the example quote to the participant in the transcript? Also add more than one representative example for each code."	<p>"Certainly! I will revise the initial open coding...</p> <p>Definition: The need for instructors to customize metrics based on individual student needs and assignment types.</p> <p>Example Quotes:</p> <p>Anna: "I would love complete customization for what we do, decide to include or not include."...</p>
Step 6: Conduct Focused Coding	Refine initial codes through analytic reasoning, allowing for theme emergence based on salience and relevance.	"Review the open codes you've created for both transcript files. Now, conduct focused coding by identifying relationships between codes, merging similar ones, elevating frequent codes, and discarding less relevant ones. Explain your decision-making process."	To conduct focused coding, I will review the open codes created for both transcript files, identify relationships between codes, merge similar codes, elevate frequent or significant codes, and discard less relevant codes... Discarding less relevant codes helped to focus the analysis on the most critical aspects of the participants' feedback, ensuring that the final set of codes provides a clear understanding of the key themes emerging from the data.
Step 7: Develop Themes	Organize data into meaningful, research-aligned themes that synthesize participant perspectives.	"Based on your focused coding, develop overarching themes. For each theme, provide a name, description, list of contributing codes, explanation of code relationships, and illustrative quotes."	<p>Theme 1: Customization and Flexibility</p> <p>Description: This theme reflects the participants' strong desire for the Writing Analytics Tool (WAT) to be customizable and flexible to meet the diverse needs of instructors and students... Illustrative Quotes:</p> <p>Kayla: "I'd be really curious to see how I would fit this into canvas...that's how I record their grades."...</p>
Step 8: Summarize by Research Question	Align thematic findings with the study's research questions for reporting and interpretation.	"Now, using the themes you created, summarize the answers to the following research questions: [To what extent and in what ways are WAT-produced metrics interpretable and useful to writing instructors and their students?]. Include representative quotes."	<p>Summary of Research Questions Based on Analysis</p> <p>1. Perceived Usefulness of WAT Metric Interpretations</p> <p>To what extent and in what ways are WAT-produced metrics interpretable and useful to writing instructors and their students?</p> <p>Summary: The perceived usefulness of WAT-produced metrics varies among instructors. While some see potential in the tool for providing automated feedback and reducing workload...</p> <p>Illustrative Quotes:</p> <p>Gavin: "The problem for me is that I don't have control of how it's been graded because the criteria there is not really clear."</p>

Note. LLM output responses were edited for brevity.

4.6.3. Human Validation and Thematic Synthesis

Following the initial GPT-4o-assisted open coding phase, three core researchers independently applied structured prompts to analyze segmented transcript data. These researchers were integrally involved in the design of the study—they co-developed the semi-structured focus group protocol, facilitated the two sessions, and conducted multiple close readings of all transcripts. Each researcher reviewed the model’s codebook outputs and identified points of convergence and divergence across segments.

To guide the analysis, we adopted a general inductive approach (Thomas, 2006), which emphasizes deriving findings from raw data in alignment with the study’s evaluation objectives, which were improving the WAT metrics, feedback, and interface to align with instructor needs. Researchers began by reviewing the AI-generated codebook and transcript excerpts produced through structured LLM prompting, then independently reanalyzed the transcripts using the same segmentation and prompt structure. Following Thomas’s (2006) recommended procedures, each researcher conducted multiple close readings of the transcripts, identified segments relevant to the evaluation objectives, labeled and grouped these into preliminary categories, and generated analytic memos to track interpretive decisions.

To clarify the sequencing and interaction of analytic activities within this phase, the human validation process unfolded through five interrelated stages (see also steps 3-7 in Table 3). The first stage, LLM-assisted open coding, involved using GPT-4o to generate preliminary codes, definitions, and representative quotations from segmented transcripts. In the second stage, independent human reanalysis, three core researchers re-coded the same transcript segments using a general inductive approach, producing individualized code lists and analytic memos. The third stage, collaborative comparison and synthesis, brought the team together to compare their independent analyses with the AI-generated codebook, merge overlapping codes, and refine category definitions through consensus. The fourth stage, external validation review (Table 3, step 6), engaged two additional researchers, who had not participated in prompt engineering or the creation, refinement, or synthesis of the AI-generated codebook. These researchers independently read, coded, and wrote analytic memos for the full transcripts, and then verified that emerging themes accurately reflected participant perspectives. Finally, in the consensus and thematic consolidation stage, the full research team discussed discrepancies, confirmed representative quotations in context, and synthesized the four overarching themes that informed the design recommendations. Although these activities occurred in distinct phases, the analytic process was recursive, with continual movement between AI-generated outputs, human interpretation, and collaborative reflection to ensure interpretive rigor and credibility.

During team meetings, researchers examined points of convergence and divergence in their individual analyses, resolved discrepancies through discussion, and refined the organization and naming of themes. Representative quotations were verified in context, and additional excerpts were selected from underrepresented participants to ensure diverse perspectives were reflected in the findings. Although the LLM served as a structured analytic starting point, final theme development was guided by human interpretation, analytic consensus, and alignment with the study’s formative design goals.

Through this comparative review, we identified four recurring themes in the AI-generated output, reflecting key areas of focus in the data: (1) clarity and interpretation of metrics and feedback, (2) need for customization and flexibility, (3) integration with existing instructional tools, and (4) potential for pedagogical improvement. These themes served as the foundation for subsequent human refinement. Drawing on standard qualitative procedures, we synthesized overlapping codes, expanded descriptions, verified all representative quotations in context, and

added new excerpts from underrepresented participants. In this way, the research team used the AI output as a structured analytic starting point but relied on iterative human interpretation to finalize the themes presented in the Findings section.

4.6.4. Quantitative Data Analysis

Quantitative data from the post-session survey were used to support the primary qualitative findings of instructors' perceptions of WAT's perceived usefulness, perceived ease of use, behavioral intention to adopt, and perceived feasibility for classroom integration. Given the small sample size ($N = 9$), descriptive statistics were calculated to summarize patterns in responses to each Likert-scale item associated with the TAM constructs. Metric ranking items were summarized using median ranks and response distributions to illustrate perceived instructional priorities across lexical, syntactic, and writing style features. Open-ended responses were reviewed to identify commonly expressed concerns, suggestions, or points of emphasis. These findings were integrated with the qualitative analysis during the interpretation phase to contextualize and complement the qualitative analysis.

4.6.5. Integration of Quantitative and Qualitative Findings

To support integration within the convergent mixed-methods design, we employed a process of convergent validation, a form of triangulation in which qualitative and quantitative results were analyzed independently and then compared for alignment and divergence (Creswell & Clark, 2017; Fielding, 2012). GPT-4o was used to assist in this step by comparing final qualitative themes with participants' survey responses, including Likert-scale ratings and metric rankings. The model was prompted to identify areas of convergence between participants' stated perceptions and the patterns observed in the focus group transcripts. These AI-assisted comparisons were then reviewed and interpreted by the research team to strengthen the validity and completeness of the integrated findings.

5. FINDINGS

5.1. FROM AI OUTPUT TO HUMAN-REFINED THEMES

The following themes were derived through a collaborative process that blended GPT-4o-generated outputs with interpretive human analysis. As described in the Methods section, each of the three researchers independently used GPT-4o to generate themes from transcript segments. We then reviewed the model's outputs and met to compare, critique, and consolidate the AI-suggested themes. This iterative review process resulted in four refined themes that retained some of the initial model structure but diverged in their emphasis, coding, and use of representative quotations. To ensure transparency and traceability, all initial codebooks developed through the human-AI workflow are available on OSF (see Table 3 and https://osf.io/8j9y6/?view_only=3a21ed900475418a8c258ea595cbe53b). The final themes presented in Table 5 were not solely generated by the LLM but reflect iterative human interpretation, consolidation, and revision based on researchers' discussions and review of the initial outputs.

Table 5: Final themes and contributing codes from qualitative data analysis.

Final Theme	Contributing Codes
Theme 1: Instructors believe the content and presentation of WAT metrics and feedback can be improved.	Redundancy in metrics Metrics require elaborated definitions and examples Metric and feedback presentation can be clearer
Theme 2: Instructors require WAT to be customizable and flexible.	Instructors want to customize metrics and feedback Agreement with metric definitions, scores, and feedback was inconsistent
Theme 3: Instructors stressed that WAT should be integrated with existing tools to improve feasibility.	Adopting WAT may be compromised by workload concerns for faculty WAT should be integrated with existing technologies
Theme 4: Instructors indicated that adjustments to WAT features and interface could improve pedagogical impact.	Improvements to the interface can help instructors access information and assign writing tasks more efficiently Improvements to the interface and functionality can help facilitate student reflection Students may be interested in learning from WAT, but it depends on their learning preferences

5.2. FINAL THEMES FROM FOCUS GROUP ANALYSIS

This section presents the four themes derived through a human-AI collaborative workflow. Initial codes were generated using GPT-4o and refined by the research team through iterative analysis and member checking. Each theme reflects instructors' perspectives on the usability, interpretability, and instructional integration of WAT. Quotes have been retained verbatim to preserve the authenticity of participants' voices.

5.2.1. Theme 1: Improve Clarity and Redundancy of Metrics and Feedback

Instructors consistently emphasized the need for clearer metric definitions and more coherent feedback presentation. Several participants expressed concerns about redundancy across metrics, difficulty distinguishing overlapping categories, and the cognitive burden imposed by too many features.

REDUNDANCY IN METRICS. Participants highlighted overlapping content across WAT's advanced metrics. Kayla found the distinction between Word Concreteness and Conventional Language unclear. Ava suggested combining Word Variety and Information Density to reduce cognitive load. Seth proposed grouping related metrics into subcategories to help "reduce the cognitive load necessary [for students] to internalize them." While instructors could choose which metrics were visible to students, many felt that simplifying the set of available metrics would improve usability.

Jason remarked on similarities between Sentence Cohesion and Sentence Structure Variety: "When sentence cohesion is low, then sentence structure variety is consistent and repetitive." Kayla noted that Information Density may not accurately capture the writing features being

assessed, stating, “I feel that by calling that [feedback] concise...I don’t love the word dense used in this sense.” Gavin and Kayla discussed replacing Information Density with terms like “supporting details” or “depth,” though they acknowledged possible overlap with Development of Ideas. Ultimately, instructors agreed that Information Density should be folded into Development of Ideas to reduce confusion.

EXAMPLES AND VISUAL ALIGNMENT TO STUDENT TEXT. Instructors urged the inclusion of examples to help students understand and act on the feedback. Ava said that “examples would be helpful to include in the ‘Read Here to Learn More’ section.” Jason noted the benefit of “just a longer definition” for each metric. Kayla emphasized the need for brief descriptions to aid interpretation.

Participants felt feedback would be more actionable if specific features in the student’s writing were visually linked to the assigned metric. Jason, drawing from experience with Grammarly, worried students would revise without understanding why: “They just change it bit by bit like a puzzle until they get a certain mark.” Seth supported the idea of clickable metric components: “So they’re seeing the connection between the concept of academic language and the specific expression in their essay.” He explained that students should “be able to click on the casual [feedback icon], see examples of casual language, click on the academic and see more... examples of academic language which is highlighted in the actual submission.”

5.2.2. Theme 2: Instructors Require a Customizable and Flexible WAT

Participants stressed that WAT should be adaptable to diverse instructional goals, classroom contexts, and writing assignments. While instructors welcomed descriptive metrics, they sought greater control over how those metrics were presented and interpreted.

METRICS AND FEEDBACK SHOULD BE CUSTOMIZABLE. Anna noted that while instructors at the institution shared “course outcomes,” they “come with different experiences. We work from our strengths.” She expressed a desire to select relevant metrics depending on the prompt and pedagogical focus. Jason shared that if students are asked to write informally, feedback suggesting an academic tone could undermine the instructional goal.

Seth noted that his curriculum includes instruction on source integration. Ideally, WAT would let him “focus on the elements... that I need at the moment.” He explained that using all metrics all the time would require him to “adjust a lot to what the tool is doing rather than sort of picking and choosing a little bit.” While current WAT metrics are fixed, some participants envisioned future versions of the tool that could incorporate GenAI to support more personalized, context-sensitive feedback. For instance, GenAI could generate revision suggestions aligned with instructor-defined priorities for a given assignment, reducing the need for instructors to manually adjust or explain mismatched feedback.

LIMITS OF CUSTOMIZATION AND NEED FOR GUIDANCE. Instructors recognized that excessive customization could affect interpretability. Gavin warned, “That could create a conflict on how the text is assessed.” He recommended allowing instructors to adjust metric targets (e.g., aiming for “Balanced” in Sentence Cohesion) rather than altering metric definitions. Anna also noted that students may misinterpret feedback without this guidance: “Allowing the instructor to set and say you should be striving for balanced [the middle of the spectrum]... they can see what they should be striving for.”

INCONSISTENT AGREEMENT WITH METRIC DEFINITIONS, METRIC VALUES, AND FEEDBACK. Gavin noted that some metrics felt unclear or misaligned with his judgment, saying, “Do I have to adjust it myself and then have to explain, or even argue with a student?” Autumn, Marcus, and Gavin questioned the definition of Sentence Cohesion. Autumn noted it involved “more

than just using similar words in your sentences.” Marcus added that it ignored “paragraph organization,” and Gavin explained, “Cohesion has to [do] with connections, transitions, and, in a way, to avoid repetition.” Jason emphasized the need for transparency: “If I’m using this and I agree with all the metrics, I would still have to scrutinize every metric to make sure... I would be in agreement with [the metric value and definition].”

5.2.3. Theme 3: WAT Must be Integrated with Existing Tools

Participants raised concerns about WAT increasing instructor workload and expressed a desire for seamless integration into existing teaching platforms.

PERCEIVED BURDEN AND WORKFLOW CONCERNS. Ava worried that using WAT would “add more to my workload.” Anna, who teaches over 125 students per semester, described setting individual metrics as “an unrealistic burden.”

DESIRE FOR LMS INTEGRATION. Several instructors advocated integrating WAT into Canvas. Anna stated, “The more [WAT] can be integrated with Canvas the better it’s going to be.” Kayla envisioned WAT being used as frequently as Turnitin: “If it was just as easy as the Turnitin plagiarism checker... I’m absolutely doing this every single time.” Even if students didn’t initially like the feedback, she said, she would assign WAT “for every single assignment.”

5.2.4. Theme 4: Improvements to WAT Interface can Enhance Pedagogical Potential

Instructors suggested improvements to WAT’s user interface to support interpretation, feedback engagement, and revision. Ava advocated for student interactivity: “To engage with the feedback. So in some way where you can actually comment on the application, their essay, because otherwise it would just be another [learning tool].” She added that students could reflect “on the information they received and how it will impact their next draft.”

Tara recommended simplifying descriptions to avoid overwhelming students. Gavin pointed out that “the color scheme may not be as informative and overload the interpretation of the score.” Kayla liked the use of segmented scoring: “That’s going to be really similar to other types of scales that they’ve seen... and how our rubrics are set up.” Kayla added that tooltips could be improved: “Maybe making that a little more apparent to [students] could be beneficial.” Ava proposed refining the flow of the teacher interface to improve navigation and usability.

5.3. QUANTITATIVE RESULTS

Instructors responded to a series of survey items evaluating the perceived usefulness, ease of use, interpretability, and adoption potential of WAT’s core features. The results are organized below by construct and survey domain, and related implementation factors.

5.3.1. Perceived Usefulness of WAT Metrics

Instructors rated the usefulness of WAT metrics for supporting writing instruction on a scale from 1 (Not useful at all) to 6 (Extremely useful). As shown in Table 6, *Development of Ideas* received the highest overall rating, followed closely by *Academic Focus* and *Varied Sentence Structure*, two of the more advanced analytic features. *Conventional Language*, *Word Concrete-ness*, and *Sophisticated Wording* received lower ratings, indicating that instructors may perceive these features as less useful in their instruction. The predictive quality scores received the lowest ratings overall, suggesting they were viewed as least useful for formative feedback, which is consistent with a previous design study for WAT (Li et al., 2022).

Table 6: Perceived usefulness of WAT metrics ranked by mean rating

Metric	Mean (SD)	Min	Max
Advanced Analytics			
Development of Ideas	4.75 (2.05)	1	6
Academic Focus	4.25 (1.75)	1	6
Varied Sentence Structure	4.22 (1.39)	2	6
Word Variety	4.00 (1.58)	2	6
Academic Writing	3.88 (1.36)	1	5
Language Variety	3.78 (1.30)	2	6
Source Integration	3.62 (1.51)	1	5
Conversational Writing Style	3.50 (1.31)	1	5
Function Word Repetition	3.33 (1.12)	2	5
Sentence Cohesion	3.22 (1.72)	1	6
Information Density	3.22 (1.30)	1	5
Sophisticated wording	3.11 (1.54)	1	6
Conventional Language	2.89 (1.54)	1	5
Word Concreteness	2.67 (1.12)	1	4
Basic Analytics			
Word Count	4.33 (1.50)	2	6
Spelling Errors	4.33 (1.50)	2	6
Paragraph Count	3.44 (1.33)	2	5
Sentence Count	2.78 (1.39)	1	5
Average Sentence Length	2.67 (1.32)	1	5
Average Word Length	1.78 (1.30)	1	4
Predictive Quality Scores			
Wording Quality (Summary Only)	3.11 (1.27)	1	5
Content Quality (Summary Only)	3.00 (1.12)	1	4
Overall Quality Score	2.33 (1.22)	1	5

Note. Ratings were on a 6-point scale ranging from 1 (Not useful at all) to 6 (Extremely useful).

In addition to evaluating the metrics themselves, instructors rated the interpretability and clarity of WAT's visual feedback displays. These items addressed the clarity and usefulness of visual features, such as the color-coded spectrum, the "Click Here to Learn" link, and the "i" icon definitions. As shown in Table 7, instructors viewed the clickable help features to be useful instructional supports. However, they expressed more mixed views about the visual spectrum: while its inclusion was seen as moderately helpful, lower ratings for label accuracy and ease of understanding suggest that this display may require redesign to better support student understanding.

Table 7: Instructor ratings of WAT metrics visual design and on-screen support

Item Description	Mean (SD)	Min	Max
-------------------------	------------------	------------	------------

“Click Here to Learn” Link Useful	4.00 (1.00)	2	6
“i” Icon Definitions Helpful	3.67 (1.12)	2	5
Visual Spectrum Helpful	3.67 (1.32)	1	5
Visual Spectrum Labels Accurate	3.11 (1.17)	1	5
Visual Spectrum Easy to Understand	2.67 (1.22)	1	4

Note. Items were rated on a scale from 1 (Strongly Disagree) to 6 (Strongly Agree).

5.3.2. Perceived Usefulness and Structure of WAT Feedback Comments

Instructors also rated the perceived usefulness and clarity of WAT’s written feedback comments, which are associated with each advanced metric. These comments are organized into a “Why–When–How” format that aims to support student revision by explaining the purpose of a metric (Why), its ideal application (When), and actionable strategies (How). As shown in Table 8, instructors generally rated the feedback comments positively. They found the comments helpful for revision and easy to understand, with the WHEN and HOW sections receiving particularly strong ratings. The overall structure of the feedback was seen as beneficial for guiding revision. However, the WHY section, which includes metric definitions, received more mixed responses, and the lowest-rated item asked whether additional detailed definitions might be helpful or overwhelming. These results suggest that simplifying or refining feedback comments and their presentation may improve accessibility for students.

Table 8: Instructor ratings of WAT feedback comments and structure

Metric	Mean (SD)	Min	Max
Feedback comments are helpful for revision	4.11 (1.17)	2	6
WHEN section clarifies when the current writing is beneficial	4.11 (0.78)	3	5
Overall “Why–When–How” structure is helpful for student revision	4.11 (0.78)	3	5
Feedback comments are easy to understand	4.0 (0.87)	3	5
HOW section provides specific revision strategies	4.0 (0.71)	3	5
Feedback comments are displayed clearly	3.89 (0.93)	3	5
Comments contain the appropriate amount of information	3.78 (0.97)	2	5
WHY section helps students decide whether to revise	3.67 (1.41)	1	5
Additional metric definitions would be helpful vs. overwhelming (reverse-coded)	2.67 (1.22)	1	5

Note. Items were rated on a scale from 1 (Strongly Disagree) to 6 (Strongly Agree).

5.3.3. Perceived Usefulness of Instructional Affordances

Instructors evaluated the perceived usefulness of specific WAT features related to instructional tasks (Table 9). These affordances were grouped into two categories: features for assigning writing tasks and features for providing feedback. Participants responded to each item using a 6-point scale ranging from 1 (Not useful at all) to 6 (Extremely useful). Ratings reflected moderate to high perceived usefulness for features that supported task customization, scaffolded revision, and selective analytics displays. The most highly rated affordances included uploading personalized rubrics and assigning multiple drafts for revision. In contrast, features related to fixed

writing task types (e.g., summary writing) and directing students to basic and advanced analytics received lower scores.

Table 9: Instructor ratings of usefulness of WAT instructional affordances

Affordance Description	Mean (SD)	Min	Max
Assigning Writing Tasks			
Uploading personalized rubrics for student assignments	4.56 (1.51)	1	6
Assigning multiple drafts for revision	4.44 (0.88)	3	6
Selecting Basic Analytics for Student View	4.22 (1.30)	2	6
Uploading source texts to writing tasks	4.11 (1.62)	1	6
Selecting Advanced Analytics for Student View	3.89 (1.69)	1	6
Creating Source-Based Writing Tasks	3.67 (1.41)	1	5
Creating Persuasive Writing Tasks	3.33 (1.66)	1	6
Creating Summary Writing Tasks	2.00 (1.50)	1	4
Providing Feedback			
Directing Students to Review Basic Analytics	3.56 (1.24)	2	5
Directing Students to Review Advanced Analytics	3.33 (1.32)	1	5
Viewing basic analytics for individual students	3.22 (0.83)	2	4
Communicating via WAT's teacher comment feature	3.22 (1.56)	1	5
Viewing advanced analytics for individual students	3.00 (1.22)	1	5

Note. Participants rated each item on a scale from 1 (Not useful at all) to 6 (Extremely useful).

5.3.4. Behavioral Intention to Adopt and Perceived Feasibility

Instructors reported moderate perceptions of WAT's usefulness and feasibility for classroom adoption (see Table 10). Ratings of the tool's potential to improve student outcomes were slightly below the scale midpoint. The overall likelihood of use and perceived usefulness for teaching also clustered near the midpoint, suggesting uncertainty about integrating the tool into regular instruction. Instructors also expressed moderate concern about adoption feasibility, particularly regarding potential impacts on instructional workload and time efficiency, as reflected in the lowest mean rating for anticipated effects on instructional time.

Table 10: Instructor perceptions of intention to use and perceived feasibility of WAT

Metric	Mean (SD)	Min	Max
Likelihood WAT Will Improve Student Outcomes	3.33 (1.41)	1	5
Concern About Adopting WAT	3.11 (1.62)	1	6
Usefulness of WAT for Teaching	3.00 (1.32)	1	5
Likelihood of Using WAT	3.00 (1.58)	1	5
Effect of WAT on Instruction Time	2.56 (0.73)	1	3

Note. Ratings were collected on a 6-point scale. Higher values indicate greater agreement or likelihood, except for "Effect of WAT on Instruction Time," where lower scores indicate greater anticipated burden.

5.4. INTEGRATED FINDINGS TO INFORM DESIGN

The final phase of analysis integrated qualitative focus group themes, quantitative survey data, and short-answer responses to identify user-centered priorities for redesigning WAT. To center instructor voices and in keeping with the study's convergent mixed-methods design, qualitative data were prioritized in the development of design recommendations (QUAL + quan), while the quantitative ratings and open-ended responses were used to corroborate and extend the emerging themes (Creswell & Clark, 2017; Fielding, 2012). Open-ended survey responses provided elaboration and clarification of key themes and offered participants an additional opportunity to articulate their perceptions and suggest improvements, enhancing the interpretive depth of the findings.

As shown in Table 11, the four major themes that emerged from focus group transcripts were supported by converging evidence from both survey ratings and written responses. Recommendations to clarify metrics and feedback (Recommendation 1) and enhance user control through customization (Recommendation 2) reflected concerns about interpretability, flexibility, and instructional alignment. Similarly, concerns about perceived ease of use and perceived feasibility (Recommendations 3 and 4) were reinforced by moderate quantitative scores for interface navigation and anticipated instructional workload, as well as multiple instructor comments requesting integration with LMS platforms and more streamlined user interface layouts.

Table 11: Integrated evidence supporting design recommendations for WAT

Recommendation	Focus Group Themes	Quantitative Support	Short-Answer Support
1. Redesign feedback and metric presentation	Theme 1: Instructors believe the content and presentation of WAT metrics and feedback can be improved.	Low usefulness ratings for metrics (e.g., Word Concreteness $M = 2.67$); Low clarity ratings for visual spectrum and WHY section	Tara: "I would suggest having fewer metrics. Though it could be informative, it also felt tedious and confusing at times to distinguish between them." Autumn: "The current display will be confusing for students, because they will assume they want to score whatever is furthest to the right, which is not always the case."
2. Introduce customization features for instructors	Theme 2: Instructors require WAT to be customizable and flexible.	High support for customizable affordances (e.g., uploading rubrics $M = 4.56$; setting student views)	Gavin: "The teacher should have the ability to set their own expectation and score based on the expectation of the teacher and not of the system." Kayla: "The more streamlined the information can be for students, and the more teachers have the chance to modify or add materials like rubrics or definitions/examples of advanced analytics, the better."

3. Reduce workload with LMS integration	Theme 3: Instructors stressed that WAT should be integrated with existing tools to improve feasibility.	Concern about instructional time ($M = 2.56$) and about adoption ($M = 3.11$)	<p>Kayla: “If you can find a way to add this into Canvas as an extension or something like that, that would be amazing! I grade my students rough drafts mostly for completion points, so if I can assign it to them in Canvas for them to turn into WAT, and grade it in Canvas, that would be a huge game changer in the ease of access and usability to this tool.”</p> <p>Ava: “If it's not integrated into Canvas, it could be a barrier to using for me.”</p> <p>Gavin: “An AI integration seems to be critical for this application should be considered relevant at the moment.”</p>
4. Improve user interface design	Theme 4: Instructors indicated that adjustments to WAT features and interface could improve pedagogical impact.	Visual design ratings varied: Spectrum clarity ($M = 2.67$), tool tip helpfulness ($M = 3.89$)	<p>Kayla: “Less information, and less clicks to get to this information.”</p> <p>Gavin: “The feedback should be streamlined into fewer visual cues and displayed in one soap location with the minimal amount of clicks possible. I visualize this page as only having this score for the metrics and a brief description, for what the metric means, click here to learn quotes on the side of the text.”</p>

Note. Qualitative themes were drawn from focus group transcripts; quantitative support reflects mean scores from relevant survey items (see Section 5.3.). Short-answer responses were used for convergence and elaboration.

6. DISCUSSION

This study makes two connected contributions to the field of educational data mining and writing analytics. The primary methodological contribution is a hybrid AI-human workflow for inductive thematic analysis that demonstrates how LLMs can assist researchers in generating and refining qualitative codes and themes, particularly in a participatory design research process. The secondary applied contribution is the use of this workflow to guide revisions to WAT through participatory analysis of instructor feedback. These contributions clarify how LLM-supported qualitative inquiry can strengthen both the rigor and practicality of participatory design research. The following sections examine the evaluation of the LLM within this workflow and the resulting design recommendations for WAT.

6.1. EVALUATION OF LLMs IN QUALITATIVE ANALYSIS AND MIXED METHODS INTEGRATION

The use of GPT-4o in this study offered both methodological and practical benefits, particularly in accelerating early-stage coding and supporting the integration of qualitative and quantitative data. The model was used to generate initial theme labels, code definitions, representative quotations, and justification logic. In some cases, GPT-4o helped surface relevant patterns across transcript segments and suggested useful groupings of themes for the joint display.

However, the outputs produced by the model required substantial researcher oversight and post-processing. Prior research suggests that while LLMs may reduce time in deductive coding tasks (Dunivin, 2024), they must be used cautiously in inductive analysis and data integration workflows. Model-generated codes in such contexts can be shallow, inconsistently applied, and difficult to validate due to limited transparency and a lack of reliable error-detection mechanisms (Bijker et al., 2024; Chen et al., 2024). Our findings are consistent with this body of research: LLMs should not be viewed as replacements for human analysis and require continuous oversight to ensure accuracy and interpretive integrity (Bijker et al., 2024; Katz et al., 2024), particularly in mixed method and participatory design research contexts.

Throughout this study, model-generated theme definitions required refinement, and illustrative quotations had to be reviewed and contextualized by human researchers. During the integration of focus group transcripts and open-ended survey responses, GPT-4o occasionally excluded or reworded quotations, even when prompted to return verbatim excerpts. As such, these inaccuracies introduced potential for misrepresentation and required full researcher oversight during the triangulation and integration phases. Another example of these limitations were evident in the prompting process illustrated in Table 4. For example, Step 5 (“Revise for Rigor”) shows how the research team prompted GPT-4o to attribute example quotations to individual participants and to provide multiple representative excerpts for each code. Without this additional direction, the model produced generic or truncated quotations that lacked sufficient context for interpretation. However, the model did not hallucinate new quotations; all excerpts were drawn from the source data. Even though the LLM operated within the bounds of the provided dataset, its outputs still required researcher oversight.

Moreover, to ensure the accuracy and contextual integrity of the findings, two external researchers independently reviewed the full focus group transcripts and wrote analytic memos to validate or challenge the model-generated themes. This step was essential for verifying that participant perspectives were interpreted faithfully and that the resulting themes reflected the full complexity of the discussion. As such, the AI-assisted analysis did not reduce the overall analytic workload. In some cases, it introduced additional layers of review and verification, offsetting potential time savings.

This experience aligns with recent research suggesting that although LLMs may serve as helpful “research assistants” because they demonstrated value in accelerating discrete tasks (e.g., open-coding, generating themes, retrieving illustrative examples, and comparing participant responses), these time savings were in part offset by the need for careful verification, contextualization, and triangulation of model outputs, which require researcher oversight (Chen et al., 2024; Turobov et al., 2024). Rather than reducing the overall analytic workload, the LLM redistributed it by shifting effort from early coding to later interpretive review. This redistribution underscores the role of generative AI not as a labor-saving replacement for qualitative researchers but as a tool that can expedite specific analytic stages while requiring researcher oversight (Hayes, 2025).

Finally, the use of the ASU CreateAI platform influenced the workflow by providing a RAG system within a secure, institutionally managed environment. This infrastructure ensured that all data were processed securely and were not used to train OpenAI models, minimizing privacy and confidentiality risks associated with commercial AI platforms. Although access was limited to approved ASU users, this restriction strengthened data security and governance, which is an essential consideration when working with human subjects data. Overall, these design choices reflect a deliberate balance between methodological innovation and ethical responsibility, reinforcing that the value of LLM-assisted qualitative analysis lies not only in its analytic potential but also in the establishment of transparent, secure, and researcher-controlled environments for its use.

Although findings indicate that GPT-4o accelerated early-stage coding and supported theme development, an important limitation is that we did not systematically capture operational metrics that would allow for a quantitative comparison of the AI-assisted workflow and a human-only baseline. Prior research highlights the value of such measures and the affordance of LLMs to reduce time-on-task for qualitative data analysis (Bijker et al., 2024; Prescott et al., 2024). While our process revealed qualitative evidence of error correction, iterative prompting, and theme refinement (e.g., the extensive clarification and verification steps reflected in Table 3, Step 5), these activities were not tracked in a way that would permit systematic evaluation of efficiency gains or losses. Thus, future work should incorporate systematic tracking procedures that could, for example, log the frequency and types of AI errors, document rates of human edits and revisions to AI outputs, and assess intercoder agreement before and after AI assistance. Such metrics could be used to more rigorously evaluate analytic performance and workflow efficiency, and also identify where LLMs introduce additional layers of labor that potentially offset efficiencies.

6.2. DESIGN RECOMMENDATIONS

Based on instructor feedback from the focus group sessions, four primary areas for improvement were identified: clarity of feedback, need for customization, reduction of workload, and user interface enhancements. Below, we outline the design recommendations developed in response to these themes.

6.2.1. Recommendation 1: Redesign Feedback and Metric Presentation

A central theme from the focus group was the need for clearer, more concise feedback and a more intuitive interface for engaging with advanced metrics. Instructors noted that certain metrics felt redundant or difficult to distinguish and that students might struggle to interpret or apply the feedback meaningfully.

In response, we revised both the structure and presentation of the feedback interface (see Figure 3). In the previous design, students and instructors had to click separately on both ends of the continuum to explore the full range of feedback. The revised version streamlines this process by allowing users to click once to access a consolidated feedback view, which includes definitions, revision guidance, and task-relevant examples. To support usability and reduce cognitive load, the feedback pop-up now follows an accordion-style structure, allowing users to reveal specific sections, such as definitions, rationale, or revision strategies, only when needed (see Figure 4).

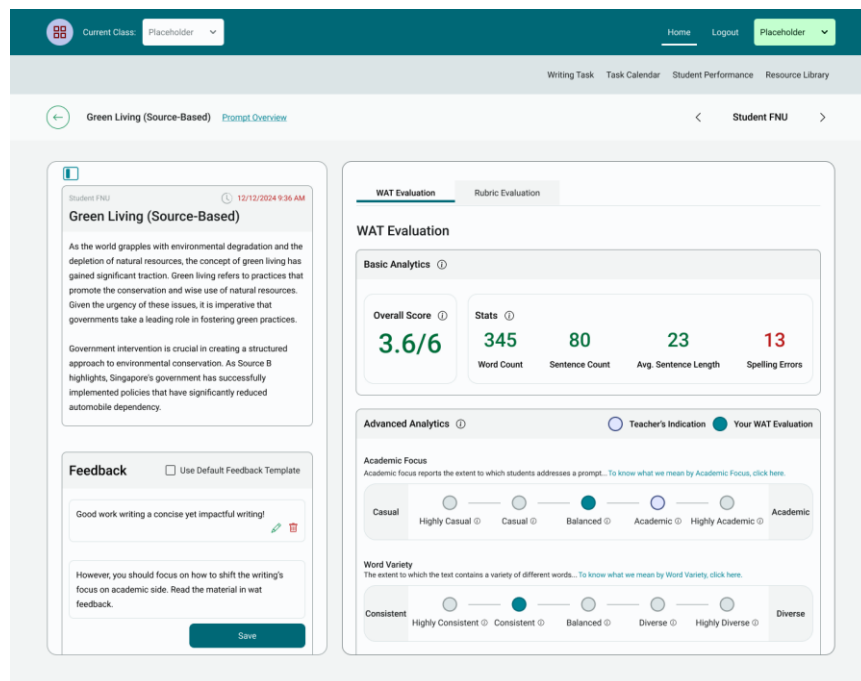


Figure 3: Revised metric feedback interface for students.

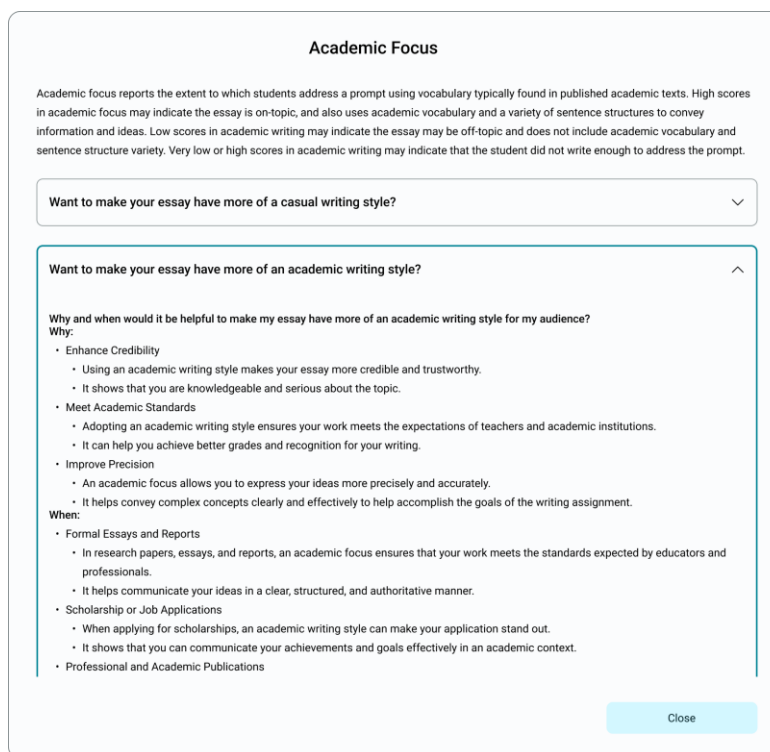


Figure 4: Revised written feedback student interface.

Additionally, the feedback language and organization were revised for clarity and precision. Based on instructor suggestions, we are also refining the taxonomy of metrics by grouping them into broader categories (i.e., lexical, syntactic, and rhetorical/style) and consolidating overlapping features. For instance, Information Density is being merged with Development of Ideas to

reduce redundancy while preserving core analytic distinctions. These updates are designed to make the tool more interpretable for instructors and more actionable for students revising their work.

6.2.2. Recommendation 2: Introduce Customization Features for Instructors

Instructors emphasized the need for greater customization to align WAT’s feedback with their specific pedagogical goals and classroom contexts. In response, we introduced features that allow instructors to set target values for each advanced metric during assignment creation (see Figure 5). These optional targets function as teacher-created benchmarks by highlighting the expected performance level on a given metric without altering the underlying model-generated evaluation.

The screenshot displays the 'Create a Writing Task' interface. At the top, there's a navigation bar with 'Current Class' and 'Placeholder' dropdowns, and links for 'Home', 'Class Settings', 'My Profile', 'Logout', and another 'Placeholder' dropdown. Below this is a secondary navigation bar with 'Writing Task', 'Task Calendar', 'Student Performance', and 'Resource Library'. The main content area is titled 'Create a Writing Task' and features a progress bar with three stages: 'Prompt' (light green), 'Metrics' (dark green), and 'Details' (grey). The 'Metrics' stage is currently active. Below the progress bar is the 'Metric Selection' section. It includes a sub-header 'Metric Selection' and a prompt: 'Please select the metrics you would like to use to measure students' writing.' There are two columns: 'Include All' and 'Description'. The 'Include All' column has a list of metrics with checkboxes: 'Overall Score' (checked), 'Basic Analytics' (checked), 'Word Count' (unchecked), 'Sentence Count' (checked), 'Paragraph Count' (unchecked), 'Spelling Error' (checked), 'Average Sentence Length' (unchecked), 'Advanced Analytics' (checked), 'Sophisticated Wording' (unchecked), 'Development of Ideas' (unchecked), 'Academic Focus' (checked), 'Sentence Coherence' (checked), 'Conversational Writing Style' (unchecked), and 'Word Concreteness' (checked). The 'Description' column provides details for the 'Academic Focus' metric, explaining that it reports the extent to which students address a prompt using academic vocabulary and sentence structures. It also includes a section for 'Indicate Desired Level' with a scale from 'Casual' to 'Academic'. The 'Academic Focus' metric is currently set to 'Academic' on this scale. A note at the bottom states: 'Note: Click and select a desired level for this metric. Students will see this target level along with actual evaluation against this metric.'

Figure 5: Revised teacher interface that allows teachers to select a desired metric goal when creating writing tasks

For example, an instructor might designate “Academic” as the target level for the Academic Focus metric, signaling to students that they should use formal, discipline-appropriate language. When enabled, target values are visually distinguished from the student’s actual metric value, allowing both instructors and students to identify areas where revisions are needed. This comparison supports formative instruction by clarifying expectations and guiding revision priorities.

To ensure feedback reliability, instructors are not permitted to modify metric definitions, which are based on statistical models of linguistic features. Preserving the core meaning of each metric ensures consistency and interpretability across writing tasks.

This instructor-facing customization is mirrored in the student interface, where system-generated values and teacher-designated targets are displayed in parallel (see Figure 3). This visual contrast encourages students to reflect on the alignment between their writing and instructional expectations. These features reflect key principles of personalized learning, which emphasize the importance of adaptable systems that empower educators to tailor tools to diverse classroom needs (Verma et al., 2024).

6.2.3. Recommendation 3: Reduce Workload

Instructors consistently noted that using WAT as a standalone tool increased their administrative workload. To address this concern, we developed a Learning Tools Interoperability (LTI) integration that allows WAT to be embedded directly within a LMS. This will enable instructors to assign writing tasks, view student submissions, and access WAT-generated scores and feedback all within the LMS interface. For example, this feature will allow assignment grades and feedback to be automatically passed back to a Canvas gradebook, which reduces manual input and aligns with instructors' grading workflows. This integration supports two key goals. First, it improves usability by aligning with instructors' existing workflows, making WAT easier to adopt and use consistently. Second, it aligns with TAM, which posits that ease of use and perceived usefulness contribute to greater technology adoption and behavioral intent (Davis, 1989).

In parallel, we are exploring the integration of LLMs into the WAT ecosystem. One current area of research and development involves leveraging WAT's automated scoring outputs as inputs for an AI-powered revision assistant. In an initial prototype, we developed a chatbot that guides students through revision strategies based on their individual WAT scores (see Figure 6). This AI assistant interprets the meaning behind the scores and suggests tailored revisions that align with the student's writing goals. While still in its experimental phase, this integration lays the groundwork for future iterations of WAT that incorporate adaptive feedback loops powered by LLMs, with the potential to significantly enhance writing instruction and student engagement.

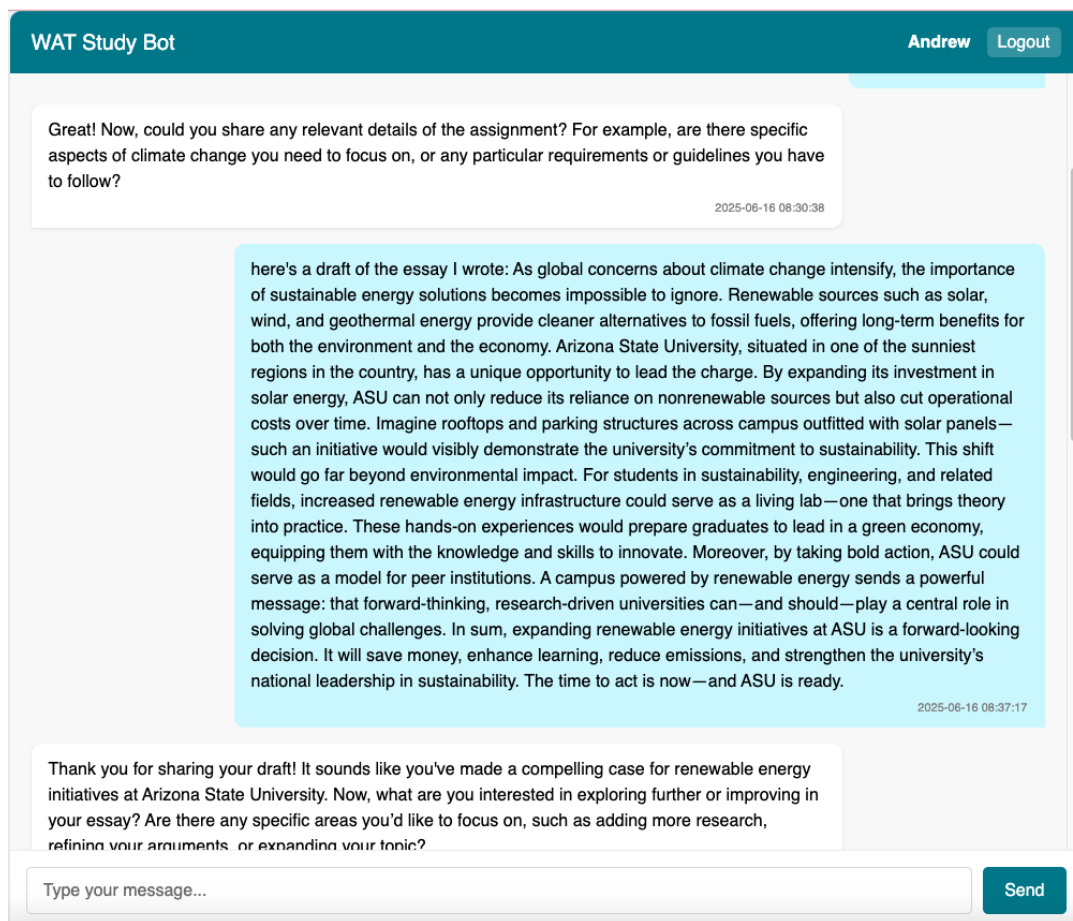


Figure 6: Alpha version of a chatbot to support students' revisions in WAT interface.

To further reduce workload, we also introduced custom rubric creation within WAT (see Figure 7). Instructors can now build rubrics tailored to their course goals and link these directly to WAT's automated metrics. This reduces the need to manually transfer scores between systems while preserving instructional flexibility. These changes respond directly to instructors' desire for more efficient feedback systems and customizable assessment tools.

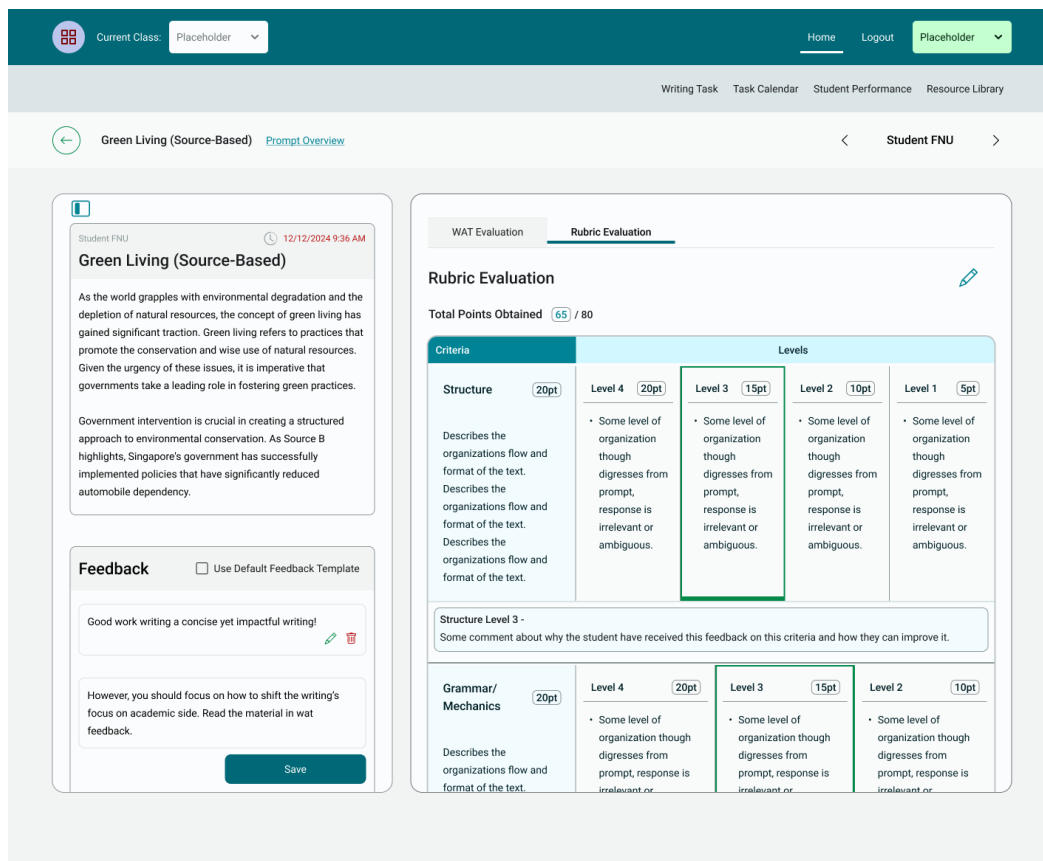


Figure 7: Revised rubric interface that integrates teacher-created rubric and associated feedback with WAT metric feedback.

6.2.4. Recommendation 4: Improve User Interface

The final theme focused on the user interface and its impact on the tool's pedagogical utility. Instructors suggested that improvements to the user interface could make the tool easier to navigate and more instructionally effective (see participant comments in Sections 5.2 and Table 11). Based on this feedback, we redesigned several user interface elements to improve clarity and reduce visual and cognitive complexity.

We replaced the previous color-coding scheme with a unified visual design that simplifies the layout and makes the metric display more accessible (see Figure 3). This revision aligns with cognitive load theory, which suggests that minimizing extraneous visual elements can help users focus on meaningful instructional tasks. In addition, we reduced the size of each metric container to allow more metrics to appear on screen at once, reducing the need for excessive scrolling and enabling faster comparison (see Figure 5).

We also redesigned the feedback interface by moving the student submission and teacher comment sections into a collapsible side panel (see Figure 7). This panel allows instructors to view WAT evaluations, student work, and rubric feedback simultaneously, supporting more efficient feedback cycles. Instructors can choose to expand or collapse the panel as needed, giving them greater control over their workspace. Finally, Figure 8 displays another teacher interface improvement that allows teachers to examine student and class progress by particular WAT metrics and teacher-created rubrics for each writing task. Collectively, these interface

updates respond directly to instructors’ feedback and aim to make WAT a more flexible, intuitive, and pedagogically aligned tool for classroom use.

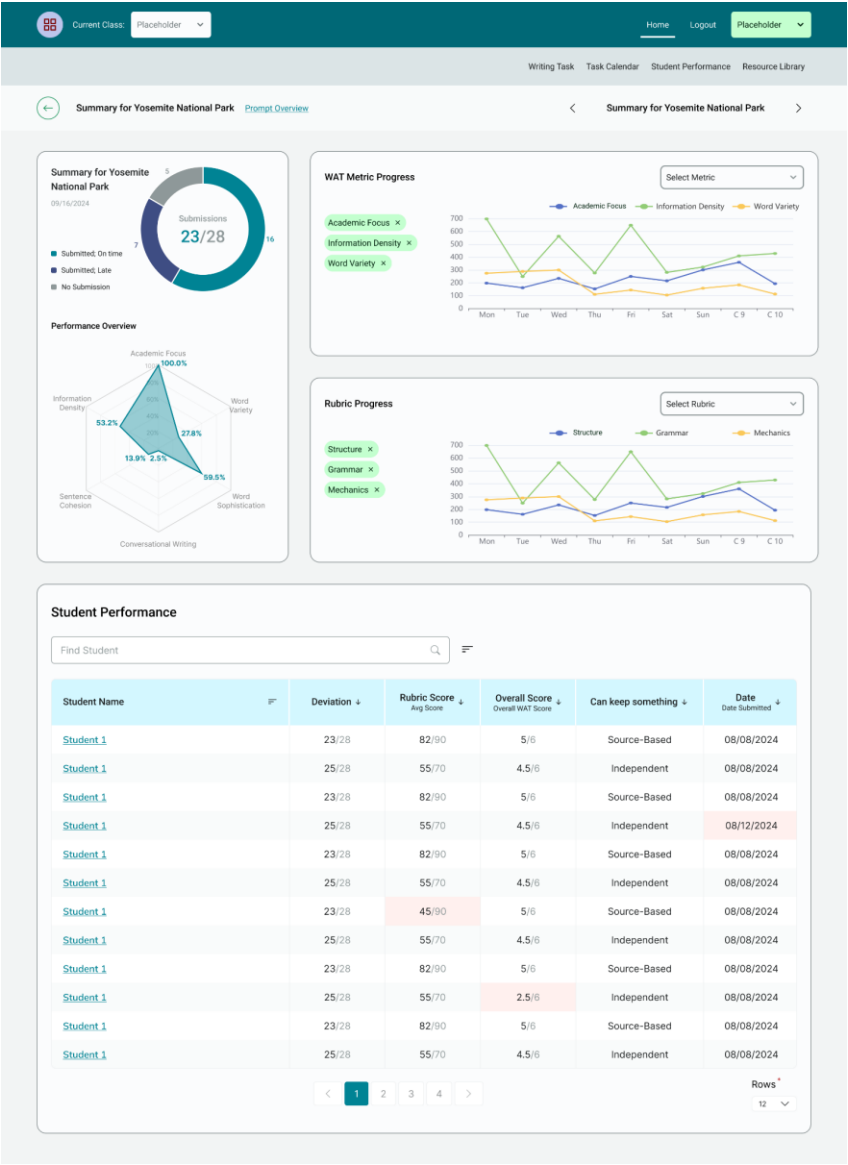


Figure 8: Teacher progress monitoring for class and individual performance per writing task.

7. CONCLUSIONS

This study explored the iterative development of WAT through a participatory design framework and a hybrid AI-human approach to qualitative data analysis. Findings from two focus group sessions with writing instructors revealed four key areas for improvement: the clarity of metrics and feedback, the need for instructor customization, workload concerns, and user interface design. In response, we implemented and proposed a series of design changes, including a simplified feedback interface, customizable features (e.g., metric targets), LMS integration, and a streamlined layout to reduce cognitive load. These revisions aim to make WAT more flexible, intuitive, and pedagogically aligned for classroom use. In doing so, this work advances the

pedagogical use of writing analytics by emphasizing descriptive, interpretable feedback that supports teacher decision-making and student learning.

Importantly, this study contributes a novel methodological approach by demonstrating how LLMs can be used to support inductive qualitative analysis within a participatory research design involving the integration of mixed method data sources. While user-centered design is essential for developing educational technologies that meet real classroom needs, it is often constrained by the high labor demands of analyzing qualitative user feedback. In this study, we employed GPT-4o in a secure, retrieval-augmented environment to assist in thematic coding of focus group transcripts. Model outputs were reviewed, refined, and triangulated by human researchers, which accelerated the early stages of coding while ensuring interpretive rigor through later rounds of human validation. In this sense, the hybrid workflow redistributed analytic labor rather than eliminating it by speeding initial organization and theme surfacing but requiring additional time for contextual verification and consensus building. This hybrid AI-human workflow was embedded within a mixed-methods research design, allowing us to integrate qualitative insights with survey data to inform tool development and design decisions. This work offers a model for how LLMs can augment, and not replace, human-centered qualitative research by increasing scalability, reducing time demands, and supporting more systematic identification of design-relevant themes. This methodological contribution also extends AI research in education beyond its prevailing detection orientation, where AI systems are primarily used to identify, classify, or predict learner characteristics, by demonstrating a generative and interpretive role for AI within human-centered research. In this context, LLMs function not as detectors but as analytic partners that enhance sensemaking, reflection, and design. Moreover, to our knowledge, this is one of the first studies to apply LLMs in this way within a participatory design context for educational tool development.

Several limitations should be noted. First, the current study focused exclusively on instructor perspectives; student voices were not included in this design phase. Future work should investigate how students interpret WAT's feedback and how it affects their revision processes. Second, although LLMs were used to analyze qualitative feedback about the tool, LLM-based tools have not yet been integrated into WAT itself as a feedback generator. Ongoing development will explore ways to embed LLMs directly into the system to enhance real-time feedback capabilities. Finally, while this study demonstrates the utility of LLMs for qualitative analysis, further research is needed to examine the experiences, assumptions, and positionalities of researchers who use LLMs in qualitative data analysis workflows, particularly regarding issues of trust, interpretation, and ethical responsibility.

Beyond its immediate implications for WAT, this study offers a transferable framework for integrating LLMs into participatory design workflows more broadly. By combining structured user engagement with AI-accelerated qualitative analysis, researchers and developers may be able to iterate on educational technologies more rapidly and responsively, without sacrificing methodological rigor or human-centered values. Additionally, this approach highlights how participatory design with educators reveals core pedagogical principles, such as feedback clarity, student agency, and instructor adaptability, that can guide the design of AI-driven writing tools.

DECLARATION OF GENERATIVE AI SOFTWARE TOOLS IN THE WRITING PROCESS

During the preparation of this work, the authors used ChatGPT-4o in the data analysis and results sections in order to demonstrate a human-AI workflow for qualitative data analysis.

After using this tool, the authors reviewed all results and edited the content as needed, and take full responsibility for the content of the publication.

ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A180261 to Arizona State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

REFERENCES

- ABRAS, C., MALONEY-KRICHMAR, D., AND PREECE, J. 2004. User-Centered Design. In *Encyclopedia of Human-Computer Interaction* (Vol. 2 pp. 763-768), W. S. Bainbridge (Ed.), Berkshire Publishing Group.
- AHMED, I., ALVARADO, P., JAIN, S., ARNER, T., REILLEY, E., AND MCNAMARA, D.S. 2025. Arizona State University CreateAI platform. In *Design recommendations for intelligent tutoring systems: Generative AI in intelligent tutoring systems* (Vol. 12), A. M. Sinatra, A. C. Graesser, P. M. Lawton, and V. Rus, Eds. US Army Combat Capabilities Development Command - Soldier Center. <https://eric.ed.gov/?id=ED672991>
- ARIZONA STATE UNIVERSITY. 2024. *Create AI Chat*. <https://ai.asu.edu/technical-foundation/create-ai-chat>
- ASPERS, P., AND CORTE, U. 2019. What is qualitative in qualitative research. *Qualitative sociology*, 42(2), 139-160. <https://doi.org/10.1007/s11133-019-9413-7>
- BÉCHARD, P., AND AYALA, O. M. 2024. Reducing hallucination in structured outputs via Retrieval-Augmented Generation. *arXiv preprint arXiv:2404.08189*.
- BIJKER, R., MERKOURIS, S. S., DOWLING, N. A., AND RODDA, S. N. 2024. ChatGPT for automated qualitative research: Content analysis. *Journal of Medical Internet Research*, 26, e59050. <https://doi.org/10.2196/59050>
- BOURKE, B. 2014. Positionality: Reflecting on the research process. *The Qualitative Report*, 19(33), 1-9. <http://www.nova.edu/ssss/QR/QR19/bourke18.pdf>
- BRAUN, V., AND CLARKE, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101. <https://doi.org/10.1191/1478088706qp063oa>
- BUTTERFUSS, R., ROSCOE, R. D., ALLEN, L. K., MCCARTHY, K. S., AND MCNAMARA, D. S. 2022. Strategy uptake in writing pal: Adaptive feedback and instruction. *Journal of Educational Computing Research*, 60(3), 696-721. <https://doi.org/10.1177/07356331211045304>
- CHEN, J., LOTSOS, A., ZHAO, L., WANG, G., WILENSKY, U., SHERIN, B., AND HORN, M. 2024. Prompts matter: Comparing ML/GAI approaches for generating inductive qualitative coding results. *arXiv*. <https://arxiv.org/abs/2411.06316>
- CHEW, R., BOLLENBACHER, J., WENGER, M., SPEER, J., AND KIM, A. 2023. LLM-assisted content analysis: Using large language models to support deductive coding. *arXiv*. <https://arxiv.org/abs/2306.14924>
- CONIJN, R., MARTINEZ-MALDONADO, R., KNIGHT, S., BUCKINGHAM SHUM, S., WAES, L. V., AND VAN ZAAENEN, M. 2022. How to provide automated feedback on the writing process? A participatory approach to design writing analytics tools, *Computer Assisted Language Learning*, 35(8), 1838-1868, <https://doi.org/10.1080/09588221.2020.1839503>

- CORLETT, S., AND MAVIN, S. 2018. Reflexivity and researcher positionality. In *The SAGE handbook of qualitative business and management research methods*, C. Cassell, A. Cunliffe, and G. Grandy, Eds. 377–389. SAGE.
- CORRENTI, R., MATSUMURA, L. C., WANG, E. L., LITMAN, D., AND ZHANG, H. 2022. Building a validity argument for an automated writing evaluation system (eRevise) as a formative assessment. *Computers and Education Open*, 3, 100084. <https://doi.org/10.1016/j.caeo.2022.100084>
- CORRENTI, R., WANG, E. L., MATSUMURA, L. C., LITMAN, D., LIU, Z., AND LI, T. 2024. Supporting students' text-based evidence use via formative automated writing and revision assessment. In *The Routledge international handbook of automated essay evaluation*, M. D. Shermis and J. Wilson, Eds., 221–243. Routledge. <https://doi.org/10.4324/9781003397618>
- CRESWELL, J. W., AND CLARK, V. L. P. 2017. Designing and conducting mixed methods research. Sage publications.
- CUMBO, B., AND SELWYN, N. 2022. Using participatory design approaches in educational research. *International Journal of Research & Method in Education*, 45, 60-72. <https://doi.org/10.1080/1743727X.2021.1902981>
- DAVIS, F. D. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319-340.
- DIKLI, S., AND BLEYLE, S. 2014. Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing*, 22, 1–17. <https://doi.org/10.1016/j.asw.2014.03.006>
- DUNIVIN, Z. O. 2024. Scalable qualitative coding with LLMs: Chain-of-thought reasoning matches human performance in some hermeneutic tasks. *Center for Complex Networks and Systems Research*, Indiana University. <https://osf.io/k4fg9>
- FIELDING, N. G. 2012. Triangulation and mixed methods designs: Data integration with new research technologies. *Journal of Mixed Methods Research*, 6(2), 124-136. <https://doi.org/10.1177/1558689812437101>
- FLECKENSTEIN, J., LIEBENOW, L., AND MEYER, J. 2023. Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence*, 6. <https://doi.org/10.3389/frai.2023.1162454>
- FOSSEY, E., HARVEY, C., MCDERMOTT, F., AND DAVIDSON, L. 2002. Understanding and evaluating qualitative research. *Australian and New Zealand Journal of Psychiatry*, 36(6), 717-732. <https://doi.org/10.1046/j.1440-1614.2002.01100.x>
- GOLDA, A., SINGH, A., RAJ, M., DEORA, V., NAYAK, G. K., AND POONIA, R. C. 2024. Privacy and security concerns in generative AI: A comprehensive survey. *IEEE Access*, 12, 48126–48144. <https://doi.org/10.1109/ACCESS.2024.3381611>
- GOLDSHTEIN, M., ALHASHIM, A. G., AND ROSCOE, R. D. 2024. Automating bias in writing evaluation: Sources, barriers, and recommendations. In *The Routledge International Handbook of Automated Essay Evaluation*, M. D. Shermis and J. Wilson, Eds., 421-444. Routledge. <https://doi.org/10.4324/9781003397618>
- GOLDSHTEIN, M., OCUMPAUGH, J., POTTER, A., AND ROSCOE, R. D. 2024. The social consequences of language technologies and their underlying language ideologies. In *Universal access in human-computer interaction*. HCII 2024. Lecture Notes in Computer Science (Vol. 14696), M. Antona and C. Stephanidis, Eds., 271-290. https://doi.org/10.1007/978-3-031-60875-9_18
- HATTIE, J., AND TIMPERLEY, H. 2007. The power of feedback. *Review of Educational Research*, 77, 81-112. <https://doi.org/10.3102/0034654302984>

- HAYES, A. S. 2025. “Conversing” With Qualitative Data: Enhancing Qualitative Research Through Large Language Models (LLMs). *International Journal of Qualitative Methods*, 24, 16094069251322346. <https://doi.org/10.1177/16094069251322346>
- HUANG, Y., PALERMO, C., AND WILSON, J. 2025. Identifying active ingredients and uptake patterns in the implementation of an AI-based writing support tool: Insights from a randomized controlled trial. *Computers and Education: Artificial Intelligence*, 6, 100479. <https://doi.org/10.1016/j.caeai.2025.100479>
- IMUNDO, M. N., WATANABE, M., POTTER, A. H., GONG, J., ARNER, T., AND MCNAMARA, D. S. 2024. Expert thinking with generative chatbots. *Journal of Applied Research in Memory and Cognition*, 13(4), 465–484. <https://doi.org/10.1037/mac0000199>
- KATZ, A., FLEMING, G. C., AND MAIN, J. 2024. Thematic analysis with open-source generative AI and machine learning: A new method for inductive qualitative codebook development. *arXiv*. <https://arxiv.org/abs/2410.03721>
- KHALID, M. T., AND WITMER, A. P. 2025. Prompt engineering for large language model-assisted inductive thematic analysis. *arXiv*. <https://arxiv.org/abs/2503.22978>
- KIRSTEN, E., BUCKMANN, A., MHAIDLI, A., AND BECKER, S. 2024. Decoding complexity: Exploring human-AI co-cordance in qualitative coding. *Max Planck Institute for Security and Privacy*. <https://arxiv.org/abs/2403.06607>
- KNIGHT, S., MARTINEZ-MALDONADO, R., GIBSON, A., AND BUCKINGHAM SHUM, S. 2017. Towards mining sequences and dispersion of rhetorical moves in student written texts. In *Proceedings of the Seventh International Learning Analytics and Knowledge Conference* (pp. 228–232). ACM. <https://doi.org/10.1145/3027385.3027433>
- KNIGHT, S., SHIBANI, A., ABEL, S., GIBSON, A., AND RYAN, P. 2020. AcaWriter: A learning analytics tool for formative feedback on academic writing. *Journal of Writing Research*, 12, 141–186. <https://doi.org/10.17239/jowr-2020.12.01.06>
- LI, T., CREER, S. D., ARNER, T., ROSCOE, R. D., ALLEN, L. K., AND MCNAMARA, D. S. 2022. Participatory Design of a Writing Analytics Tool: Teachers’ Needs and Design Solutions. In *Proceedings of the 12th International Conference on Learning Analytics and Knowledge LAK22*, A. F. Wise., R. Martinez-Maldonado, and I. Hilliger, Eds., 15-18. Online.
- LIAQAT, A., MUNTEANU, C., AND DEMMANS E., C. 2021. Collaborating with mature English language learners to combine peer and automated feedback: A user-centered approach to designing writing support. *International Journal of Artificial Intelligence in Education*, 31(4), 638-679. <https://doi.org/10.1007/s40593-020-00204-4>
- LINK, S., DURSUN, A., KARAKAYA, K., AND HEGELHEIMER, V. 2014. Towards best ESL practices for implementing automated writing evaluation. *Calico Journal*, 31(3), 323–344. <https://www.jstor.org/stable/calicojournal.31.3.323>
- LIU, L., XU, W., LI, Y., AND LIU, M. 2017. Automated essay feedback generation and its impact on revision. *IEEE Transactions on Learning Technologies*, 10(4), 502–513. <https://doi.org/10.1109/TLT.2016.2612659>
- LO, L. S. 2023. The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship*, 49(4), 102720. <https://doi.org/10.1016/j.acalib.2023.102720>
- MACARTHUR, C. A. 2016. Instruction in evaluation and revision. In *Handbook of writing research* (2nd ed.), C. A. MacArthur, S. Graham, and J. Fitzgerald, Eds., 272-287. Guilford Press.
- MCNAMARA, D. S., AND KENDEOU, P. 2022. The early automated writing evaluation (eAWE) framework. *Assessment in Education: Principles, Policy & Practice*, 29(2), 150-182. <https://doi.org/10.1080/0969594X.2022.2037509>

- MCNAMARA, D. S., AND POTTER, A. 2024. The two U's in the future of automated essay evaluation: Universal access and user-centered design. In *Handbook of Automated Essay Evaluation* (2nd ed.), M. D. Shermis and J. Wilson, Eds., 590-608. Routledge. <https://doi.org/10.4324/9781003397618>
- MINAEE, S., MIKOLOV, T., NIKZAD, N., CHENAGHLU, M., SOCHER, R., AMATRIAIN, X., AND GAO, J. 2024. Large language models: A survey. arXiv. <https://arxiv.org/abs/2402.06196>
- MULLER, M. J., & KUHN, S. 1993. Participatory design. *Communications of the ACM*, 36(6), 24-28.
- OPENAI. 2024. GPT-4o [LARGE MULTIMODAL MODEL]. <https://openai.com/index/hello-gpt-4o>
- PALERMO, C., AND THOMSON, M. M. 2018. Teacher implementation of self-regulated strategy development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students. *Contemporary Educational Psychology*, 54, 255-270. <https://doi.org/10.1016/j.cedpsych.2018.07.002>
- POTTER, A., AND WILSON, J. 2022. Statewide implementation of automated writing evaluation: analyzing usage and associations with state test performance in grades 4-11. *Educational Technology Research and Development*, 69(3), 1557-1578. <https://doi.org/10.1007/s11423-021-10004-9>
- POTTER, A., WILSON, J., ROSCOE, R.D., ARNER, T., AND MCNAMARA, D.S. 2025. Computer-based writing instruction. In *Handbook of writing research* (3rd ed.) C. A. MacArthur, S. Graham, and J. Fitzgerald, Eds., 255-270. Guilford Press.
- PRESCOTT, M. R., YEAGER, S., HAM, L., RIVERA SALDANA, C. D., SERRANO, V., NAREZ, J., AND MONTOYA, J. 2024. Comparing the efficacy and efficiency of human and generative AI: Qualitative thematic analyses. *JMIR AI*, 3, e54482. <https://doi.org/10.2196/54482>
- ROSCOE, R. D., CRAIG, S. D., AND DOUGLAS, I. 2018. *End-user considerations in educational technology design*. IGI Global Scientific Publishing. <https://doi.org/10.4018/978-1-5225-2639-1>
- ROSCOE, R. D., WILSON, J., JOHNSON, A. C., AND MAYRA, C. R. 2017. Presentation, expectations, and experience: Sources of student perceptions of automated writing evaluation. *Computers in Human Behavior*, 70, 207-221. <https://doi.org/10.1016/j.chb.2016.12.076>
- SALDAÑA, J. 2014. Coding and analysis strategies. In P. Leavy (Ed.), *The Oxford handbook of Qualitative Research*, 581-605. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199811755.013.001>
- SHERMIS, M. D., AND WILSON, J. 2024. Introduction to automated essay evaluation. In *The Routledge international handbook of automated essay evaluation*, M. D. Shermis and J. Wilson, Eds., 3-22. Routledge. <https://doi.org/10.4324/9781003397618>
- STEEN, M. 2011. Tensions in human-centered design. *CoDesign*, 7, 45-60. <https://doi.org/10.1080/15710882.2011.563314>
- STONE, M. L., KENT, K. M., ROSCOE, R. D., CORLEY, K. M., ALLEN, L. K., AND MCNAMARA, D. S. 2018. The design implementation framework: Iterative design from the lab to the classroom. In *End-user considerations in educational technology design*, R. D. Roscoe, S. D. Craig, and I. Douglas, Eds., 76-98. Hershey, PA: IGI Global. <https://doi.org/10.4018/978-1-5225-2639-1.ch004>
- STROBL, C., AILHAUD, E., BENETOS, K., DEVITT, A., KRUSE, O., PROSKE, A., AND RAPP, C. 2019. Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education*, 131, 33-48. <https://doi.org/10.1016/j.compedu.2018.12.005>
- TEN HOLTER, C. 2022. Participatory design: lessons and directions for responsible research and innovation. *Journal of Responsible Innovation*, 9(2), 275-290. <https://doi.org/10.1080/23299460.2022.2041801>

- THEELEN, H., VREULS, J., AND RUTTEN, J. 2024. Doing research with help from ChatGPT: Promising examples for coding and inter-rater reliability. *International Journal of Technology in Education*, 7, 1–18. <https://doi.org/10.46328/ijte.537>
- THOMAS, D. R. 2006. A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, 27(2), 237–246. <https://doi.org/10.1177/1098214005283748>
- TOLEDO, C., AND SHANNON-BAKER, P. 2023. Choosing a qualitatively oriented mixed methods research approach: Recommendations for researchers. In *Handbook of mixed methods research in business and management*, R. Cameron and X. Golenko, Eds., 41–54. Edward Elgar Publishing. <https://doi.org/10.4337/9781800887954.00011>
- TUHKALA, A. 2021. A systematic literature review of participatory design studies involving teachers. *European Journal of Education*, 56(4), 641–659. <https://doi.org/10.1111/ejed.12471>
- TUROBOV, A., COYLE, D., AND HARDING, V. 2024. Using ChatGPT for thematic analysis. *arXiv*. <https://arxiv.org/abs/2405.08828>
- VENKATESH, V., AND DAVIS, F. D. 2000. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, 46(2), 186–204. <https://doi.org/10.1287/mnsc.46.2.186.11926>
- VERMA, R. K., GUPTA, S., AND ILLINICH, S. 2024. Technology-enhanced personalized learning in higher education. In *Advances in technological innovations in higher education* (1st ed.), A. Garg, B. V. Babu, & V. E. Balas, Eds., 71–92. CRC Press. <https://doi.org/10.1201/9781003376699>
- WACNIK, P., DALY, S. R., AND VERMA, A. 2025. Participatory design: a systematic review and insights for future practice. *Design Science*, 11, e21. doi:10.1017/dsj.2025.10009
- WANG, E. L., MATSUMURA, L. C., CORRENTI, R., LITMAN, D., ZHANG, H., HOWE, E., MAGOODA, A., AND QUINTANA, R. 2020. eRevis(ing): Students’ revision of text evidence use in an automated writing evaluation system. *Assessing Writing*, 44, 100449. <https://doi.org/10.1016/j.asw.2020.100449>
- WARR, M., AND HEATH, M. K. (2025). Uncovering the Hidden Curriculum in Generative AI: A Reflective Technology Audit for Teacher Educators. *Journal of Teacher Education*, 76(3), 245–261. <https://doi.org/10.1177/00224871251325073>
- WEI, J., WANG, X., SCHUURMANS, D., BOSMA, M., XIA, F., CHI, E., LE, Q.V., AND ZHOU, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
- WILSON, J., DELGADO, A., PALERMO, C., CRUZ CORDERO, T. M., MYERS, M. C., EACKER, H., POTTER, A., COLES, J., AND ZHANG, S. 2024. Middle school teachers’ implementation and perceptions of automated writing evaluation. *Computers and Education Open*, 7, 100231. <https://doi.org/10.1016/j.cao.2024.100231>
- WILSON, J., AND MACARTHUR, C. 2024. Exploring the role of automated writing evaluation as a formative assessment tool supporting self-regulated learning and writing. In *Routledge international handbook of automated essay evaluation*, M. D. Shermis and J. Wilson, Eds., 197–220. Routledge. <https://doi.org/10.4324/9781003397618>
- WILSON, J., AND ROSCOE, R. D. 2020. Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, 58, 87–125. <https://doi.org/10.1177/0735633119830764>
- WILSON, J., ZHANG, S., PALERMO, C., CRUZ CORDERO, T., ZHANG, F., MYERS, M. C., POTTER, A., EACKER, H., AND COLES, J. 2024. A Latent Dirichlet Allocation approach to

- understanding students' perceptions of automated writing evaluation. *Computers and Education Open*, 6, 100194. <https://doi.org/10.1016/j.caeo.2024.100194>
- WISNIEWSKI, B., ZIERER, K., AND HATTIE, J. 2020. The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, 487662. <https://doi.org/10.3389/fpsyg.2019.03087>
- XIAO, Z., YUAN, X., LIAO, Q. V., ABDELGHANI, R., AND OUDEYER, P.-Y. 2023. Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23 Companion)* (pp. 1–6). ACM. <https://doi.org/10.1145/3581754.3584136>
- XU, Z., JAIN, S., AND KANKANHALLI, M. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv*. <https://arxiv.org/abs/2401.11817>
- YANG, Y., ALBA, C., WANG, C., WANG, X., ANDERSON, J., AND AN, R. (2024). GPT models can perform thematic analysis in public health studies, akin to qualitative researchers. *Journal of Social Computing*, 5(4), 293–312. <https://doi.org/10.23919/JSC.2024.0024>
- ZAMBRANO, A. F., LIU, X., BARANY, A., BAKER, R. S., KIM, J., AND NASIAR, N. 2023. From nCoder to ChatGPT: From automated coding to refining human coding. In *Advances in quantitative ethnography: ICQE 2023* (Vol. 1895), G. Arastoopour Irgens and S. Knight, Eds., 470–485. Springer. https://doi.org/10.1007/978-3-031-47014-1_32
- ZHANG, H., WU, C., XIE, J., KIM, C., AND CARROLL, J. M. 2023. QualiGPT: GPT as an easy-to-use tool for qualitative coding. *arXiv*. <https://arxiv.org/abs/2310.07061>
- ZHANG, H., WU, C., XIE, J., LYU, Y., CAI, J., AND CARROLL, J. M. 2023. Redefining qualitative analysis in the AI era: Utilizing ChatGPT for efficient thematic analysis. *arXiv*. <https://arxiv.org/abs/2309.10771>
- ZHANG, H., WU, C., XIE, J., RUBINO, F., GRAVER, S., KIM, C., AND CAI, J. 2024. When qualitative research meets large language model: Exploring the potential of QualiGPT as a tool for qualitative coding. *arXiv*. <https://arxiv.org/abs/2407.14925>
- ZHAO, F., YU, F., AND SHANG, Y. 2024. A new method supporting qualitative data analysis through prompt generation for inductive coding. In *Proceedings of the 2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)* (pp. 164–169). IEEE. <https://doi.org/10.1109/IRI62200.2024.00043>