Evaluating Generative AI as a Supportive Analytic Partner in Qualitative Coding of Metacognitive Student Reflections

Rebecca Marrone
Adelaide University
Adelaide, Australia
Rebecca.Marrone@adelaide.edu.au

Abhinava Barthakur
Adelaide University
Adelaide, Australia
Abhinava.Barthakur@adelaide.edu.au

David Randall
Adelaide University
Adelaide, Australia
David.Randall@adelaide.edu.au

Nazanin Mottaghi
Adelaide University
Adelaide, Australia
Nazanin.Mottaghi@adelaide.edu.au

Vitomir Kovanovic
Adelaide University
Adelaide, Australia
Vitomir.Kovanovic@adelaide.edu.au

Maarten De Laat
Adelaide University
Adelaide, Australia
Maarten.DeLaat@adelaide.edu.au

Qualitative data provides rich insight into student and educator thinking but remains difficult to analyse systematically at scale. The rise of generative artificial intelligence (GenAI) introduces new opportunities for pattern recognition and interpretive support, while also raising questions about how such systems can be responsibly embedded in educational research workflows. This study investigates the application of a fine-tuned GenAI model to classify metacognitive elements in student self-reflections and examines the methodological and epistemological implications of this process.

In partnership with a secondary school, a university, and a state education department, the study analysed more than 14,000 student reflection artefacts collected between 2018 and 2023. A total of 4,631 samples were manually coded for four sub-elements of metacognition—Goal Setting, Strategy Choice, Reflection on Learning, and Effort Regulation—which were then used to train and evaluate a fine-tuned GPT-4o-mini model that achieved 80.98% classification accuracy. However, our analysis also raises critical questions. While the model could detect the existence of metacognitive constructs, it lacked the pedagogical and contextual grounding to assess their quality or relevance and required significant human effort. These findings highlight the need to reconceptualise GenAI not as a replacement for human judgement, but as a supportive analytic partner. We argue that co-design processes between educators, researchers, and developers are essential to ensure AI systems are trustworthy, theoretically grounded, and practically useful. The approach outlined in this study provides a roadmap for extending the use of GenAI to other complex educational constructs, ensuring that AI is framed not merely as a tool but as a supportive analytic partner whose design reflects the needs and values of all system actors.

Keywords: metacognition, human-in-the-loop, artificial intelligence, qualitative research, K-12 education

1. INTRODUCTION

Metacognition is understood as the awareness and regulation of one's own cognitive processes (Flavell, 1979). It is widely recognised as a cornerstone of self-regulated learning (Winne & Hadwin, 1998) and a critical determinant of academic and lifelong learning success (Zimmerman, 2002). However, capturing metacognitive thinking remains a persistent challenge (Azevedo, 2020). Traditional tools such as self-report questionnaires, while scalable, are limited in their capacity to capture the nuanced and context-specific ways in which students engage in metacognitive thinking (Pintrich et al., 2000; Veenman & Spaans, 2005). In response, researchers have increasingly turned to qualitative data, particularly reflective writing, as a means of accessing students' cognitive processes (Ozturk, 2017; Schraw, 2000). Qualitative data, particularly when drawn from naturally occurring classroom artefacts, offers educators and researchers a powerful lens into how students articulate, regulate, and adapt their learning (Lichtman, 2023). Nevertheless, the process of analysing such data poses persistent challenges, especially ensuring consistency, transparency, and scalability of coding decisions (Belotto, 2018). Coding requires iterative interpretation, which is often grounded in theoretically derived frameworks, and typically relies on extensive human collaboration to establish shared meaning, interpretation and reliability (Gisev et al., 2013; Harwood & Garry, 2003; Strauss & Corbin, 1998).

The emergence of generative artificial intelligence (GenAI) presents new possibilities in data analysis, as pattern recognition and analysis can be conducted in a fraction of the time it takes a human (Garg et al., 2024; Prescott et al., 2024). The capabilities of GenAI are being explored in educational research, particularly in the context of automating aspects of qualitative coding (see Barany et al., 2024; Zambrano et al., 2023). However, much of this work has focused on testing the technical performance of models across datasets or comparing model outputs to human-coded benchmarks (Parker et al., 2023). What remains under-examined are the methodological and epistemological implications of integrating GenAI into qualitative research workflows, especially when the aim is not only to replicate human coders but to interrogate how human and AI interpretations align, diverge, and co-evolve through iterative collaboration (Agbon, 2024; Schroeder et al., 2025). Whilst some researchers acknowledge these limitations (Agbon, 2024; Li et al., 2025), there remains a pressing need for continued investigation within K–12 education, where qualitative data is essential for understanding and advancing student learning and development, and where constructs such as metacognition play a pivotal role in shaping long-term academic outcomes.

This paper presents a case study conducted in partnership with a secondary school, a university, and a state department of education to examine the role of GenAI in qualitative research, not merely as a tool, but as a component in a supportive human–AI partnership that shaped the analytic process. We use the term "partnership" deliberately to avoid implying symmetrical collaboration or shared agency. Our conception aligns more closely with recent work on human–AI teaming in interpretive research workflows (e.g., Schroeder et al., 2025), where AI systems surface patterns that inform, but do not replace, human meaning-making. We explore how GenAI can function as an interpretive actor within educational systems, shaping how complex constructs like metacognition are identified, understood, and acted upon in real-world contexts. We examined the use of a fine-tuned GenAI model to analyse over 14,000 student-generated comments drawn from six years (2018-2023) of classroom reflections. While GenAI presents new opportunities for analysing large-scale qualitative data, its integration into

educational workflows must be approached with caution. Without deliberate alignment to theory and ongoing human oversight, GenAI risks becoming a black box: efficient, but epistemically unaccountable, and potentially unethical (Li et al., 2025). This concern is especially relevant when working with complex constructs such as metacognition, where surface-level pattern recognition may obscure deeper interpretive meaning. To address this, our approach prioritises methodological transparency, sustained educator involvement, and theory-driven annotation practices to ensure the system supports, rather than distorts, educational decision-making.

While our primary focus was on identifying metacognitive elements (such as goal-setting, strategy use, and reflection of learning), we approached the use of GenAI not simply as a tool to replicate human coders but as a supportive analytic partner in the interpretive process. Our human coders coded data in a theoretically grounded manner and engaged in iterative negotiation to refine shared meaning. In contrast, the GenAI model learned patterns from this human-labelled data, but without access to the theoretical commitments and contextual judgements that shape human coding practices.

As such, this paper interrogates how methodological decisions were made, how training data were constructed, and where tensions emerged between human and AI interpretations. We do not aim to provide a best-practice blueprint for structuring an ideal human–AI partnership or conduct a comparative performance analysis. Instead we problematise the common "human-in-the-loop" framing (Siemens et al., 2022) and propose an alternative view: that AI should be treated as a supportive analytic partner whose inclusion requires careful epistemic positioning, rather than simply as an assistant that mimics or speeds up human work. We suggest that introducing AI into the qualitative coding process should be approached in a manner similar to inducting a new human coder, through shared dialogue, theory alignment, and critical reflection on the limits of both human and machine sense-making. In qualitative coding, agreement between human coders is typically achieved through theory-informed discussion, reflexive judgement, and negotiation of meaning. When discrepancies arise, coders examine assumptions, clarify construct boundaries, and revisit theoretical definitions in light of contextual evidence (Saldaña, 2015). These deliberations are central to qualitative validity, as they make explicit how meaning is constructed and justified. GenAI systems cannot participate in these processes in the same way, as they do not reason with theory or exercise reflexive judgement. However, they can contribute in adjacent ways. By consistently applying learned patterns at scale, GenAI can surface recurrent ambiguities, boundary cases, and systematic disagreements that may prompt further human discussion. In this sense, while the negotiation of meaning remains a human responsibility, AI outputs can act as inputs to that negotiation by revealing where constructs are unstable, underspecified, or inconsistently applied.

In doing so, we propose the following research questions to guide our exploration:

1. In what ways can GenAI augment, rather than replace, human judgement in analysing complex constructs such as metacognition, and what design principles support this collaboration?
2. What design choices, training processes, and performance tradeoffs were involved in fine-tuning a GenAI model to classify metacognitive elements in student reflections?

2. BACKGROUND

2.1. METACOGNITION

Metacognition, first introduced by Flavell (1979) refers to an individual's ability to reflect on, monitor, and regulate their own thinking. It is widely recognised as a critical component of self-regulated learning (SRL), functioning as the mechanism through which learners plan, monitor, and adjust their learning strategies to achieve academic goals (Azevedo, 2020; Winne, 2017). Through iterative cycles of reflection and refinement, skilled learners continuously evaluate their understanding, identify knowledge gaps, and adopt new strategies to enhance their performance (Schunk & Greene, 2017). Importantly, the effectiveness of metacognition hinges on the learner's capacity to accurately assess their cognitive processes and exercise control over those processes in real time. These capabilities are enacted through metacognitive strategies, deliberate behaviours that support planning, monitoring, and evaluating one's learning. In this sense, metacognition is not merely an internal state but a dynamic process that becomes visible through purposeful actions and reflections (Azevedo, 2020; Flavell, 1979). Given its central role in supporting academic resilience and long-term learning success, there is interest in how K-12 learners actively engage with metacognitive processes and how educators can support the development of these skills (Ohtani & Hisasaka, 2018). Nevertheless, assessing metacognition in authentic classroom settings remains a complex task (Gascoine et al., 2017).

2.2. CHALLENGES IN ASSESSING METACOGNITION

Despite strong theoretical and empirical support for metacognition as a key driver of learning, its assessment presents significant methodological challenges (Azevedo, 2020). As an internal cognitive process, metacognition is not always readily observable. Teachers and researchers must rely on indirect indicators such as student behaviour, self-report measures, or written reflections, to infer the presence and quality of metacognitive thinking (Chen & McDunn, 2022).

Reflective writing has emerged as a valuable window into students' metacognitive processes (Ismail & Tawalbeh, 2014; Zsigmond et al., 2025). In educational contexts, reflection and metacognition are deeply intertwined as both involve stepping back from the task at hand to examine one's own learning, behaviours, and decisions. Reflection enables learners to critically evaluate their strategies, confront misconceptions, and articulate how and why they learn, a process closely aligned with metacognition (Silver, 2023). Studies have shown that metacognitive elements can be reliably identified within student reflections, particularly when writing is scaffolded with explicit prompts or frameworks (Alt & Raichel, 2020; Ramadhanti et al., 2020).

Nevertheless, analysing student reflections is a labour-intensive and often ambiguous process (Kovanović et al., 2018). Students may be unaccustomed to articulating their thought processes in writing, or they may use language that is vague, inconsistent, or contextually dependent (Kovanović et al., 2018). As a result, identifying specific metacognitive elements, such as strategy selection, monitoring, or effort regulation, requires a deep understanding of the context and interpretive judgment. This makes consistent coding across students and contexts difficult, especially when dealing with large datasets.

The practical demands of analysing metacognitive processes also place a considerable burden on educators. Existing research indicates that such assessments are typically carried out by trained researchers rather than classroom teachers, or, where teachers are involved, they often require specialised training in metacognition to conduct the analysis effectively (Gascoine et al.,

2017). Teachers, already facing high workloads and administrative responsibilities, often lack the time and resources to engage deeply with open-ended, qualitative student work (Creagh et al., 2025). Unlike standardised assessments, which can be graded using pre-determined rubrics, reflections require attention to nuance, tone, and context. These are attributes that do not lend themselves easily to automated scoring or quick interpretation.

In response to these challenges, researchers have begun exploring computational approaches to assist with the analysis of reflective writing (Barthakur et al., 2022; Ullmann, 2019). Automated methods offer the potential to consistently apply coding criteria across large datasets and can augment, rather than replace, human judgment (Gibson et al., 2016; Kovanović et al., 2018). With the emergence of large language models (LLMs), the field now faces new possibilities (and questions) about how GenAI might support the identification of complex constructs such as metacognition in student reflections.

2.3. Qualitative Coding Using LLMs

The rise of GenAI, particularly LLMs has introduced new possibilities for augmenting qualitative research in education. These models, trained on vast amounts of data, demonstrate an unprecedented capacity to detect patterns, infer meaning from unstructured data, and adapt outputs based on contextual cues. Recent studies have demonstrated that LLMs can streamline various research processes, including formative feedback generation, essay grading, the classification of affective states or learning strategies, and thematic analysis (e.g., Barany et al., 2024; Nguyen-Trung, 2025; Prescott et al., 2024). Specific to qualitative approaches, Katz et al. (2023) show that GenAI can assist with tasks such as labelling and collaborative coding. Building on this work, Schroeder et al. (2025) conducted interviews with 20 qualitative researchers to explore their views on the use of AI in qualitative research. While participants identified promising applications of AI across various stages of the research process, they expressed concerns about safeguarding participant interests, ensuring the performance and reliability of AI tools, and assessing the appropriateness of their use. Notably, the researchers highlighted a significant gap in existing norms and guidelines for the ethical use of LLMs in qualitative research, and the authors offer a comprehensive set of recommendations for future work in addressing these ethical tensions.

These concerns are particularly salient as GenAI techniques continue to evolve. In qualitative research, some common AI approaches include zero shot learning, few-shot learning, and chain of thought reasoning. Zero-shot learning involves prompting a model to perform a task with no prior examples in the prompt; it relies entirely on the model's pre-trained knowledge to generate a response (Wei et al., 2023). In comparison, few-shot learning provides a small number of annotated examples to help the model learn the task within the prompt itself (Brown et al., 2020; Zhao et al., 2021). Chain-of-thought (CoT) prompting encourages the model to explain its reasoning step by step (White et al., 2023). All of these have benefits and can be enhanced through optimisation techniques. However, implementing these techniques requires significant technical expertise, creating barriers for many educational researchers and practitioners (Zamfirescu-Pereira et al., 2023). At the same time, researchers such as Zhang et al. (2025) argue that current work tends to prioritise model accuracy and coherence over explainability and transparency. The lack of explainability and transparency is particularly problematic in education, where trust and accountability are central to practice, especially in K–12 contexts. While LLMs excel at approximating human-like output, they lack the grounding in educational theory, classroom context, and pedagogical intent that shapes how teachers and researchers interpret student work (Agbon, 2024; Li et al., 2025). As a result, their classifications may reflect

surface-level statistical patterns rather than deeper conceptual understanding, especially for complex constructs like metacognition.

One of the most important gaps in the current literature on AI integration in education is not just how we conceptualise the role of AI or the researchers, but how we view study participants. Students and educators are often treated primarily as data sources, and while research ethics rightly emphasise the protection of their privacy and autonomy, far less attention is given to establishing a reciprocal relationship (Holstein & Aleven, 2022). In many cases, little effort is made to ensure that findings are returned in ways that can inform practice or contribute to meaningful change within the communities from which the data was drawn.

In qualitative research, human analysts play a central interpretive role that extends beyond pattern identification. Human coding involves theoretically informed judgement, reflexive reasoning, and dialogue between coders to negotiate meaning, resolve ambiguity, and refine construct boundaries. These processes are inherently situated, relying on disciplinary knowledge, contextual awareness, and epistemic accountability (Saldaña, 2015, Miles et al. 2014). GenAI systems cannot perform these functions. They do not reason with theory, engage in reflexive dialogue, or adjudicate conceptual disagreements. Instead, their contribution lies in detecting statistical regularities at scale, surfacing boundary cases, inconsistencies, and distributional patterns that may not be readily visible in, human-only analyses. Making these distinctions explicit is essential for positioning AI appropriately within qualitative research workflows and avoiding inflated claims about interpretive capacity.

To summarise, current approaches to support qualitative analysis are most evident in deductive analysis (Siiman et al., 2023). However, qualitative research is also inductive and is fundamentally an interpretive act. It involves the systematic categorisation of textual data into conceptual units and codes that are often developed from, or closely aligned with, theoretical frameworks (Strauss & Corbin, 1998; Saldaña, 2015). In the context of educational research, this process is typically guided by a priori constructs such as metacognitive strategies, engagement behaviours, or self-regulatory phases, allowing researchers to observe how these constructs manifest in students' language. To use AI responsibly in qualitative research, then, we must ask not only whether it "works," but *how* it works, *why* it works in particular ways, and *when* its outputs are trustworthy. Doing so requires adapting the role of the AI beyond enhancing accuracy to supporting human interpretive work in ways that are transparent, interrogable, and epistemically constrained. It also necessitates new forms of documentation and reflexivity: how were training datasets constructed? How was theory operationalised in the annotation process? What assumptions are embedded in the model's outputs, and how do these shape the conclusions drawn?

In the sections that follow, we present a case study exploring these questions in depth. We describe how our team developed and fine-tuned a GPT-4o-mini model to identify metacognitive elements in student reflections, and we reflect on the methodological choices, interpretive dilemmas, and collaborative dynamics that emerged throughout the process. Our goal is not to prescribe 'best practice', but to surface the kinds of epistemic, pedagogical, and ethical considerations that should accompany the integration of GenAI into qualitative educational research.

3. METHODS

3.1. STUDY CONTEXT

The dataset comprised 14,100 individual reflections written by students aged 15–18 years attending a metropolitan secondary school in South Australia, collected between 2018 and 2023. Table 1 provides the demographics of the students from 2019-2023. Note that demographic data was not publicly available for 2018. These reflections capture students' personal insights and responses to their learning across a range of subjects. Prompts vary, but commonly include reflections on standardised tests (e.g., PAT testing), personal learning goals, and end-of-term evaluations. For example, a student studying mathematics might be asked to reflect on their goals for the subject, their performance on recent assessments, or their progress over the semester. Similarly, students enrolled in biology might reflect on a specific assignment, their understanding of a core topic, or their aspirations for improvement. Students in the senior years choose many of their subjects, which results in highly individualised reflections spanning a wide range of disciplines—including English, history, social sciences, and the physical sciences. In addition to subject-specific reflections, students are also prompted to write about broader learning experiences, such as their future goals or progress across multiple subjects.

Initial screening revealed several data quality issues requiring preprocessing, including substantial missing values across multiple fields, 38 exact duplicate rows, HTML entity encoding, corrupted Unicode characters and placeholder entries where students did not provide authentic comments. Data was preprocessed using custom Python scripts to address these issues: to ensure data quality for model training, sentences shorter than five words were excluded as they typically lacked sufficient context for reliable classification, and duplicate text appearing across multiple reflections was filtered out to prevent data leakage.

Reflections varied in length, with an average of 11.17 sentences per entry (SD = 4.66) and a mean character count of 1,445.86 per reflection (SD = 577.56). The dataset constitutes a rich source of naturalistic evidence, reflecting authentic student engagement with learning materials over time. Students at the school are familiar with reflective writing and are introduced to its importance upon entry. This practice is underscored by the inclusion of metacognitive reflection forms as a fundamental component of the school's strategic plan. Reflections are typically written ad hoc, either spontaneously or in response to teacher prompts. In some instances, students reflect on their overall learning journey (e.g., for report card comments), while in others, they are asked to reflect on specific learning tasks or assessment topics. Given the nature of the reflections, we were provided with deidentified data and therefore cannot determine how many reflections each individual student contributed to the dataset. However, given the number of students enrolled each year in comparison to the total number of reflections, it is clear that each student contributes a significant number of reflections per year and is therefore very familiar with the notion of reflective writing.

Table 1: School demographics from 2019 to 2023.

Year	#Students enroled	Gender	English as first language
2019	373	212 Boys 161 Girls	69%

Year	#Students enroled	Gender	English as first language
2020	365	214 boys 151 girls	65%
2021	366	206 boys 160 girls	67%
2022	388	227 boys 161 girls	67%
2023	401	227 boys 174 girls	71%

3.2. DESIGN CHOICES

The design and development of this study followed a nine-stage process between May and November 2024 and took approximately 744 hours. This iterative workflow combined educational research methods with data science practices, drawing on expertise from researchers, educators, and state department employees. Table 2 outlines the full sequence of stages and associated team responsibilities. The research team comprised the first author and two subject matter experts in metacognition. The educator team included three school leaders, namely, the principal and two assistant principals from the participating secondary school. The data science team consisted of one data scientist with expertise in artificial intelligence development. Finally, the state education department team included two staff members from the Strategy and Measurement Division.

Table 2: The different design stages of the project.

Stage	Action Step	Team involved	Hours Spent
Stage 1	Review the data provided by the school	Research team, educators, education department staff	2
Stage 2	Generate coding book	Research team (2 subject matter experts)	2
Stage 3	Sense check coding book	Educators	2
Stage 4	Complete coding	Research team	150
Stage 5	Build the algorithm	Data science team (data scientist)	300
Stage 6	Test the algorithm	Data science team	250
Stage 7	Check the quality of the output	Educators and the research team	3
Stage 8	Explore next steps: how do we make this useful for a teacher?	Educators, education department staff Research team	10
Stage 9	Finalise outputs	Data science team	25

To ensure the AI model was developed in a theoretically grounded and trustworthy way, we adopted an iterative, collaborative workflow involving researchers, educators, and the AI system. Figure 1 below outlines this process across Stages 2 to 6, highlighting the continuous refinement of both the codebook and the model prompts. Key activities included collaborative coding, educator sense-checking, independent double coding with inter-rater reliability checks, and prompt adjustments informed by model outputs. Notably, Stage 4 and Stage 5 were not linear steps but were repeated multiple times as human and AI inputs informed one another. Following each round of model training, samples of AI classifications were reviewed by the research team and educators, with particular attention paid to misclassified or ambiguous cases. These reviews prompted discussion, refinement of construct definitions, and revisions to the codebook, which in turn informed subsequent rounds of human coding and model retraining. This figure illustrates how human judgement and AI outputs were brought into dialogue to enhance the accuracy, interpretability, and contextual relevance of the model.

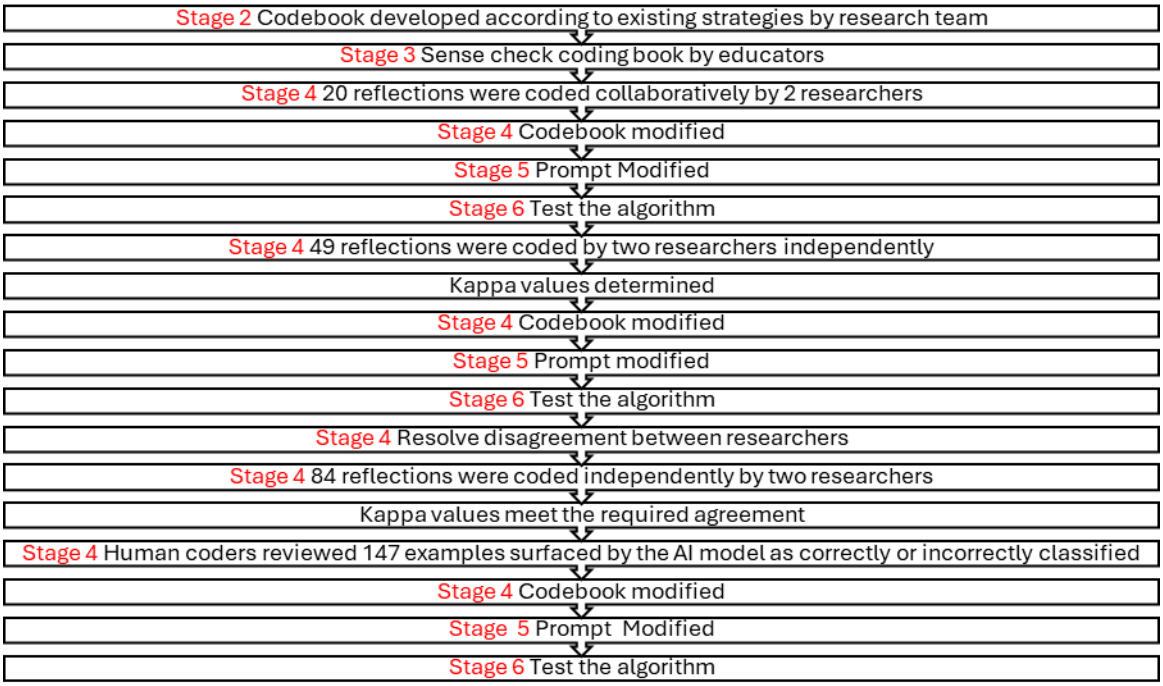

Figure 1: Coding, prompt design, and human–AI partnership workflow. This figure illustrates the iterative process undertaken across stages 2 to 6. The bidirectional refinement process from stages 4, 5 and 6 demonstrates how human expertise and AI outputs informed one another over multiple cycles during model development.

The project commenced with the three groups (educator team, education department, and research team) reviewing the available dataset of student reflections and clarifying how the concept of metacognition was understood and operationalised within the school context.

Following this meeting, the research team began drafting a codebook based on theoretical frameworks derived from the self-regulated learning literature. Drawing on the work of Zimmerman (2002) and Pintrich et al. (2000), the team identified four key sub-elements of metacognition to guide the analysis. Each is well-supported in the literature and represents a critical component of metacognitive awareness and self-regulated learning:

- **Effort Regulation**: The ability to persist with a task despite difficulty, boredom, or low motivation. This construct plays a key motivational role in maintaining attention and perseverance.
- **Goal Setting**: The process of establishing specific, measurable learning objectives that guide monitoring and adjustment of learning behaviours. It is a foundational element of self-regulated learning.
- **Reflection on Learning**: The evaluation of one's learning experiences, strategies, and outcomes in order to inform future actions and improve performance.
- **Strategy Choice**: The deliberate selection and application of learning strategies, such as time management, planning, or help-seeking. These behaviours signal active regulation of the learning process.

The data analysis began with the research team collaboratively coding 20 reflections. The coders worked at the sentence level within each reflection, using character-position-based text selection to precisely mark spans that demonstrated metacognitive thinking. As illustrated in Table 3, multiple labels could be applied to the same sentence when reflections demonstrated overlapping metacognitive constructs—an important capability for capturing the complex, multifaceted nature of authentic student metacognitive thinking.

After the 20 reflections, the two subject matter experts from the research team independently coded an additional 49 reflections (50 lines, one missing), after which inter-rater reliability was assessed. Kappa values from the initial round of coding (49 reflections) indicated low to moderate agreement across constructs, suggesting a need for refinement of the coding framework. Agreement was highest for Strategy Choice ($\kappa = 0.51$), followed by Effort Regulation ($\kappa = 0.44$), Goal Setting ($\kappa = 0.35$), and Reflection on Learning ($\kappa = 0.32$). In response, the research team held a calibration meeting to discuss discrepancies, resolve disagreements and clarify coding definitions.

A second round of independent coding was then undertaken using 84 reflections (100 lines, 16 missing). Additional coding resulted in improved agreement across all categories. Perfect agreement was achieved for Goal Setting ($\kappa = 1.00$), while Effort Regulation increased to $\kappa = 0.57$, Strategy Choice to $\kappa = 0.58$, and Reflection on Learning to $\kappa = 0.52$.

The remaining inconsistencies were primarily due to ambiguity in the students' phrasing. To address this, the research team further refined the codebook that included key terms, example sentences, and decision rules. One major decision was to code Strategy Choice only when an explicit learning strategy was mentioned. Inferred strategies were excluded to ensure transparency and replicability.

Following this refinement, the research team reached full agreement across all four constructs in the third round. This agreement was critical for ensuring the theoretical consistency and reliability of the training data. The research team then presented the updated codebook to the education department staff and the education team for validation. During this process, the label "Effort Regulation" was renamed "Sustained Effort" based on educator feedback about clarity and accessibility. No further changes were required. For the sake of this paper, we have opted to default to typical metacognitive language and use the term effort regulation.

Table 3 is an example of the subset of our code book for strategy choice. Appendix A provides a more comprehensive list of our codebook, along with examples. Where students explicitly mentioned school-specific factors, we have omitted them from the final list due to concerns about privacy and ethics.

Table 3: Examples of students' reflections coded as strategy choice.

Student Reflection	Strategy choice present (yes or no)
After analysing my results, I have identified that I need more practice with algebraic equations. I plan to do some work in my spare time to make sure I am up to where I need to be.	This reflection has goal setting, reflection on learning and strategy choice.
I plan to do this by trying to read at least 5 books over the semester as reading is an awesome way to improve these areas.	This reflection involves goal setting and strategy choice.
Through my continued hard work my grades have stayed fairly steady throughout the semester and this can be pinpointed to strategising my homework and assignments better and earlier.	This reflection is on learning and strategy choice.
By going through the resources at my own pace, I am able to develop a deeper understanding of the concepts.	This reflection is on learning and strategy choice.

3.3. AUTOMATED CODING

Following codebook validation, a total of 4,631 reflection artefacts were manually coded by the research team across the metacognitive sub-elements. The reflections were segmented into individual sentences using character position mapping to extract labelled text segments. Therefore, a total of 18,883 manually coded samples were used for model training, comprising both single-label (96.1%) and multi-label (3.9%) examples, with sentences ranging from brief statements to extended reflective passages, distributed across the five categories: None (47.4%), Strategy Choice (18.9%), Reflection on Learning (16.5%), Goal Setting (12.4%), and Effort Regulation (8.7%).

We selected GPT-4o-mini (gpt-4o-mini-2024-07-18) as the base model. GPT-4o-mini supports supervised fine-tuning through OpenAI's managed API service, unlike some larger proprietary models with restricted fine-tuning access. This enabled direct weight updates on our domain-specific classification task rather than relying solely on prompting strategies. Recent comparative studies demonstrate that fine-tuning smaller language models outperforms zero-shot, few-shot, and chain-of-thought prompting approaches for domain-specific classification tasks, particularly when ground truth labelled data is available (Latif & Zhai, 2024; Ramanathan et al., 2025). For a multi-label classification task with five well-defined categories and 18,883 training examples, including 733 multi-label samples demonstrating overlapping metacognitive constructs, the model's capacity was appropriate—large enough to learn nuanced metacognitive distinctions and handle complex multi-label patterns, yet small enough to avoid overfitting.

To automatically code the data, the experiment was divided into multiple stages with different sizes of training data using GPT models. At each stage of model training, we conducted tests using 100, 500, and 1,000 reflection samples, with each configuration repeated three times

to ensure reliability. For each sample size, we randomly shuffled to ensure samples from different time periods, students, and reflection contexts were distributed across all three sets, minimising temporal or contextual bias. The multi-stage training experiments were designed to examine the effect of dataset size on performance. The small subsample runs were intended to isolate the effects of data volume rather than optimise each condition. Accordingly, we used the provider's default ("auto") settings so the OpenAI platform would select reasonable, size-aware hyperparameters, sufficient for observing data effects even if not fully optimal.

In the final stage, the entire dataset was split into training (80%), validation (10%), and testing (10%) sets, with the full 18,883 samples used. This progressive approach enabled systematic evaluation of how training data affect model performance, with the final stage incorporating multi-label classification capability to handle the 733 examples demonstrating overlapping metacognitive constructs.

During the final model training, we conducted hyperparameter tuning on the entire training dataset to optimise the learning rate multiplier (LRM), one of the few user-configurable parameters in OpenAI's managed fine-tuning service. This optimisation step aimed at identifying the optimal learning rate multiplier for the final model. We tested LRM values of 0.5, 1.0, and 1.5 on the same full training dataset using 2 training epochs. Table 4 highlights the performance of the model using various LRM.

Table 4: Performance of the model using different LRM.

LRM	Macro-F1	Overall Accuracy	Strategy Choice	None	Goal Setting	Reflection on Learning	Effort Regulation
0.5	0.7541	80.40%	80.82%	87.46%	83.08%	68.44%	55.56%
1.0	0.7552	80.40%	80.82%	87.46%	83.08%	68.44%	55.56%
1.5	0.7541	79.85%	78.77%	84.77%	91.28%	66.78%	62.22%

Based on these results showing negligible differences in overall performance (Macro-F1 range: 0.7541-0.7552), we selected LRM 1.0 for final model training as it represents the default multiplier and achieved a marginally higher Macro-F1 (0.7552). Batch size and number of epochs were left on automatic, allowing the service to optimise these parameters based on dataset characteristics. This automatic optimisation is recommended by OpenAI for datasets of this size, as the service dynamically adjusts batch size based on available GPU memory and dataset complexity, and determines the optimal number of epochs through validation loss monitoring and early stopping to prevent overfitting. The overall similarity likely reflects the robustness of the fine-tuning process to modest learning rate variation at this dataset scale, consistent with findings in comparable studies (Latif & Zhai, 2024), where small LRM differences produce minimal performance shifts once sufficient training data are available.

4. RESULTS

In this section, we present the results of the automated coding of the metacognitive elements using different sampling sizes. Table 5 presents the results of training the model with 100

samples across three runs. Similarly, Table 6 and Table 7 present the results of training the model with 500 and 1,000 samples across three runs, respectively.

Table 5: Accuracy of the model across three runs using 100 random samples.

	Accuracy		
Construct	**Run 1**	**Run 2**	**Run 3**
No metacognition	79.22%	83.44%	64.78%
Goal Setting	77.84%	84.54%	90.72%
Reflection on Learning	57.91%	25.90%	65.83%
Strategy Choice	74.60%	64.13%	76.83%
Effort Regulation	64.62%	76.15%	67.69%
Overall	73.97%	70.89%	70.01%

Table 6: Accuracy of the model across three runs using 500 random samples.

	Accuracy		
Construct	**Run 1**	**Run 2**	**Run 3**
No metacognition	90.22%	86.33%	86.33%
Goal Setting	78.35%	93.81%	93.81%
Reflection on Learning	51.44%	54.68%	54.68%
Strategy Choice	79.37%	81.27%	81.27%
Effort Regulation	45.38%	55.38%	55.38%
Overall	77.93%	79.20%	79.20%

Table 7: Accuracy of the model across three runs using 1000 random samples.

	Accuracy		
	Run 1	**Run 2**	**Run 3**
No metacognition	83.44%	89.00%	87.89%
Goal Setting	80.41%	87.11%	80.93%
Reflection on Learning	75.18%	58.99%	76.26%
Strategy Choice	82.54%	77.78%	71.75%
Effort Regulation	63.85%	53.85%	46.92%
Overall	80.30%	79.75%	79.64%

Fine-tuning the model and training the model on 80% of the total coded data, evaluating its performance on another 10% of the data, and finally testing it on the remaining 10% of the data, the model achieved an overall performance accuracy of 81.94%. In Table 8, we present the accuracy of the model trained on the full dataset (18,883 samples), along with the average accuracy across 100, 500, and 1,000 samples.

Table 8: Accuracy of the model across the total dataset.

	Average accuracy			**Accuracy on the entire dataset**
Construct	**100 samples**	**500 samples**	**1000 samples**	
Effort regulation	69.49%	52.05%	54.87%	60.74%
Goal Setting	84.37%	88.66%	82.82%	91.28%
Strategy Choice	71.85%	80.64%	77.63%	81.16%
Reflection on learning	49.88%	53.06%	70.14%	67.44%
No metacognition	75.81%	87.63%	86.78%	88.24%
Overall accuracy	71.62%	78.78%	79.90%	81.94%

5. DISCUSSION

This study examined the development and application of a fine-tuned GenAI model to identify metacognitive elements in student reflections. By drawing on a large dataset of naturally occurring classroom artefacts and engaging in iterative co-design with educators and researchers, the project sought to explore the interpretive and methodological implications of integrating GenAI into qualitative educational research. The final model achieved an overall accuracy of 81.94% on 18,883 samples. While this level of performance meets accepted thresholds for reliability in educational research (see Cronbach, 1951), a closer analysis of classification trends across different sample sizes reveals key insights. For instance, Goal Setting achieved high accuracy even with a smaller number of examples, improving from 84.4% at 100 examples to 91.28% with the full dataset. In contrast, constructs such as Reflection on Learning and Effort Regulation exhibited marked variability across sample sizes. Reflection on Learning never reached an accuracy rate exceeding 72%, even with extensive training data and Effort Regulation was most accurate on only 100 samples. Strategy Choice initially improved with more examples but showed diminishing returns at scale. This experience highlights a broader consideration for the adoption of AI in educational research and practice. In applications involving qualitative student data, accuracy must be understood not simply as a numerical threshold, but as a function of the trust, interpretability, and contextual alignment that it enables. At the same time, there must be a recognition that some constructs resist 'easy' classification and that human interpretive insight remains essential. Progress in this space requires balancing ambition with pragmatism and knowing when methodological rigour must be complemented by practical decision-making.

Our findings invite a reconceptualisation of how human–AI partnerships are approached in educational research and practice. In qualitative research, human coders are rarely treated as fixed instruments. Instead, they are trained, calibrated, and supported through rounds of dialogue, codebook development, theoretical discussion, and iterative sense-making. Discrepancies are not simply corrected; instead, they are explored, and interpretations are refined over time in relation to theory and context. While our workflow is iterative and may resemble typical "human-in-the-loop" approaches, we argue that this framing obscures the deeper epistemic dynamics at play. In typical human-in-the-loop systems, the human either supervises or validates model decisions. By contrast, we positioned GenAI as a supporting

analytic partner—one that surfaces recurrent patterns, inconsistencies, and boundary cases that prompt renewed human interpretation. We use the term 'partner' deliberately to signal the AI's active role, but not to imply equality in agency, accountability, or interpretive capacity. GenAI shaped human coding through its outputs, but it did not reason with theory or take part in meaning-making. The distinction is not procedural but conceptual: our emphasis is not on efficiency or correction but on how AI outputs provoke further sense-making and theory refinement.

If we work with human coders in this collaborative way, should we not engage with AI in a similar way? Rather than positioning educators or AI as post-hoc validators in a "human-in-the-loop" system, we advocate for treating GenAI as a supportive analytic partner, one that supports, challenges, and evolves with human interpretation across all stages of the process (Jiang et al., 2021; Schroeder et al., 2025). In our process, this same principle guided how we interacted with the GenAI model. We began by providing a set of manually coded examples and gradually expanded this training data to help the model better reflect our intended constructs. As the model produced outputs, we reviewed samples and identified patterns of misclassification, particularly where construct boundaries were unclear or inconsistently applied. These instances led us to revise the codebook, clarify definitions, and improve annotation consistency. The model did not interpret meaning or adjudicate theoretical debates. However, its consistent application of learned patterns across scale allowed latent issues in our coding schema to become visible. By making these issues explicit, rather than leaving them latent in distributed human interpretation GenAI shaped the sensemaking process. Meaningful integration and collaboration of this kind requires careful alignment between the model's learning processes and the theoretical constructs it is intended to recognise. It also requires transparent annotation practices, clear coding frameworks, and ongoing dialogue with practitioners (Schroeder et al., 2025). In this view, the role of AI is not to replace human—or more specifically, teacher—judgment, but to amplify it by handling routine classification tasks and surfacing patterns that allow educators to focus on higher-order reflection, support, and decision-making (Khosravi et al., 2022). This aligns with interpretivist traditions that treat coding as a socially negotiated process (Strauss & Corbin, 1998), and it reflects our commitment to a human–AI partnership where meaning is shaped over time through interaction.

Bryda and Sadowski (2024) suggest that a hybrid approach, which integrates the strengths of both manual and automated coding, offers a promising pathway to achieving both depth and efficiency in qualitative analysis. In this model, the authors suggest that the AI is used to perform the initial categorisation of data, providing a broad structure that can be refined through targeted manual review. Human analysts then examine selected subsets of the data, adding contextual nuance and interpretive depth that automated systems may overlook. In our case, human input shaped the training data and theoretical framing, while AI generated feedback was integrated to enable scaled classification across the full dataset offering a practical balance between interpretability and efficiency. This process can be further strengthened through iterative cycles, in which AI-generated classifications are revisited and refined, allowing the system to learn from human input and progressively enhance the accuracy and relevance of its coding.

Our results suggest that GenAI may be well suited to identifying the *presence* of key constructs, while leaving open space for teachers and researchers to interpret their *quality*, *depth*, and *relevance*. This blended approach, which combines the efficiency and scalability of AI with the contextual and pedagogical sensitivity of educators, offers a promising way to make qualitative analysis more timely, equitable, and actionable. While this study focused on metacognition, the approach could be extended to other complex constructs such as critical thinking, emotional intelligence, or collaboration. These areas often defy standardised

measurement and rely on qualitative student expression. GenAI, when integrated thoughtfully and ethically, may help address these challenges by broadening access to meaningful assessment and reducing reliance on reductive or overly rigid rubrics (İpek et al., 2023).

5.1. Human–AI Partnerships in Real-World Educational Systems

This study offers early insights into how AI-generated outputs might support educational decision-making if integrated thoughtfully into classroom workflows. While researchers often emphasise model performance and classification accuracy (Zhang et al., 2025), less attention is given to how these outputs are translated into actionable insights within classroom practice. In this project, we engaged with educators and state education department staff across every stage of development, from co-constructing the codebook to reviewing and validating outputs. These collaborative processes not only helped ensure alignment with pedagogical priorities but also revealed broader opportunities for improving the trust, relevance, and usability of AI in schools. To be practically valuable, models trained on student reflections must be designed not just for technical accuracy but also with feedback and decision-making in mind. For instance, classification outputs could be embedded into teacher-facing dashboards (Barthakur et al., 2025) that summarise metacognitive patterns across a cohort or highlight students whose reflections suggest vague goal setting or limited strategy use. Student-facing tools might provide real-time, personalised prompts based on model classifications, supporting the development of more specific, reflective, or actionable learning goals. These design choices reflect a broader commitment to co-constructing systems alongside educators. By involving system actors throughout the project, we were able to build tools that enhanced trust and usability while also clarifying the infrastructure and resourcing required for broader implementation (Calisto et al., 2023; Zhang et al., 2025).

A critical insight from this process was the sheer investment required to build the model. Nearly 500 hours of specialist labour were dedicated to algorithm development, prompting a necessary reflection on the true value AI brings to educational research and practice. If a model can simply locate a sentence indicating goal setting, many experienced teachers are already capable of doing so, particularly when they know their students well. The value, then, cannot rest solely on replicating teacher judgment at the sentence level. Instead, we argue that the return lies in what this foundational work makes possible. The annotated datasets, codebooks, validation procedures, and educator partnerships all contribute to building the infrastructure required for scalable and sustainable use. In our case, this work also helped demonstrate to the education department what is involved in developing trustworthy and pedagogically relevant AI systems. By surfacing these resourcing demands and showing how they translate into usable tools, we were able to make a case for future investment.

5.2. Limitations

Several limitations should be acknowledged. First, while the AI model performed well within this specific dataset, its generalisability to other schools, age groups, or educational systems has yet to be tested. Variability in school-specific language, disciplinary expectations, or student writing ability may limit transferability. Second, the five metacognitive codes used in training were derived from a specific theoretical framework; alternative coding schemes could yield different interpretive results. Third, despite high inter-rater agreement among human coders, the coding process itself was theory-driven and therefore subject to bias and interpretive drift biases, which the AI may have inherited during training. Additionally, the model was not designed to provide qualitative feedback or nuanced explanations of its decisions, limiting transparency.

Without explainability mechanisms or confidence indicators, there remains a risk that educators could over-rely on AI classifications without adequate critical engagement. Finally, the choice of AI model and prompting strategy may have influenced the results. Different models or prompt designs could produce substantially different outcomes. To reduce this risk, we trained the selected model thrice using randomly sampled data at each sample size. However, this variability highlights an important limitation. Researchers using AI in educational contexts must carefully consider how model selection and configuration affect both findings and their interpretation.

5.3. Future Research Directions

Future work should expand on several fronts. Methodologically, research is needed to examine how AI models perform across diverse educational contexts, particularly in low-resourced or multilingual settings. Greater attention should also be paid to designing explainable AI systems that offer confidence scores, rationales, or traceable logic for their classifications (Calisto et al., 2023).

From a pedagogical perspective, there is a need to explore how educators interpret and act on AI-generated codes. What support do teachers need to integrate AI insights into their formative assessment practices? How do they reconcile AI classifications with their own interpretations of student work? Furthermore, how might co-design processes with educators shape the development of AI tools that are contextually grounded, pedagogically useful, and ethically robust? A central strength of the project was its use of naturally occurring classroom artefacts. By analysing reflections already embedded in existing assessment and reporting practices, the study avoided placing additional burdens on students or teachers. This approach not only enhances ecological validity but also increases the replicability and sustainability of AI-based analysis across different educational contexts.

Finally, future research should move beyond identifying the *presence* of learning constructs toward assessing their *quality*, *trajectory*, and *impact*. For instance, how can AI be trained not only to identify that a student has set a goal, but to evaluate whether it is developmentally appropriate, aligned to instructional targets, or likely to support academic growth? Addressing these questions will require deeper integration of AI into educational research workflows, not as a black box, but as a transparent and interrogable partner in the sense-making process.

6. Conclusion

This study examined the development and application of a fine-tuned GenAI model to classify metacognitive elements in student reflections artefacts. While the model achieved an overall accuracy of 80.98%, the deeper value of the work lies in demonstrating how AI can complement, rather than replace, educators through scalable and context-sensitive analysis of qualitative learning data. Meaningful integration of AI into education requires more than strong performance metrics. It depends on collaborative design with educators, sustained investment in infrastructure, and alignment with the practical conditions of classroom learning. In this project, ongoing engagement with stakeholders ensured that the model remained relevant, interpretable, and grounded in pedagogical needs. Although our focus was on metacognition, the same principles can be applied to other complex learning constructs such as critical thinking or emotional understanding. Advancing AI in education means paying close attention not just to what the models can achieve, but to how they are used, and how they serve the learners and educators at the centre of the system.

Ultimately, the value of GenAI in this context lies not in replacing educators, but in augmenting their practice. This project demonstrates that meaningful integration of AI into education depends on more than achieving strong performance metrics. It requires participatory design, sustained investment in foundational infrastructure, and a deliberate effort to align technological development with the practical demands of classroom learning across all learning contexts.

Declaration of Generative AI Software Tools in the Writing Process

During the preparation of this work, the authors used ChatGPT 4.0 in the background section in order to support clarification in writing. After using this tool the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Agbon, G. 2024. Who speaks through the machine? Generative AI as discourse and implications for management. *Critical Perspectives on Accounting*, *100*, 102761.

Alt, D., and Raichel, N. 2020. Reflective journaling and metacognitive awareness: Insights from a longitudinal study in higher education. *Reflective Practice*, *21*, 2, 145–158. https://doi.org/10.1080/14623943.2020.1716708

Azevedo, R. 2020. Reflections on the field of metacognition: Issues, challenges, and opportunities. *Metacognition and Learning*, *15*, 91–98.

Barany, A., Nasiar, N., Porter, C., Zambrano, A. F., Andres, A. L., Bright, D., Shah, M., Liu, X., Gao, S., Zhang, J., Mehta, S., Choi, J., Giordano, C., and Baker, R. S. 2024. ChatGPT for Education Research: Exploring the Potential of Large Language Models for Qualitative Codebook Development. In *Artificial Intelligence in Education*, A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, and I. I. Bittencourt, Eds. Springer Nature Switzerland 134–149. https://doi.org/10.1007/978-3-031-64299-9_10

Barthakur, A., Joksimovic, S., Kovanovic, V., Mello, R. F., Taylor, M., Richey, M., and Pardo, A. 2022. Understanding Depth of Reflective Writing in Workplace Learning Assessments Using Machine Learning Classification. *IEEE Transactions on Learning Technologies*, *15*, 5, 567–578. https://doi.org/10.1109/tlt.2022.3162546

Barthakur, A., Marrone, R., Esnaashari, S., Kovanovic, V., and Dawson, S. 2025. Advancing Holistic Decision-Making Systems in Schools: Insights From Academic Research and Practical Applications. *Journal of Computer Assisted Learning*, *41*, 3, e70021. https://doi.org/10.1111/jcal.70021

Belotto, M. J. 2018. Data analysis methods for qualitative research: Managing the challenges of coding, interrater reliability, and thematic analysis. *The Qualitative Report*, *23*, 11, 2622–2633. https://doi.org/10.46743/2160-3715/2018.3492

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., and Askell, A. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Bryda, G., and Sadowski, D. 2024. *From words to themes: AI-powered qualitative data coding and analysis*. 309–345.

Calisto, F. M., Fernandes, J., Morais, M., Santiago, C., Abrantes, J. M., Nunes, N., and Nascimento, J. C. 2023. Assertiveness-based Agent Communication for a Personalized

Medicine on Medical Imaging Diagnosis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–20. https://doi.org/10.1145/3544548.3580682

CHEN, S., AND MCDUNN, B. A. 2022. Metacognition: History, measurements, and the role in early childhood development and education. *Learning and Motivation*, *78*, 101786. https://doi.org/10.1016/j.lmot.2022.101786

CREAGH, S., THOMPSON, G., MOCKLER, N., STACEY, M., AND HOGAN, A. 2025. Workload, work intensification and time poverty for teachers and school leaders: A systematic research synthesis. *Educational Review*, *77*, 2, 661–680. https://doi.org/10.1080/00131911.2023.2196607

CRONBACH, L. J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

FLAVELL, J. H. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*, 10, 906–911.

GARG, R., HAN, J., CHENG, Y., FANG, Z., AND SWIECKI, Z. 2024. Automated Discourse Analysis via Generative Artificial Intelligence. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, 814–820. https://doi.org/10.1145/3636555.3636879

GASCOINE, L., HIGGINS, S., AND WALL, K. 2017. The assessment of metacognition in children aged 4–16 years: A systematic review. *Review of Education*, *5*, 1, 3–57. https://doi.org/10.1002/rev3.3077

GIBSON, A., KITTO, K., AND BRUZA, P. 2016. Towards the Discovery of Learner Metacognition From Reflective Writing. *Journal of Learning Analytics*, *3*, 2, 22–36. https://doi.org/10.18608/jla.2016.32.3

GISEV, N., BELL, J. S., & CHEN, T. F. (2013). Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy, 9*(3), 330-338.

HARWOOD, T. G., AND GARRY, T. 2003. An overview of content analysis. *The Marketing Review*, *3*, 4, 479–498.

HOLSTEIN, K., AND ALEVEN, V. 2022. Designing for human–AI complementarity in K-12 education. *AI Magazine*, *43*, 2, 239–248.

İPEK, Z. H., GÖZÜM, A. İ. C., PAPADAKIS, S., AND KALLOGIANNAKIS, M. 2023. Educational Applications of the ChatGPT AI System: A Systematic Review Research. *Educational Process International Journal*, *12*, 3. https://doi.org/10.22521/edupij.2023.123.2

ISMAIL, N. M., AND TAWALBEH, T. I. 2014. Effectiveness of a Metacognitive Reading Strategies Program for Improving Low Achieving EFL Readers. *International Education Studies*, *8*, 1, p71. https://doi.org/10.5539/ies.v8n1p71

JIANG, J. A., WADE, K., FIESLER, C., AND BRUBAKER, J. R. 2021. Supporting Serendipity: Opportunities and Challenges for Human-AI Collaboration in Qualitative Analysis. *Proceedings of the ACM on Human-Computer Interaction*, *5*, CSCW1, 1–23. https://doi.org/10.1145/3449168

KATZ, A., WEI, S., NANDA, G., BRINTON, C., AND OHLAND, M. 2023. *Exploring the Efficacy of ChatGPT in Analyzing Student Teamwork Feedback with an Existing Taxonomy* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2305.11882

KHOSRAVI, H., SHUM, S. B., CHEN, G., CONATI, C., TSAI, Y.-S., KAY, J., KNIGHT, S., MARTINEZ-MALDONADO, R., SADIQ, S., AND GAŠEVIĆ, D. 2022. Explainable Artificial Intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, 100074. https://doi.org/10.1016/j.caeai.2022.100074

KOVANOVIĆ, V., JOKSIMOVIĆ, S., MIRRIAHI, N., BLAINE, E., GAŠEVIĆ, D., SIEMENS, G., AND DAWSON, S. 2018. Understand students' self-reflections through learning analytics. In

Proceedings of the 8th International Conference on Learning Analytics and Knowledge, 389–398. https://doi.org/10.1145/3170358.3170374

LATIF, E., AND ZHAI, X. 2024. Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100210. https://doi.org/10.1016/j.caeai.2024.100210

LI, B., BHATTARAI, A., AND DING, Z. 2025. In search of humanness: Professional identities of qualitative research educators in the age of generative AI. *Learning, Media and Technology*, 1–13. https://doi.org/10.1080/17439884.2025.2521547

LICHTMAN, M. 2023. *Qualitative research in education: A user's guide*. Routledge.

MILES, M., HUBERMAN, A.M., AND SALDANA, J. 2014. *Qualitative Data Analysis: A Methods Sourcebook* (4th. ed.). Sage.

NGUYEN-TRUNG, K. (2025). ChatGPT in thematic analysis: Can AI become a research assistant in qualitative research?. *Quality & Quantity*, *59*(6), 4945-4978.

OHTANI, K., AND HISASAKA, T. 2018. Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance. *Metacognition and Learning*, *13*, 2, 179–212. https://doi.org/10.1007/s11409-018-9183-8

OZTURK, N. 2017. Assessing Metacognition: Theory and Practices. *International Journal of Assessment Tools in Education*, 134–134. https://doi.org/10.21449/ijate.298299

PARKER, J., RICHARD, V., AND BECKER, K. 2023. Guidelines for the Integration of Large Language Models in Developing and Refining Interview Protocols. *The Qualitative Report*. https://doi.org/10.46743/2160-3715/2023.6801

PINTRICH, P. R., WOLTERS, C. A., AND BAXTER, G. P. 2000. *Assessing metacognition and self-regulated learning*. https://digitalcommons.unl.edu/burosmetacognition/3/

PRESCOTT, M. R., YEAGER, S., HAM, L., RIVERA SALDANA, C. D., SERRANO, V., NAREZ, J., PALTIN, D., DELGADO, J., MOORE, D. 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

RAMADHANTI, D., GHAZALI, A. S., HASANAH, M., HARSIATI, T., AND YANDA, D. P. 2020. The Use of Reflective Journal as a Tool for Monitoring of Metacognition Growth in Writing. *International Journal of Emerging Technologies in Learning (iJET)*, *15*, 11, 162. https://doi.org/10.3991/ijet.v15i11.11939

RAMANATHAN, S., LIM, L. A., MOTTAGHI, N. R., & BUCKINGHAM SHUM, S. (2025, March). When the prompt becomes the codebook: Grounded prompt engineering (groproe) and its application to belonging analytics. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 713-725).

SALDAÑA, J. 2015. *The Coding Manual for Qualitative Research* (3rd. ed.). Sage, Newcastle upon Tyne.

SCHRAW, G. 2000. Assessing Metacognition: Implications Of The Buros Symposium. In *Issues in the measurement of metacognition,* G. Schraw, and J. C. Impara, Eds. Lincoln, Nebraska: Buros Institute of Mental Measurements, 297–321.

SCHROEDER, H., AUBIN LE QUÉRÉ, M., RANDAZZO, C., MIMNO, D., AND SCHOENEBECK, S. 2025. Large Language Models in Qualitative Research: Uses, Tensions, and Intentions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3706598.3713120

SCHUNK, D. H., AND GREENE, J. A. 2017. Historical, contemporary, and future perspectives on self-regulated learning and performance. In *Handbook of self-regulation of learning and performance*. Routledge, 1–15.

SIEMENS, G., MARMOLEJO-RAMOS, F., GABRIEL, F., MEDEIROS, K., MARRONE, R., JOKSIMOVIC, S., AND DE LAAT, M. 2022. Human and artificial cognition. *Computers and Education: Artificial Intelligence*, *3*, 100107.

SIIMAN, L. A., RANNASTU-AVALOS, M., PÖYSÄ-TARHONEN, J., HÄKKINEN, P., AND PEDASTE, M. 2023. Opportunities and Challenges for AI-Assisted Qualitative Data Analysis: An Example from Collaborative Problem-Solving Discourse Data. In *Innovative Technologies and Learning*. Y.-M. Huang and T. Rocha. Eds., 14099. Springer Nature Switzerland, 87–96. https://doi.org/10.1007/978-3-031-40113-8_9

SILVER, N. 2023. Reflective Pedagogies and the Metacognitive Turn in College Teaching. In *Using Reflection and Metacognition to Improve Student Learning* (1st ed.), D. LaVaque-Manty, D. Meizlish, N. Silver, M. Kaplan, and J. Rhem, Eds. Routledge, 1–17. https://doi.org/10.4324/9781003448570-1

STRAUSS, A., AND CORBIN, J. 1998. *Basics of qualitative research techniques*. Sage Publications, Inc.

ULLMANN, T. D. 2019. Automated Analysis of Reflection in Writing: Validating Machine Learning Approaches. *International Journal of Artificial Intelligence in Education*, *29*, 2, 217–257. https://doi.org/10.1007/s40593-019-00174-2

VEENMAN, M. V. J., AND SPAANS, M. A. 2005. Relation between intellectual and metacognitive skills: Age and task differences. *Learning and Individual Differences*, *15*, 2, 159–176. https://doi.org/10.1016/j.lindif.2004.12.001

WEI, X., CUI, X., CHENG, N., WANG, X., ZHANG, X., HUANG, S., XIE, P., XU, J., CHEN, Y., AND ZHANG, M. 2023. Chatie: Zero-shot information extraction via chatting with chatgpt. *arXiv Preprint arXiv:2302.10205*.

WHITE, J., FU, Q., HAYS, S., SANDBORN, M., OLEA, C., GILBERT, H., ELNASHAR, A., SPENCER-SMITH, J., AND SCHMIDT, D. C. 2023. *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT* (No. arXiv:2302.11382). arXiv. https://doi.org/10.48550/arXiv.2302.11382

WINNE, P. H. 2017. Cognition and metacognition within self-regulated learning. In *Handbook of self-regulation of learning and performance*. Routledge, 36–48.

WINNE, P. H., AND HADWIN, A. F. 1998. Studying as Self-Regulated Learning. In *Metacognition in Educational Theory and Practice*. Routledge.

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*, *1895*, G. Arastoopour Irgens, and S. Knight, Eds. Springer Nature Switzerland, 470–485. https://doi.org/10.1007/978-3-031-47014-1_32

ZAMFIRESCU-PEREIRA, J. D., WONG, R. Y., HARTMANN, B., AND YANG, Q. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. I*n Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21. https://doi.org/10.1145/3544548.3581388

ZHANG, H., WU, C., XIE, J., LYU, Y., CAI, J., AND CARROLL, J. M. 2025. Harnessing the power of AI in qualitative research: Exploring, using and redesigning ChatGPT. *Computers in Human Behavior: Artificial Humans*, *4*, 100144. https://doi.org/10.1016/j.chbah.2025.100144

ZHAO, Z., WALLACE, E., FENG, S., KLEIN, D., AND SINGH, S. 2021. *Calibrate before use: Improving few-shot performance of language models*. 12697–12706.

ZIMMERMAN, B. J. 2002. Becoming a Self-Regulated Learner: An Overview. *Theory Into Practice*, *41*, 2, 64–70. https://doi.org/10.1207/s15430421tip4102_2

ZSIGMOND, I., METALLIDOU, P., MISAILIDI, P., IORDANOU, K., AND PAPALEONTIOU-LOUCA, E. 2025. Metacognitive monitoring in written communication: Improving reflective practice. *Education Sciences*, *15*, 3, 299.

7. APPENDIX A

Table 9: Examples of student reflections where the Goal Setting metacognitive code was present.

Goal setting comment from the student reflection	Is GS present, yes or no?
I will do that	No
I'm going to do	No
I am planning	Yes
I plan to do this by trying to read at least 5 books over the semester as reading is an awesome way to improve these areas.	Yes
A goal for myself from the start of this year has to be to improve and expand vocabulary and I hope to achieve this by reading more fictional books.	First sentence is goal setting plus strategy choice.

Table 10: Examples of student reflections where the Reflection on Learning metacognitive code was present.

Reflection on Learning comment from the student reflection	Is ROL present, yes or no?
Based on the results of the PAT-M test, I need to improve my algebraic skills	Yes
I noticed that the questions I answered for the reading test were mostly wrong, which surprised me	Yes
"I have learned..." "I noticed..." "The results show that..." "I realized that..." "I have identified..."	Yes
"I improved capability X by doing course/project Y"	No, a simple statement is not enough to indicate the student has actually reflected

Table 11: Examples of student reflections where the Effort Regulation metacognitive code was present.

Effort Regulation comment from the student reflection	Is ER present, yes or no?
I want to procrastinate less and I want to feel like everything is sorted and controlled	Yes
Even though I didn't know anyone in most groups, I pushed myself to get involved and contribute.	Yes
Last year, I was a bit lazy with my assignments and would sometimes hand them in late and procrastinate often.	Yes, effort regulation and reflection on learning
Although I still have a lot of areas I'm struggling in, The Energy Equation, and other areas more focused on maths, perseverance is also a key skill I have developed this semester.	Yes, effort regulation and reflection on learning