Using LLMs to Identify Indicators of Persistence from Students' Dialogues with a Pedagogical Agent

Teresa M. Ober
ETS
Princeton, NJ, USA
tober@ets.org

Shan Zhang
University of Florida
Gainesville, FL, USA
zhangshan@ufl.edu

Diego Zapata-Rivera
ETS
Princeton, NJ, USA
dzapata@ets.org

Noah L. Schroeder
University of Florida
Gainesville, FL, USA
schroedern@ufl.edu

Anthony F. Botelho
University of Florida
Gainesville, FL, USA
abotelho@coe.ufl.edu

Conversational learning systems offer new opportunities to examine learning processes through chat log data. Constructs such as persistence, self-efficacy, interest, perceived challenge, and prior knowledge are known predictors of student performance but are challenging to detect at scale using traditional methods. This study explores the use of Large Language Models (LLMs) to automatically code indicators of these constructs from student chat logs collected through a conversation-based assessment (CBA) for middle school mathematics. Indicators included observable behaviors such as students' expressions of challenge, help-seeking, goal-setting, and self-regulatory strategies evident in their conversational interactions within the CBA. We evaluated multiple configurations of ChatGPT4o, varying temperature settings (0, .3, .7, 1) and model types (mini vs. regular), against human expert coders. The dataset comprised over 10,000 student turns collected from 107 middle school students classified as English learners as they interact with the CBA. Reliability was assessed within and between LLM configurations and humans. Results reveal systematic patterns: constructs with moderate theoretical coherence benefited from higher temperatures, while well-defined constructs required deterministic settings. Self-efficacy showed the highest human-LLM alignment. These findings illustrate the challenges of measuring complex psychological constructs and highlight the promise of human-LLM collaboration to enhance qualitative coding efficiency and validity in educational research. Supplemental materials are available online here: https://doi.org/10.17605/osf.io/s85ck.

Keywords: construct extraction, persistence, large language models (LLMs), model configuration, conversation-based assessment (CBA), human-LLM collaboration, qualitative analysis, educational data mining, construct validity, temperature settings

1. INTRODUCTION

When students engage with conversational learning systems, their chat exchanges create a digital trail of learning processes that can predict academic success (Zapata-Rivera & Forsyth, 2022). Critical psychological factors, including persistence, self-efficacy, interest, perceived challenge, and prior knowledge, consistently emerge as powerful predictors of learning outcomes (Nuutila et al., 2021; O'Reilly et al., 2019). Yet extracting these insights from student conversations has traditionally required painstaking manual analysis that cannot keep pace with today's data-rich educational environments. While more traditional statistical and machine learning approaches have been used to analyze natural language data in educational settings (Botelho et al., 2023; Crossley et al., 2015; Dowell & Kovanović, 2022), large language models (LLMs) present a transformative opportunity to automatically detect these complex psychological constructs within student dialogue, potentially unlocking real-time understanding of learning processes at unprecedented scale. However, critical questions remain about how to harness LLMs for reliable, valid, and consistent analysis of educational constructs (Liu et al., 2025).

Capturing authentic learning processes requires a delicate balance by gathering meaningful data without disrupting the natural flow of learning. Conversation-based assessments (CBAs) may strike this balance by transforming everyday student interactions into rich data sources (Zapata-Rivera et al., 2015). Digital learning platforms such as CBAs provide rich trace data that can reveal evidence of students' persistence patterns and associated learning outcomes (see Kai et al., 2018), where chat logs reveal genuine indicators of engagement and learning behaviors as they unfold (Zapata-Rivera et al., 2023). When LLMs can be used to systematically code such naturally occurring conversations, researchers may gain a lens into the psychological dynamics that drive learning while preserving the authentic, unfiltered nature of student expression. In the present study, we explore the potential for leveraging LLMs to extract indicators of persistence and related constructs from factors from conversational data gathered from a CBA. Despite the promise of using LLMs for capturing rich information from conversational data, we lack systematic understanding of which psychological constructs are amenable to LLM coding and under what configurations LLMs are able to do so, particularly for constructs with varying theoretical clarity.

1.1. PERSISTENCE AS CENTRAL TO SELF-REGULATED LEARNING

Among the psychological constructs that emerge from student conversations, some play a more central role than others in shaping students' ongoing engagement (cf. Clark & Saxberg, 2018). In particular, the capacity to remain engaged in learning, even when tasks are complex or feedback is ambiguous, is a critical component of academic success. This quality, often referred to as persistence, serves as both a signal of current motivational state and a predictor of long-term learning outcomes (Du et al., 2023). Understanding how persistence is reflected in student dialogue offers a meaningful entry point for studying the mechanisms of learning in action.

To situate this construct theoretically, we turn to the concept of self-regulated learning (SRL), which positions persistence as a core process through which learners actively orchestrate their cognition, motivation, and behavior to achieve academic goals (Zimmerman, 2002). Rather than simply a product of learning, persistence represents the dynamic capacity to sustain effort and remain engaged despite obstacles, embodying a learner's ability to manage motivation and maintain focus when confronted with difficulty, ambiguity, or failure (Skinner et al., 2020; Zimmerman, 2002). From the perspective of SRL, when a student is faced with a challenging task,

behaviors that emerge such as help-seeking (Karabenick & Gonida, 2017) or active problem decomposition (Zimmerman & Moylan, 2009) may reflect a strategic form of adaptive persistence in which a student is intentionally using available resources or knowledge to overcome a challenge by making it more tractable.

The construct of persistence operates across multiple levels of educational experience. At institutional and classroom levels, persistence manifests as continued academic participation despite structural barriers or personal adversity (Stewart et al., 2015; Tinto, 2017). At the task level, it emerges through sustained engagement in cognitively demanding activities and the determination to continue working through challenges (Battle, 1965; Skinner & Pitzer, 2012; Sparks et al., 2025). Students who demonstrate persistence consistently tend to outperform their peers, showing greater gains in learning and more positive academic trajectories over time (Gignac et al., 2021; Jozsa et al., 2014; Meindl et al., 2019; Porter et al., 2020). This pattern extends beyond academics, where sustained effort toward long-term goals also distinguishes high achievers across competitive professional fields (Lechner et al., 2019). A learner who demonstrates persistence not only monitors performance on a given task but also deploys motivational resources, such as self-efficacy and interest, to remain engaged through difficulties (Bandura, 1977; Hidi & Reninger, 2011; Schunk, 1985). Persistence has been regarded as a multifaceted construct shaped by personal dispositions, contextual conditions, and the characteristics of the task at hand (Skinner & Pitzer, 2012).

Persistence emerges from the interplay between enduring individual characteristics and situational factors that can be strategically influenced. As an enduring quality, persistence may be viewed as a relatively stable source of individual differences like conscientiousness. Students who are naturally more organized, dependable, and goal-oriented demonstrate greater perseverance across learning contexts (Martin et al., 2013; Minbashian et al., 2010). Yet understanding how persistence connects with malleable, situational factors proves crucial for educators seeking to cultivate productive persistence habits, particularly when students encounter academic obstacles. For example, situational or task-contingent persistence can be activated by characteristics of the task or environment, such as intrinsic interest or perceived relevance (Sansone & Thoman, 2005). Research also shows that achievement goals and utility value interventions can enhance students' interest and persistence in specific academic settings (Harackiewicz et al., 2000; Hulleman et al., 2010).

To inform potential interventions to support students' persistence, we focus on it in relation to constructs thought to be malleable or task-dependent. Previous research suggests that persistence is closely influenced by several interrelated psychological constructs (Sparks et al., 2025), notably self-efficacy, interest, perceived challenge, as well as prior knowledge, skills, and attitudes (KSAs). Recent meta-analyses further highlight the importance of prior knowledge in supporting persistence (Simonsmeier et al., 2022).

Self-efficacy, or the belief in one's own ability to succeed in specific tasks, has been consistently linked to higher levels of persistence (Bandura, 1977; Lent et al., 1984; Schunk, 1985). Learners who perceive themselves as competent are more likely to take on challenging tasks, expend greater effort, and rebound from setbacks with resilience (Lent et al., 1984; Schunk, 1985). Meta-analytic evidence supports the connections between self-efficacy and both academic persistence and achievement (Multon et al., 1991). At the same time, it is also important to note that self-efficacy is situation dependent and even likely to vary over the course of task completion (Bernacki et al., 2015).

As noted above, interest also plays a significant motivational role in persistence. Both situational interest, arising spontaneously from engaging or novel aspects of a task, and individual interest, which reflects deeper, long-term personal value placed on a topic, have been shown to

increase students' willingness to persevere (Reninger & Hidi, 2011; Tulis & Fulmer, 2013). Additional research suggests that interest mediates the relationship between motivation and persistence, supporting academic success (Ainley et al., 2002; Krapp, 2002). When learners are intellectually or emotionally invested in the content, they are more likely to maintain attention and effort, even during complex or repetitive tasks.

In addition, perceptions of perceived challenge play a critical role; students are more likely to persist when they perceive challenges as surmountable and aligned with their skill level (Eccles & Wigfield, 2020; Wigfield & Eccles, 2000). The belief that one can overcome difficult tasks is central to persistence and academic success (Bandura, 1977; Pintrich & De Groot, 1990).

Finally, the learners' prior KSAs serve as critical cognitive resources that influence persistence. Students who possess relevant background knowledge or more developed academic skills are better able to comprehend and engage with difficult material (Simonsmeier et al., 2022; Tobias, 1994). Studies have shown that prior knowledge influences learning approaches and supports sustained engagement and goal pursuit (Dochy & Alexander, 1995).

1.2. Toward Automating the Measurement of Persistence

Given the importance of non-cognitive factors associated with learning, such as persistence and related constructs, developing a methodology that can quickly and accurately identify such factors through varied data sources has great potential to provide insights into student learning, as well as the development of personalized learning tools to support it. However, the measurement of psychological constructs such as persistence and related constructs in educational contexts represents a significant methodological challenge due to their inherent complexity (Sparks et al., 2025). These constructs are dynamic, context-dependent phenomena that interact with environmental features and emerge through subtle behavioral and linguistic expressions (Skinner et al., 2020). Such constructs need to be clearly defined with information about what constitutes evidence for them in ways that are based on existing theoretical and empirical research on assessment. Automating the measurement of such constructs could be used in the design of adaptive learning systems that respond to real-time fluctuations in learner states (Mellon et al., 2024).

Traditional natural language processing (NLP) approaches have demonstrated efficacy in inferring various learner states from textual data, including engagement, affect, and metacognition (Zapata-Rivera & Forsyth, 2022). However, these approaches often required extensive manual annotation and struggled with the nuanced, context-dependent nature of most psychological constructs. In response to these limitations, recent advancements in the use of LLMs in qualitative coding open new possibilities for identifying complex constructs at scale and offering the potential to accelerate the annotation processes for open-ended student data, tasks that were previously labor-intensive and subject to significant inter-rater variability.

While LLMs offer significant promise for qualitative coding in educational research, their adoption introduces several methodological challenges that require systematic attention. Recent reviews highlight both the opportunities and pitfalls of LLMs in educational contexts, emphasizing the need for careful methodological design (Kasneci et al., 2023). For example, Barany et al. (2024) explored how ChatGPT can assist in qualitative codebook development in education research, demonstrating that while LLMs can generate useful initial codes and suggest conceptual relationships, human oversight remains critical to ensure conceptual clarity and contextual appropriateness. In this section, we identify three primary areas which warrant further inquiry to improve the use of LLMs for qualitative analysis in educational research.

First, certain constructs are inherently complex and context-dependent, requiring nuanced interpretation that extends beyond simple pattern recognition. Traditional coding relies heavily on human judgment to navigate ambiguous meanings and cultural contexts (Charmaz, 2006), while LLMs may struggle with subtle contextual cues and implicit meanings that human coders intuitively understand (Qiao et al., 2024; Ziems et al., 2023). For example, comparable to human inter-rater reliability, some studies have revealed significant limitations when using LLM-based approaches to coding more abstract concepts or culturally specific content (Kuzman & Ljubešić, 2025; Törnberg, 2025). Research by Dunivin (2025) found that ChatGPT-4 achieved at least acceptable intercoder reliability (Cohen's $\kappa \geq 0.6$) for 8 of 9 socio-historical codes when using chain-of-thought reasoning, while ChatGPT-3.5 performed poorly on the same tasks, highlighting the importance of model selection for complex interpretive work.

Second, the non-deterministic and black-box nature of LLM-based coding approaches introduces transparency and reproducibility concerns. Unlike traditional qualitative methods where coding decisions can be traced and justified, LLM outputs often lack clear explanatory pathways (Ober et al., 2026; Ouyang et al., 2024). This opacity challenges fundamental qualitative research principles of transparency and methodological rigor.

Third, model selection and parameter tuning introduce additional variability that can significantly impact coding outcomes. Studies reveal substantial variability in LLM performance across different prompting strategies and model configurations, with some approaches yielding contradictory results even when applied to identical datasets (Anagnostidis & Bulian, 2024; Razavi et al., 2025). Prior studies have found that ChatGPT-4 achieved higher overall accuracy than human coders when using certain techniques (e.g., chain-of-thought, assertion enhanced few-shot learning), with particularly marked improvements in identifying nuanced categories (McClure et al., 2024). Such analyses have revealed considerable differences in model output based on prompting strategy, highlighting how prompting can determine coding outputs regardless of underlying model capabilities (Dunivin, 2025).

Model type may represent yet another critical factor, though its effects may be more nuanced than commonly assumed. Furthermore, the interactive effects of model type and prompt strategies are not well understood. While larger models generally produce more coherent and well-articulated themes, there is a point of diminishing returns where mid-sized models combined with sophisticated prompting may achieve results nearly comparable to the largest models but with significantly lower computational cost. For example, Zhang et al. (2025) developed an LLM-based framework to classify and correct students' programming knowledge using a predefined taxonomy, finding that mid-sized models with well-designed prompts achieved performance close to larger models but with lower computational costs. In particular, LLM prompts that make use of rubrics with clear definitions and examples for what should count as evidence may support effective prompting engineering. Another study by Yoshida (2025) found little difference in the accuracy of scored written responses when using more detailed v. simplified rubric of three of the four LLMs tested in the study (Claude 3.5 Haiku, Gemini 1.5 Flash, ChatGPT4o-mini, and Llama 3 70B Instruct), with one of these LLMs (Gemini 1.5 Flash) even showing decreased performance when a more detailed rubric was used. Such findings challenge the "bigger is better" narrative and suggest that strategic, though not necessarily more detailed, prompting may be more critical than raw model scale for many qualitative tasks. These findings may offer practical implications, particularly for researchers with limited computational resources.

Temperature settings require equally careful calibration, as they fundamentally alter the deterministic nature of model outputs. Though temperature defaults range by LLM, lower temperatures (e.g., 0–0.3) have been found to produce more deterministic, focused outputs ideal

for deductive coding tasks, while higher temperatures (e.g., 0.8–1.5) are thought to produce more creative outputs (Peeperkorn et al., 2024), generating more diverse and novel text suitable for exploratory thematic analysis. However, even setting temperature to 0 does not guarantee perfect reproducibility, as model updates, non-deterministic software elements, and API-specific configurations can still produce variations across sessions and platforms. The technical complexity extends beyond temperature to other sampling parameters: top-*p* and top-*k* sampling create distinct qualitative effects, with top-*k* imposing hard limits on token choice while top-*p* provides probability-based filtering that can dramatically alter coding consistency.

These technical decisions require systematic evaluation to ensure methodological choices produce outcomes likely to be reliable and interpretable by researchers (Heseltine & von Hohenberg, 2024). At the same time, relatively few studies to date have explored how different model parameters interact with features of constructs to affect the likelihood of reliable codes. Construct characteristics may be as important as technical parameter choices in determining coding success, yet this interaction remains poorly understood. Without systematic frameworks for matching parameter configurations to construct types, researchers risk applying inappropriate technical settings that may systematically bias results toward certain types of constructs while undermining the reliability of others. Although prior work shows LLMs can code well-defined constructs, relatively few studies have systematically examined how construct characteristics interact with model parameters to affect reliability (Liu et al., 2025), especially for complex constructs like persistence.

1.3. Construct Analysis

Recognizing that construct characteristics shape coding reliability, it may be important to consider how specific features of qualitative constructs interact with LLM performance. While generalization is limited due to the qualitative nature of this study, developing a taxonomy of construct dimensions may help explain variability in LLM-generated codes across different settings. Accordingly, we used construct conceptualizations to identify key features of each construct.

In one study, researchers assessed ChatGPT-4's coding performance across three educational datasets using five construct dimensions: clarity, concreteness, objectivity, granularity, and specificity (Liu et al., 2025). The researchers found ChatGPT-4 generally aligned well with human coders, with performance influenced by construct characteristics and prompting strategy. Constructs high in clarity and concreteness were reliably coded using zero-shot prompting, while more complex constructs benefited from few-shot or context-based prompts, highlighting the need to align coding methods with construct features.

While the dimensions proposed by Liu et al. (2025) are useful for evaluating coding reliability, we proposed an alternative framework (see Table 1) to support the evaluation of construct characteristics with an emphasis on measurement. Grounded in educational psychology measurement practices and theories of validity (e.g., Cronbach & Meehl, 1955; Kane, 2013; Messick, 1995), our framework includes operational definition, empirical grounding, and conceptual integrity. These dimensions aim to ensure constructs are clearly defined, empirically validated, and conceptually distinct, aligning with a causal perspective on validity (Borsboom et al., 2004) and addressing issues like construct conflation in self-efficacy and motivation research.

Table 1: Dimensions and sub-dimensions for construct analysis.

Dimension	Sub-dimension	Definition
Operational Definition	Clarity	Precise, unambiguous articulation enabling consistent interpretation across researchers and contexts (Messick, 1995).
	Specificity	Clear boundaries distinguishing the construct from related concepts within its nomological network, preventing construct-irrelevant variance (Cronbach & Meehl, 1955; Shepard, 2016).
	Granularity	Level of detail in defining a construct's components, capturing its full complexity while avoiding construct underrepresentation (Shepard, 2016).
Empirical Grounding	Concreteness	Observable manifestation of the construct that can be systematically documented, reflecting the causal theory that valid constructs exist independently of measurement (Borsboom et al., 2004).
	Objectivity	Verifiability through systematic observation and replicable measurement procedures, minimizing subjective interpretation and researcher bias (Kane, 2013).
	Measurability	Capability to be quantified through reliable instruments producing consistent scores across contexts, supporting structural validity (Messick, 1995).
Conceptual Integrity	Theoretical Coherence	Consistency with established theoretical frameworks and empirical findings, strengthening the construct's position in its nomological network (Cronbach & Meehl, 1955; Loevinger, 1957).
	Context sensitivity	Adaptability of operationalization to remain meaningful across different tasks, settings, and populations (Zumbo, 2009).
	Sensitivity to Individual Sources of Variability	Capacity to account for and interpret differences from individual characteristics, developmental stages, or group membership (Messick, 1995; Meredith, 1993).

This construct evaluation framework offers several advantages. It broadens theoretical application by addressing nomological networks (Cronbach & Meehl, 1955), strengthens research integration through clearer theory-measurement links (Loevinger, 1957), and enhances validity by reducing construct-irrelevant variance (Shepard, 2016). It also provides practical guidance for assessment development (Newton & Shaw, 2014) and emphasizes context sensitivity, acknowledging that constructs like persistence may vary across learning environments (Zumbo, 2009).

We reviewed the literature for each construct to evaluate alignment with sub-dimensions and reviewed qualitative ratings assigned by the first author (high, moderate, low) based on the dimensions and sub-dimensions listed in Table 1. Reliability of construct ratings was not calculated. This step was done to aid in the interpretation of LLM codes and outputs based on different features of constructs only. As such this approach enabled a high-level cross-construct comparisons, highlighting relative strengths and weaknesses in definition, measurement, and theoretical integration. While our interpretations of the construct analysis may reflect a degree of arbitrariness or confirmation bias, the analysis should be viewed as offering a descriptive foundation to interpret the results of the present study for future refinement of prompts and construct definitions.

Self-efficacy emerged as a reasonably well-defined construct, receiving high ratings for the operational definition. Bandura's (1977, 2006) domain-specific definition is widely cited and distinguishes self-efficacy from related constructs. However, its context sensitivity remains high, as beliefs vary by task and situation (Pajares, 1996), objectivity is only moderate due to varied instruments (Klassen & Usher, 2010). In contrast, interest showed more variability, receiving moderate to low ratings for clarity and consistency. Although distinctions between situational and individual interest are well established (Reninger & Hidi, 2011), definitional

ambiguity persists (Krapp, 2002), and measurement approaches are inconsistent (Ainley & Hidi, 2014). Additionally, its conceptual overlap with intrinsic motivation complicates boundary clarity (Wigfield et al., 2021).

Perceived challenge of task lacked clarity and specificity of its operational definition, frequently overlapping with constructs such as difficulty, complexity, or self-efficacy (Efklides, 2011). The high degree of context-specificity also limited comparability across studies (Silvia, 2005). Prior KSAs presented even greater challenges, scoring low on clarity and consistency. Its multidimensional nature makes it difficult to define precisely (Alexander, 2003), and while often linked to expertise and experience (Dochy & Alexander, 1995), our focus on K–12 students emphasized culturally and community-derived KSAs. These are highly task- and topic-specific, further complicating measurement (O'Reilly et al., 2019; Shapiro, 2004).

Finally, persistence shared many of these issues, with low ratings for clarity, specificity, and measurement consistency. Its definition often overlaps with other related constructs , such as effort, engagement, or determination (Sparks et al., 2025), and behavioral indicators vary widely due to task-specific constraints (DiCerbo, 2014; Zhou & Kam, 2017). These similarities suggest that both constructs may benefit from clearer conceptual delineation and more consistent operationalization.

Based on this construct analysis, persistence emerged as one of the relatively less well-defined constructs, characterized by lower clarity given definitional boundaries vary across theoretical frameworks (DiCerbo, 2014), lower specificity considering behavioral indicators are context-dependent and often overlap with related constructs like self-regulation (Zhou & Kam, 2017), as well as lower granularity given there are challenges defining persistence across different task types (Sparks et al., 2025). In contrast, constructs such as self-efficacy demonstrated higher operational clarity with precise definitions (Bandura, 1977), clear boundaries from related constructs (Bandura, 2006), and detailed specifications of its components (Schunk & DiBenedetto, 2016). Such variation in construct clarity allowed us to examine a relatively unexplored line of inquiry into whether agreement patterns between and among human and LLM coders reflect characteristics of the construct itself. Specifically, we predicted that constructs with lower operational clarity would show greater divergence in coding approaches, even when individual coders (human or LLM) maintain internal consistency. When LLMs demonstrate high self-consistency (test-retest reliability) but diverge from human coders, this pattern may indicate not LLM failure, but rather consistent failure to understand context, nuance, or another dialogue-based indicator of the underlying construct. This distinction is particularly important for psychological constructs that vary in operational clarity. Examining agreement patterns across constructs with different levels of definitional precision may support a better understanding of both the capabilities and limitations of LLM-based coding.

1.4. Research Aims

There appears to be great potential for using LLMs to analyze unobtrusive data sources like chat logs, thus automating the assessment of persistence and related constructs in real time without disrupting the learning experience. A systematic investigation of how different LLM configurations affect the detection of persistence could contribute to an understanding of methods that will identify which psychological constructs are most amenable to LLM-facilitated coding, while addressing critical gaps in understanding how to ensure reliability, validity, and unbiased analysis when using such tools.

With these current gaps in mind, the goal of this investigation is to evaluate the reliability and accuracy of LLM-generated coding of key psychological and behavioral constructs,

including students' persistence, as captured in chat log data from a conversation-based assessment. This study specifically examines the potential for LLMs to replicate human judgment in labeling constructs relevant to student engagement and learning processes during interactive problem-solving tasks. Toward this aim, the study is guided by several key research questions (**RQs**) and associated hypotheses (*Hs*).

- **RQ1**: *Which LLM configurations (model type: regular v. mini; temperature: 0, 0.3, 0.7, 1) produce the most self-consistent set of codes across the entire dataset?*
  We hypothesized that lower temperature settings (0, 0.3) would produce higher test-retest reliability than higher temperatures (0.7, 1) due to reduced randomness in token selection (*H1a*), and that non-mini models (GPT4o) would demonstrate higher self-consistency than mini models (GPT4o-mini) due to greater model capacity (*H1b*).
- **RQ2:** *Which pairs of different LLM model configurations (model type: regular v. mini; temperature: 0, 0.3, 0.7, 1) produce the most reliable set of codes?*
  We hypothesized that configuration pairs using the same base model would show higher agreement than cross-model pairs (*H2a*), and that agreement would decrease as temperature difference increases (*H2b*).
- **RQ3:** *To what extent are the different types of LLM-generated codes consistent with human-derived codes?*
  We hypothesized that human-LLM agreement would be lower than human-human agreement due to differences in interpreting contextual information (*H3a*), but that LLMs would show greater self-consistency than humans due to the application of deterministic rather than situated rules (*H3b*).
- **RQ4:** *To what extent does consistency of codes generated by the LLM configurations and human raters vary across psychological constructs including Self-efficacy, Interest, Prior KSAs, Perceived Challenge, and Persistence? Are there any notable patterns based on the qualities of the construct?*
  Drawing on the construct analysis (see Table S1 in Supplemental Materials), we hypothesized that agreement would be greater for constructs with higher operational clarity ratings (i.e., Self-efficacy), lower for constructs with moderate clarity ratings (Interest, Prior KSAs), and lowest for constructs with still lower clarity ratings (Persistence, Perceived Challenge) (*H4a*). We further hypothesized that lower construct clarity ratings would have a greater impact on codes derived by human coders more than LLMs (*H4b*), and that constructs with higher concreteness would show higher agreement regardless of overall clarity (*H4c*).
- **RQ5:** *Of the LLM configurations found to be most reliable and consistent, to what extent are codes generated from turns produced by students for each construct by each coding approach correlated with estimates of task performance?*
  We hypothesized that Persistence and Perceived Challenge would show negative correlations with performance (*H5a*), while Self-efficacy and Prior KSAs would show positive correlations (*H5b*), and Interest would show weak or null correlations (*H5c*). Despite disagreement on specific instances, we further hypothesized that all coders would show the same direction of correlation, demonstrating convergent construct validity (*H5d*).

2. METHODOLOGY

To examine the potential of LLMs for automated coding of student dialogue, we analyzed previously collected data from research developing a CBA to evaluate English and math proficiency among middle school students, all of whom were English learners ($N = 107$; $Mean_{age} = 12.4$ years, $SD_{age} = 1.03$; see Lopez et al., 2021). Based on responses of participants who provided demographic information (see Table S2 in Supplemental Materials), the sample included approximately the same number of females (50.68%) and males (49.32%). Most participants were in 6th grade (54.79%), with somewhat smaller proportions in 7th (27.40%) and 8th grade (17.81%). Participants reflected linguistically diverse households, with nearly half speaking both English and Spanish (47.95%) at home, and over a third speaking only Spanish (36.99%), with somewhat smaller percentages indicating that other languages, such as Gujarati (4.11%), were spoken at home.

The CBA was designed to integrate principles of naturalistic dialogue, peer interaction, and scaffolded problem solving to simulate authentic, collaborative classroom experiences within a digital environment (Lopez et al., 2021). It comprised a sequence of six structured, interactive conversations between the student and three pedagogical agents, a teacher agent and two peer agents. The conversational format is designed to simulate small-group mathematical problem solving, encouraging students to engage in turn-taking, interpret information, and respond meaningfully to context-specific prompts. The system uses NLP to analyze student input and adjust the dialogue path based on characteristics of the student input, using a match score system. This design allows the conversation to flexibly adapt in response to student performance.

The assessment content reflected upper elementary and middle school math curricula. Students completed up to seven different tasks, which targeted the following six key competencies: (1) understanding directions; (2) comprehending short passages; (3) finding and interpreting information in tables; (4) understanding mathematical language related to ratios and proportions; (5) converting contextual problems into mathematical expressions (e.g., ratios, unit rates); and (6) demonstrating proportional reasoning by solving quantitative problems.

Each task is presented from a first-person perspective with embedded video clips to provide visual and contextual scaffolding. The assessment is designed to elicit conversational language in mathematical contexts, potentially allowing for analysis of multiple constructs including those mathematical concepts which were the focus of the task as well as other cognitive and non-cognitive constructs such as Persistence, Self-efficacy, Interest, Perceived Challenge, and Prior KSAs. The resulting chat log data provides a rich, multi-turn corpus of student language suitable for both human and machine coding. Overall, the participation in the conversation varied between the human student and the pedagogical agent, with average counts of turn per user ranging from about 8 to 36 and average words per turn generally between 7 and 17 (see Tables S3 and S4 in Supplemental Materials). In particular, students produced fewer turns on average ($mean = 24.55$, $SD = 8.28$) with shorter responses (about 7 words per turn) compared to pedagogical agents (peer agent 1: $mean = 36.34$ turns, $SD = 10.65$; ~15 words/turn; peer agent 2: $mean = 34.92$ turns, $SD = 11.97$; ~17 words/turn). Across tasks, students consistently have lower average words per turn (e.g., in Section 1, 6.09 words on average for the student vs. 9.66 for peer agent 1 and 15 for peer agent 2), underscoring their briefer contributions relative to the pedagogical agents.

2.1. LLM CODING

We explore the potential for LLMs to identify indicators of Persistence and related constructs (Self-efficacy, Interest, Perceived Challenge, Prior KSAs) from chat log data by focusing on the consistency of LLM-generated coding across different ChatGPT4o models varying based on model type (regular v. mini) and temperature settings (0, 0.3, 0.7, 1), and with respect to human evaluation and coding. Across all LLM conditions, the same prompt was used (see Appendix B in Supplemental Materials). The construct definitions and prompt were developed in collaboration with the research team and iteratively refined through group discussions until a consensus was reached based on definitions from research in these areas and construct operationalizations found in research. This refinement process focused primarily on clarifying the operational definitions of each construct (see Appendix C for Version 1 and Version 2 of coding instructions), while the core prompt structure and model parameters remained relatively constant across iterations. We then re-ran the same prompt using the same model parameters to evaluate reliability of the same model configuration on two different occasions. Though past research has found certain LLMs, particularly those in the ChatGPT family, to be relatively stable across multiple runs (Shah et al., 2025), we wanted to confirm that variation in outputs was not unduly influenced by randomness. The first batch was run between March 30–April 2, 2025, while the second batch was run between May 28–30, 2025.

To assess the quality of the output, human reviewers evaluated the appropriateness of the codes with a subset of data that was human-coded to calculate reliability. For example, it was found in several instances that the regular (i.e., ChatGPT4o, not ChatGPT4o-mini) model with temperature set at 1 applied codes related to but not otherwise included in the original prompt, such as "Curiosity" and "Engagement." Instances of codes such as these were removed from the dataset prior to analysis.

2.2. HUMAN CODING

To evaluate the LLM-generated codes, we compared model outputs to a human-annotated benchmark created under tightly aligned conditions. Two trained human coders with over five years of experience each in assessment development were provided with coding instructions similar to the one used to elicit outputs from the LLMs, along with spreadsheet template for recording codes. This approach was designed to mirror the LLM task structure and output format as closely as possible, thereby increasing the comparability between human- and machine-generated labels. The human coders applied construct-level codes (i.e., Self-efficacy, Interest, Persistence, Perceived Challenge, Prior KSAs) to turns within the chat log dataset.

To ensure the two trained human coders were applying codes consistently, a series of meetings occurred during which the two coders and researchers met to discuss the decisions leading to the use of codes, identified patterns in their decisions, and refined the documentation and definitions of constructs to make them more consistent in applying codes. After finishing an initial round of coding, the instructions for the human coders were then revised to reflect actual coding decisions based on mutual agreement. Appendix C in Supplemental Materials provides the exact instructions provided to the human coders initially (Version 1) and after undergoing revisions that emerged through conversation during the meetings (Version 2). These revisions involved adding specific behavioral indicators (e.g., for Persistence: 're-attempting, asking for clarification, or expressing a desire to keep working') and clarifying construct boundaries, but did not alter the fundamental coding categories or introduce new constructs. This process of training and calibration reflects current best practices for scoring or otherwise identifying

qualities of cognitively complex assessment data (Bauer & Zapata-Rivera, 2020; McCaffrey et al., 2021). One of the human coders (human1) coded the data set in its entirety (N = 10,919), while the other (human2) only coded a subset of randomly assigned task section-level conversations (N = 2,092).

2.3. Comparing LLM-LLM and Human-LLM Coding Outputs

All outputs, including those by each LLM and the human coder, were evaluated as if they were produced by independent coders. To assess the reliability and consistency of codes generated by different "coders," we employed several commonly used agreement metrics. First, we calculated percent agreement as a basic indicator of how often the LLMs produced the same labels as the human benchmark. While this metric offers a straightforward view of overall alignment, it does not account for agreement occurring by chance (Krippendorff, 2011). To address this limitation, we also calculated Krippendorff's Alpha (α), a measure of reliability that adjusts for chance agreement and supports comparisons across multiple coders and categories (Krippendorff, 2004). Krippendorff's α was computed at the construct level, aggregating across all LLM models and turns, to estimate the overall consistency of LLM coding relative to human expectations. In addition, we computed pairwise Cohen's Kappa (κ) scores (Cohen, 1960) between different LLM model configurations (e.g., comparing ChatGPT4o/temperature=0.3 to ChatGPT4o-mini/temperature=0.3) and between LLMs and the human coders. Each of these metrics provide insight regarding the alignment across different coders and constructs.

Table 2: Count and proportion of codes per construct assigned by each "coder."

Coder / LLM Configuration		Self-efficacy		Interest		Prior KSAs		Perceived Challenge		Persistence		Count of Messages with Codes
Model Type	Temperature	N	%	N	%	N	%	N	%	N	%	
ChatGPT4o	0	2179	32.1%	2749	40.5%	1485	21.9%	1647	24.3%	1191	17.5%	6788
ChatGPT4o	0.3	2173	32.1%	2730	40.3%	1488	22.0%	1599	23.6%	1251	18.5%	6773
ChatGPT4o	0.7	2178	32.7%	2661	40.0%	1616	24.3%	1537	23.1%	1205	18.1%	6660
ChatGPT4o	1	2139	31.7%	2769	41.1%	1738	25.8%	1535	22.8%	1174	17.4%	6745
ChatGPT4o-mini	0	2950	48.8%	1834	30.3%	2576	42.6%	1773	29.3%	355	5.9%	6049
ChatGPT4o-mini	0.3	2923	46.5%	1710	27.2%	2811	44.7%	1923	30.6%	407	6.5%	6288
ChatGPT4o-mini	0.7	2921	46.3%	1699	27.0%	2841	45.1%	1876	29.8%	400	6.3%	6303
ChatGPT4o-mini	1	2859	44.8%	1737	27.2%	2913	45.6%	1851	29.0%	471	7.4%	6382
Human Coder 1	–	2331	47.5%	1981	40.4%	0	0.0%	0	0.0%	1498	30.5%	4908
Human Coder 2	–	846	53.5%	609	38.5%	0	0.0%	14	0.9%	164	10.4%	1581

Note: Multiple codes could appear in the same turn and thus sum of percentages could exceed 100.

3. Results

Analyses focused on the interrater reliability of LLM outputs across the five constructs: Perceived Challenge, Interest, Self-efficacy, Prior KSAs, and Persistence. Table 2 provides an indication of how frequently constructs were identified by different LLMs as well as by the human coders. Note that the counts reflect the codes assigned by the models on the first run, rather than the second run which was primarily used to evaluate the stability of the LLM outputs. The overall frequency that certain constructs were identified varied considerably between raters, particularly between LLM and human raters. However, self-efficacy appeared as the most frequently coded construct (see Table S5 in Supplemental Materials).

We further examined how model architecture (specifically model type: regular v. mini) and temperature settings had affected coding consistency through pairwise comparisons using Cohen's Kappa and percent agreement metrics. Findings are organized by research question, highlighting construct-specific patterns that inform optimal LLM configurations for educational coding contexts.

3.1. Consistency of LLM Configurations on Separate Occasions (RQ1)

To examine the stability of LLM-generated codes within individual model configurations, we calculated reliability using Cohen's κ across two independent generation runs with identical parameters (i.e., model type, model temperature, and prompt). We refer to these analyses as "within-configuration" reliability. Tables S6 and S7 in Supplemental Materials provide a summary of the within-configuration reliability estimates overall and by construct. Results of this within-configuration reliability analysis suggested that LLM configurations demonstrated overall consistency when re-coding the same dataset (see Table S6 in Supplemental Materials). All eight configurations achieved Cohen's κ values above 0.86 and percent agreement exceeding 88%. Specifically, the configurations ranged from $\kappa = 0.865$ to $\kappa = 0.889$, with ChatGPT4o/temperature=0 showing the highest within-configuration reliability ($\kappa = 0.890$, 90.52% agreement) and ChatGPT4o-mini/temperature=1 showing the lowest ($\kappa = 0.865$, 88.28% agreement). Temperature effects on within-configuration reliability showed minimal systematic patterns, with all configurations maintained strong within-configuration reliability regardless of temperature parameter. In general, temperature effects were minimal and non-monotonic. Within ChatGPT4o, consistency ranged only from 0.882 to 0.890 (0.8% variation), and temperature=0.7 and temperature=1 performed nearly identically ($\kappa = 0.882$). Within ChatGPT4o-mini, temperature=0.3 actually outperformed temperature=0 (0.874 vs. 0.866) (*H1a partially supported*). Across model types, ChatGPT4o configurations consistently outperformed ChatGPT4o-mini configurations. The four ChatGPT4o configurations showed high inter-rater reliability (κ = 0.882–0.890), though all ChatGPT4o-mini configurations tended to perform slightly lower overall ($\kappa = 0.865$–0.874) (*H1b supported*).

Certain construct-specific patterns emerged when examining LLM within-configuration agreement (see Table S7 in Supplemental Materials). For example, Perceived Challenge demonstrated the overall strongest within-configuration reliability across all configurations, with κ values ranging from 0.831 to 0.893 and percent agreement from 92.80% to 96.05%. Self-efficacy, Interest, and Prior KSAs showed reasonably good within-configuration reliability, with κ values consistently above 0.75 across all configurations. For Self-efficacy, reliability ranged from $\kappa = 0.791$ to $\kappa = 0.831$, with ChatGPT4o/temperature=1 achieving the highest consistency ($\kappa = 0.831$, 92.78% agreement). Interest showed similar patterns, ranging from $\kappa = 0.756$ to $\kappa = 0.886$, with ChatGPT4o/temperature=0 demonstrating good reliability (κ = 0.886, 94.48% agreement). Prior KSAs maintained consistent performance across configurations ($\kappa = 0.796$ to $\kappa = 0.830$), with ChatGPT4o-mini/temperature=0.3 achieving the highest reliability ($\kappa = 0.830$, 91.6% agreement). Persistence presented the most variable within-configuration reliability patterns, with reliability ranging from $\kappa = 0.701$ to $\kappa = 0.818$. Notably, while Persistence showed lower κ values compared to other constructs, it achieved the highest percent agreement rates (93.9% to 97.15%), suggesting that disagreements occurred primarily in low-frequency coding decisions. The ChatGPT4o/temperature=1 configuration demonstrated the strongest within-configuration reliability for Persistence ($\kappa = 0.818$) as well as high agreement (94.82% agreement).

3.2. Pairwise Reliability Between LLM Configurations (RQ2)

Beyond self-consistency, we examined agreement between different LLM configurations to identify which combinations produced the most similar patterns of codes. Within-model type pairs showed 3.3–4.0% higher agreement than cross-model pairs at the same temperature (e.g., ChatGPT4o/temperature=0 vs. ChatGPT4o/temperature=0.3: $\kappa = 0.876$; ChatGPT4o/temperature=0 vs. ChatGPT4o-mini/temperature=0: $\kappa = 0.841$). This pattern held across all temperature levels, suggesting that differences in model type (in this case regular v. mini ChatGPT4o models) are likely to produce different coding outputs even with identical prompts and temperature settings (*H2a supported*).

Examining differences between model configurations at different temperatures revealed several interesting findings. Within each model type, agreement decreased monotonically as temperature difference increased (*H2b supported*). In addition, pairings consisting of model configurations with lower temperature settings consistently produced the highest inter-model reliability across most constructs. The ChatGPT4o/temperature=0 vs. ChatGPT4o/temperature=0.3 comparison yielded exceptional agreement for Perceived Challenge ($\kappa = 0.884$, 95.76% agreement) and Interest ($\kappa = 0.863$, 93.44% agreement). These findings suggest that deterministic settings may create more stable interpretative frameworks that could generalize across similar parameter configurations. Pairings with model configurations at mid-range temperature settings were overall less consistent though appeared to produce reasonably consistent coding across certain constructs. For example, Self-efficacy achieved its highest pairwise reliability in the ChatGPT4o/temperature=0.3 vs. ChatGPT4o/temperature=0.7 comparison ($\kappa = 0.798$, 90.8% agreement), while Prior KSAs performed best in ChatGPT4o/temperature=0.7 vs. ChatGPT4o/temperature=1 pairings ($\kappa = 0.741$, 88.06% agreement). These patterns align with our dimensional analysis, where constructs with higher theoretical coherence benefited from moderate temperature variation. In further contrast, pairings involving model configurations with higher temperature settings generally produced lower reliability, with notable exceptions for constructs with low operational clarity. Prior KSAs showed its strongest performance in ChatGPT4o-mini/temperature=0.7 vs. ChatGPT4o-mini/temperature=1 comparisons ($\kappa = 0.809$, 90.5% agreement), suggesting that abstract constructs may require higher temperature settings to capture their varied expressions.

3.3. Consistency with Human-Derived Codes (RQ3)

Evaluation of LLM-human alignment revealed substantial divergences across all constructs and configurations, highlighting fundamental differences in how humans conceptualize and LLMs identify these psychological constructs. For comparison purposes, human-human coding results revealed a generally strong and consistent pattern of coding the constructs among the human raters with the exception of Perceived Challenge (see Table S8 in Supplemental Materials). Specifically, Self-efficacy demonstrated reasonably good inter-rater reliability with a Cohen's kappa of 0.886 and 94.39% agreement, while Interest and Persistence also showed moderate agreement (see Landis & Koch, 1977), reflected in κ values of 0.675 and 0.514 respectively, both with over 84% agreement. Both human raters did not detect instances of Prior KSAs and thus achieved perfect 100% agreement, and thus κ could not be calculated. Follow-up conversations with human coders revealed that they adhered to the definition indicated in the coding instructions and regarded only in turns that referred to knowledge accumulated from experience outside of the task context as evidence of Prior KSAs. Similarly, there were relatively few instances of Perceived Challenge noted by one human coder and none noted by another human

coder and thus had 98.94% agreement with a $\kappa = 0$. The human coders indicated that unless there was a direct reference from a student or agent regarding the difficulty of an item or the task overall, Perceived Challenge was not assumed. For both these constructs, the human coders were much more discerning, relying on information explicitly conveyed in the turns within the task context rather than presumed evidence of these constructs based on linguistic features of the turns. Yet at the same time, human-human agreement (average $\kappa = 0.644$) was actually lower than the best LLM-LLM agreement (average $\kappa = 0.856$) across constructs. (*H3a partially supported*).

In contrast, while LLMs achieved high inter-model agreement for certain configurations, human-LLM alignment remained low (κ ranging from 0.000 to 0.419). The most striking finding was the lack of alignment in identifying Prior KSAs, where every human-LLM comparison yielded $\kappa = 0$ despite agreement ranging from 52.22% to 83.96%. This pattern indicates that humans and LLMs may be identifying specific answers to questions as evidence of prior knowledge, a breakdown reflecting the difficulties LLMs face in detecting the construct's context-dependent nature (*H3b supported*). While LLMs tended to code explicit knowledge statements such as those reflecting an answer to a question, human coders noted the subtle indicators such as domain-specific vocabulary usage or references to past experiences as indicators of prior KSAs, which were largely absent in the dataset.

Another notable pattern was the finding that codes for Interest also showed inconsistent identification patterns between human coders and LLMs, with multiple negative κ values (as low as $\kappa = -0.284$ for human1 vs. ChatGPT4o-mini/temperature=0). Negative κ indicates that LLM-human agreement was worse than chance, suggesting systematic misalignment in what constitutes the expression of interest. This pattern emerged across configurations: while LLMs achieved strong internal agreement (κ up to 0.863 between ChatGPT4o/temperature=0 and ChatGPT4o/temperature=0.3), human alignment remained poor (κ ranging from -0.284 to 0.198). This may indicate that the LLMs applied internally consistent but fundamentally incorrect heuristics for interest detection, possibly conflating engagement behaviors with genuine interest or misinterpreting task-focused language as affective expression.

Perceived Challenge showed near-zero κ values ($\kappa = 0.000$ to 0.012) when compared with both humans despite moderate percent agreement (62.99% to 70.17%). This pattern may indicate that while humans and LLMs agreed on the majority of instances where evidence of Perceived Challenge was lacking, they fundamentally disagreed on what constituted evidence in favor of the construct. This was true even as human coders themselves showed minimal agreement, suggesting the construct may be inherently difficult to operationalize consistently with this type of data and thus may point to the types of constructs that lack the definitional clarity for LLM coding to be appropriately used.

Persistence exhibited a similar pattern: high percent agreement (78.23% to 92.41%) but consistently low κ values ($\kappa = 0.043$ to 0.426). This suggests that while humans and LLMs agreed on most instances, they disagreed systematically on the positive cases influenced κ calculations. The pattern further varied by human coder: human1 showed moderate alignment ($\kappa = 0.246$ to 0.430), while human2 showed less consistent alignment ($\kappa = 0.043$ to 0.223). This inter-human variability suggests that persistence operationalization lacks clarity even among expert coders, making it particularly challenging for LLMs to detect consistent patterns.

Self-efficacy demonstrated the best human-LLM alignment among all constructs, with κ values ranging from 0.138 to 0.418. However, this "best" performance still represents poor to fair agreement by conventional standards. The pattern revealed systematic differences between human coders: human2 showed consistently lower alignment with LLMs ($\kappa = 0.160$ to 0.418) compared to human1 ($\kappa = 0.246$ to 0.419). Interestingly, non-mini models consistently

outperformed mini models in human alignment for self-efficacy. For example, human2 vs. ChatGPT4o/temperature=0 achieved $\kappa = 0.418$, while human2 vs. ChatGPT4o-mini/temperature=0 reached only $\kappa = 0.192$.

3.4. Construct-Specific Patterns (RQ4)

Across all constructs, mid-range temperature settings (0.3–0.7) yielded the most consistent and reliable outputs. Lower temperatures (0.0–0.3) produced more deterministic results, while higher settings (1.0) introduced variability that in some cases enhanced interpretive constructs but reduced overall agreement for more structured tasks. Mini models demonstrated notably different patterns compared to their regular-type counterparts, particularly at higher temperature settings. In Supplemental Materials, Table S9 reports the overall consistency (α) across all coders for each construct and Figures S1–S5 in Supplemental Materials illustrate the highest pairwise consistency (κ) for each construct. Constructs with high clarity (Self-efficacy: $\alpha = 0.580$) and moderate clarity (interest: $\alpha = 0.584$) showed higher agreement than low-clarity constructs (Persistence: $\alpha = 0.498$) (*H4a largely supported*), while low construct clarity appeared to have a more negative effect on agreement for human coders than LLMs (*H4b supported*).

Krippendorff's α indicated moderate reliability across all coders (including humans and LLMs) for self-efficacy ($\alpha \approx 0.580$ overall), reflecting some inconsistency in coding definitions across raters, particularly between LLMs and human coders ($\kappa = 0.160$–0.418). Despite this, Self-efficacy had the highest between-configuration reliability among all constructs, with optimal performance at moderate temperatures. The ChatGPT4o/temperature=0.7 configuration yielded the strongest reliability ($\kappa = 0.847$, 92.1% agreement), while mini models showed more variable performance across temperature settings. This pattern aligns with the construct's high clarity and specificity ratings, suggesting that well-defined constructs benefit from moderate temperature settings that balance determinism with interpretive flexibility.

Perceived challenge had the highest Krippendorff's α among constructs ($\alpha \approx 0.603$), indicating relatively strong internal consistency among LLM coders. However, human-LLM agreement was notably low ($\kappa \approx 0.00$–0.012), suggesting divergent interpretation criteria despite moderate percent agreement. Across configurations, perceived challenge showed high reliability, particularly in low-temperature settings. The ChatGPT4o/temperature=0 configuration achieved $\kappa = 0.823$ with 94.2% agreement, and ChatGPT4o/temperature=0.3 showed similar performance ($\kappa = 0.819$, 93.8% agreement). The construct's moderate clarity and more concrete manifestations appear to support stable LLM identification across runs.

Interest had relatively low Krippendorff's α (≈ 0.461), reflecting moderate reliability with notable temperature sensitivity. Human-LLM κ values were often negative or near zero, indicating significant misalignment in construct interpretation between humans and LLMs. Low-temperature configurations performed best, with ChatGPT4o/temperature=0 achieving $\kappa = 0.756$ (89.3% agreement). Higher temperatures introduced substantial variability, particularly in mini models, where ChatGPT4o-mini/temperature=1 dropped to $\kappa = 0.542$. This pattern may reflect the construct's moderate clarity but high context-sensitivity, requiring more deterministic settings for stable identification.

The Prior KSAs construct had moderate Krippendorff's α (≈ 0.482) but showed the most variable between-configuration reliability, ranging from $\kappa = 0.423$ to $\kappa = 0.734$ across configurations. Mini models at higher temperatures showed optimal performance (ChatGPT4o-mini/temperature=1: $\kappa = 0.734$, 87.1% agreement) for Prior KSAs, in contrast to other constructs. Despite surface-level agreement above 80%, all human-LLM comparisons resulted in $\kappa = 0.000$, indicating a complete divergence in what raters considered evidence of prior

knowledge. This aligns with the construct's low clarity rating and suggests that less deterministic settings may help capture its diverse manifestations.

Persistence had the lowest Krippendorff's α among constructs (α ≈ 0.438), indicating poor reliability despite high percent agreement rates. Human-LLM κ values were also low (max κ = 0.423), suggesting conceptual misalignment. The best-performing configuration (ChatGPT4o-mini/temperature=0.7) achieved only κ = 0.687 with 95.8% agreement, reflecting high surface-level consistency but substantial disagreement on positive cases. This pattern corresponds with the construct's low specificity rating and measurement challenges identified in our dimensional analysis. With a high concreteness and low clarity rating, coding for persistence reflected only moderate human agreement (κ = 0.514), suggesting that concreteness alone is insufficient when specificity and granularity are low (*H4c partially supported*).

3.4.1. Model type and temperature differences

Mini models consistently applied significantly more codes across all constructs, with particularly pronounced differences at higher temperatures. For instance, ChatGPT4o-mini/temperature=1 applied codes to 44.8% of turns for Self-efficacy compared to 31.7% for ChatGPT4o/temperature=1, and this pattern was consistent across constructs. While non-mini models generally performed best at low-to-moderate temperatures, mini models often achieved optimal reliability at higher temperature settings. For Prior KSAs, the highest pairwise reliability occurred between ChatGPT4o-mini/temperature=0.7 and ChatGPT4o-mini/temperature=1 (κ = 0.809, 90.5% agreement), whereas non-mini models peaked at ChatGPT4o/temperature=0.7 vs. ChatGPT4o/temperature=1 (κ = 0.793, 92.11% agreement). Similarly, for Persistence, mini models achieved their highest reliability at ChatGPT4o-mini/temperature=0.7 vs. ChatGPT4o-mini/temperature=1 (κ = 0.752, 96.99% agreement). When comparing mini to non-mini models at identical temperature settings, agreement dropped substantially, particularly at higher temperatures. For example, Self-efficacy showed strong within-type agreement (ChatGPT4o/temperature=0.7 vs. ChatGPT4o/temperature=1: κ = 0.803; ChatGPT4o-mini/temperature=0.7 vs. ChatGPT4o-mini/temperature=1: κ = 0.810), but cross-type comparisons at the same temperatures yielded much lower agreement (ChatGPT4o/temperature=0.7 vs. ChatGPT4o-mini/temperature=0.7: κ = 0.462; ChatGPT4o/temperature=1 vs. ChatGPT4o-mini/temperature=1: κ = 0.464). This pattern suggests that model type introduces systematic interpretation differences that become more pronounced as temperature increases, potentially reflecting different approaches to handling ambiguous or context-dependent expressions of psychological constructs.

3.4.2. Patterns by Construct Dimensions

Analysis of construct performance patterns in relation to our dimensional framework (see Table S1) revealed potential relationships between construct characteristics and LLM coding reliability. These patterns may provide insights into which types of psychological constructs are most amenable to qualitative coding with certain LLM configurations. With respect to the quality of the *operational definition* of the constructs, those rated high in clarity, such as Self-efficacy, consistently achieved higher reliability across most metrics. For example, Self-efficacy demonstrated κ values up to 0.418 in human-LLM comparisons and maintained robust LLM-LLM agreement (κ up to 0.846 among mini model configurations). In contrast, constructs with low clarity, such as Prior KSAs, showed universally poor performance, with all human-LLM κ values equal to zero despite percent agreement exceeding 80% in some cases. Constructs with moderate clarity, including Persistence, Interest, and Perceived Challenge, exhibited

intermediate reliability and notable sensitivity to temperature settings, as reflected in the variability of κ values across configurations (e.g., Interest κ values ranging from –0.284 to 0.198; Perceived Challenge κ near zero for human-LLM pairs). With respect to the specificity of the construct, Self-efficacy, which had higher specificity tended to have more consistent ratings across configurations, with minimal variability in reliability metrics. In contrast, Persistence, rated low in specificity, exhibited high variability in both reliability and human alignment. For instance, Persistence κ values for human-LLM pairs ranged from 0.043 to 0.426, and percent agreement, while high (up to 92.41%), did not translate to interpretive consistency. These results suggest that clear construct boundaries are essential for reliable automated coding, as ambiguity in operational definitions leads to inconsistent model behavior and poor alignment with human judgment.

With respect to the empirical grounding of the constructs, those thought to have higher concreteness (i.e., Persistence, Self-efficacy, Prior KSAs) generally achieved high percent agreement between raters. For example, Persistence reached up to 92.41% agreement. However, these constructs also showed variable κ values, with Persistence κ values ranging widely. On the other hand, Perceived Challenge, which had only a moderate concreteness rating, had a percent agreement which ranged from 62.99% to 70.17% for human-LLM pairs with κ values near zero. This pattern indicates that while observable manifestations of constructs support surface-level consistency, they may not guarantee interpretive accuracy or conceptual alignment between humans and LLMs.

Finally, we considered how differences in the reliability of constructs could be partly explained by the extent of the conceptual integrity of the construct. In particular, constructs with high theoretical coherence, such as Self-efficacy, demonstrated more stable performance patterns across LLM configurations and human comparisons. Self-efficacy maintained moderate to high κ values and percent agreement, while constructs with only moderate theoretical coherence, such as Interest and Perceived Challenge, showed greater sensitivity to parameter variations and less consistent reliability. This suggests that well-established theoretical frameworks provide a more reliable foundation for automated coding, reducing the impact of model and parameter selection on coding outcomes.

The observed variation in reliability across different tasks reflects differences in the conceptual integrity of the constructs. Constructs with higher theoretical coherence and clearer operational definitions, such as Self-efficacy and Perceived Challenge, generally maintained more stable reliability across sections, particularly in non-mini models. For example, Perceived Challenge in ChatGPT4o/temperature=0.3 showed high and consistent κ values (0.804–0.975) regardless of the task, indicating that its behavioral markers were robust to contextual changes within the assessment. In contrast, constructs with lower conceptual integrity, those that are less clearly defined or more context-dependent, such as Persistence and Interest, exhibited greater fluctuations in reliability across sections, especially in mini models. For instance, Persistence κ values in ChatGPT4o-mini/temperature=0.3 ranged widely (0.359–0.992), suggesting that the model's ability to identify a construct such as Persistence was highly sensitive to the specific demands or discourse patterns of each task. These patterns indicate that constructs with strong theoretical and operational foundations are less affected by contextual variation, resulting in more reliable automated coding across different parts of the task. Conversely, constructs that are more ambiguous or context-sensitive show greater variability, highlighting the importance of both construct clarity and context consideration when applying LLM-based coding in educational settings. In addition, we noted that non-mini models generally tended to outperform mini models for constructs with higher clarity and specificity (esp. Self-efficacy), while mini models

appeared to more reliably code constructs with somewhat lower operational clarity (esp. Prior KSAs).

We further note that LLMs demonstrated equal or higher self-consistency than consistency between human coders across all constructs. These findings may suggest divergent but internally consistent interpretive frameworks for identifying certain constructs. For example, when constructs lack operational clarity, both human and LLM coders may develop stable coding schemes, even though these schemes are likely to differ. Notably, Perceived Challenge showed zero human-human agreement ($\kappa = 0.000$) despite high LLM self-consistency ($\kappa = 0.831$–0.893), indicating that even trained human coders could not establish a shared interpretation of a relatively ambiguously defined construct. At the same time, we note with emphasis that while the LLMs tended to demonstrate greater internal consistency, there remains the important question of whether such consistency matters if the application of definitions and labels are not theoretically aligned with the underlying construct.

3.5. Predictive Validity: Correlations with Task Performance (RQ5)

To address RQ5, we examined the extent to which LLM-identified expressions of Persistence and related constructs predicted students' performance on the ELLA-Math assessment (Lopez et al., 2021). Correlations were computed between each construct (Persistence, Self-efficacy, Interest, Perceived Challenge, and Prior KSAs) and students' total scores, as well as their performance on each of the four assessment sections. Examination of relationships between construct coding frequency and performance on the task revealed certain patterns. Overall correlations between construct counts and student performance showed that all five constructs exhibited negative correlations, indicating that higher frequencies of these constructs were generally associated with lower performance outcomes.

The frequency of codes for Perceived Challenge appeared to have the strongest negative correlation with performance across all coders ($r = –0.25$, $p < .001$), followed by Persistence ($r = –0.19$, $p < .001$). These findings align with theoretical expectations, as expressions of challenge and continued struggle may indicate students encountering difficulty rather than productive engagement (*H5a supported*). Self-efficacy ($r = –0.15$, $p < .001$) and Prior KSAs ($r = –0.14$, $p < .001$) showed moderate negative correlations (*H5b not supported*), while Interest exhibited the weakest association ($r = –0.08$, $p < .001$) (*H5c not supported*).

3.5.1. Variation by Task

The relationship between constructs and task performance varied dramatically across different sections of the assessment, revealing the context-dependent nature of these motivational indicators. In Section 2 of the task, both Self-efficacy ($r = 0.33$, $p < .001$) and Interest ($r = 0.31$, $p < .001$) showed strong positive correlations with performance, suggesting that early expressions of confidence and engagement were predictive of success. However, this pattern reversed in later sections, particularly Section 5, where Perceived Challenge, which was almost solely identified by the LLMs ($r = –0.23$, $p < .001$), and Persistence ($r = –0.15$, $p < .001$) demonstrated significant negative associations with performance. This shift suggests that continued expressions of challenge and persistence in later sections may indicate unproductive struggle rather than adaptive engagement. As shown in Figure 1, notable shifts from positive correlations in early sections to negative correlations in later sections highlight the context-dependent nature of motivational constructs during problem-solving.

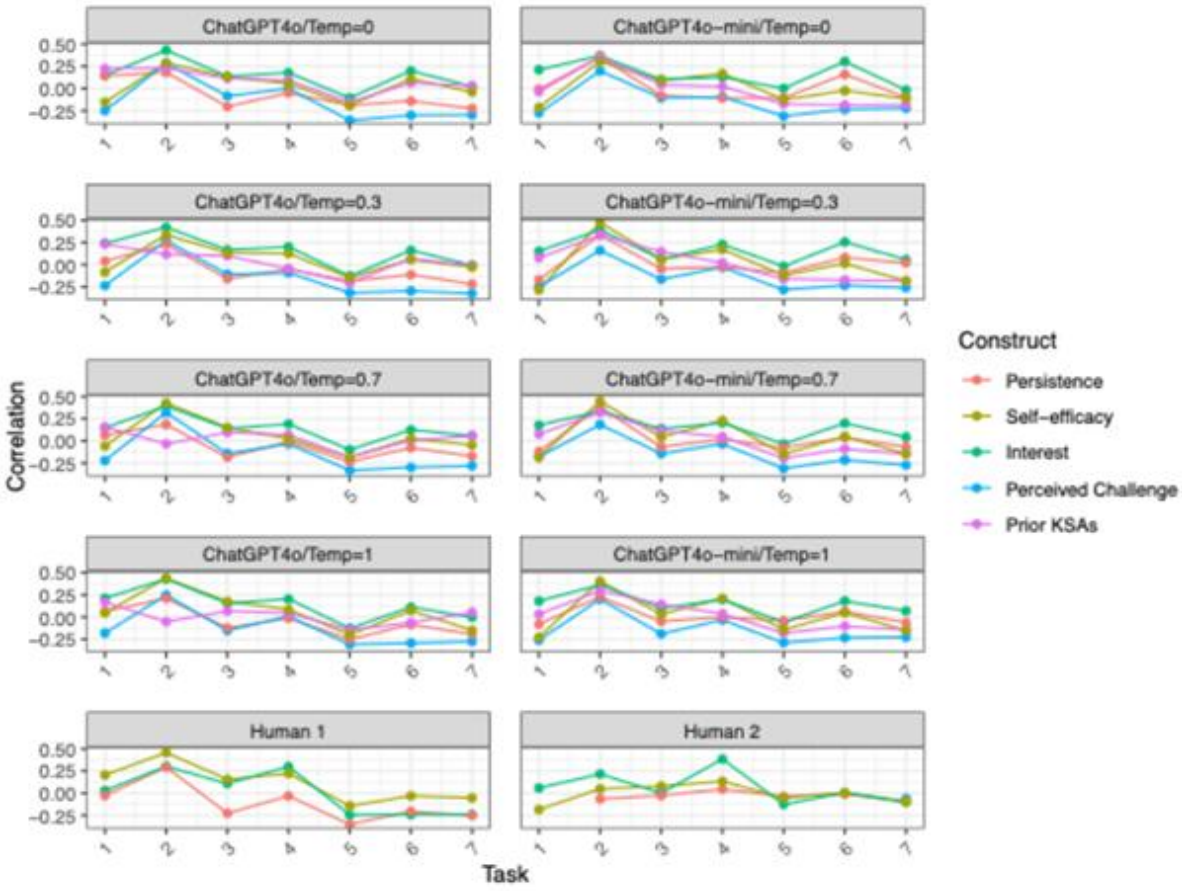

Figure 1: Line graphs showing correlations between construct counts and performance across tasks.

3.5.2. Rater-specific Patterns

Correlations with task performance varied notably between human and LLM coders, with human coders showing more conservative coding patterns. For instance, Persistence coded by human1 demonstrated a strong negative correlation ($r = –0.32$, $p < .001$), while the same construct coded by LLM model ChatGPT4o/temperature = 0.3 showed a similar but slightly weaker pattern ($r = –0.25$, $p < .001$). Figure 2 illustrates differences between the strength of the association between the frequency of certain codes on students' messages and students' scores between different coders, where darker colors indicate stronger negative correlations, with Perceived Challenge and Persistence showing the most consistent negative associations across raters. Notably, correlations are not shown for human coders with respect to certain constructs (shown in gray cells in Figure 2), reflecting the lack of codes assigned by the human coders for those constructs.

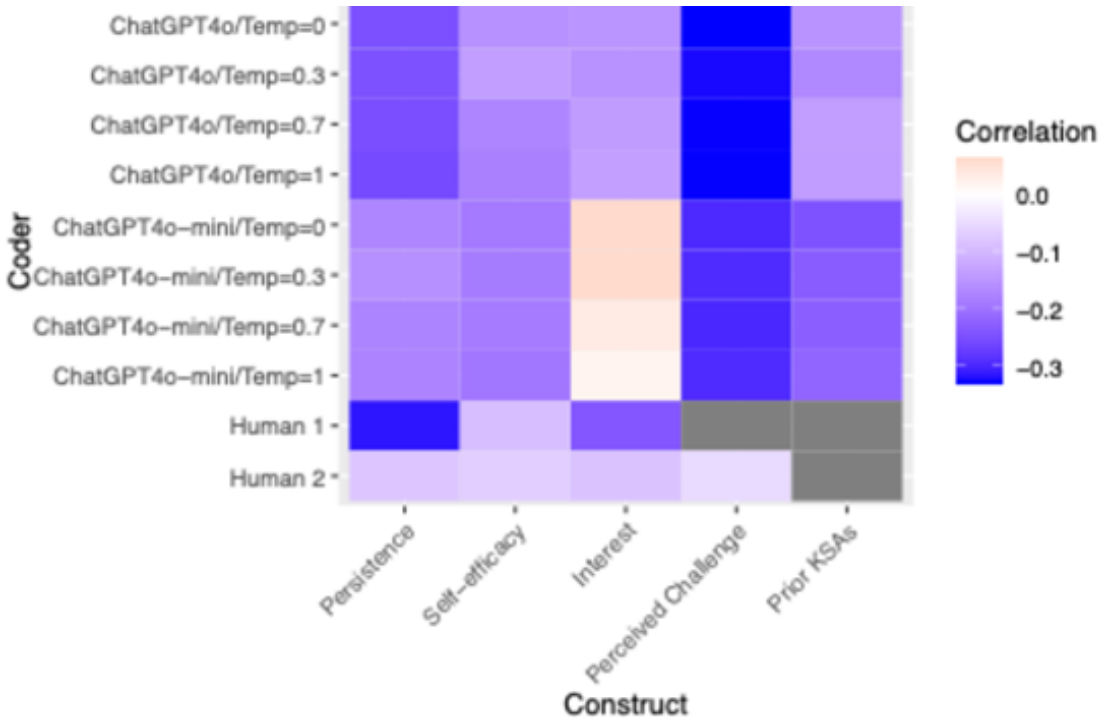

Figure 2: Heat map displaying correlations between construct counts and student performance across different raters and LLM configurations. Note that the gray boxes reflect the lack of codes.

3.5.3. Temperature and Model Effects

Different LLM temperature settings revealed varying sensitivity to construct-performance relationships (see Table S10 in Supplemental Materials). For Perceived Challenge, correlations remained consistently strong across temperature settings, ranging from $r = –0.30$ (ChatGPT4o-mini/temperature=0) to $r = –0.33$ (ChatGPT4o/temperature=0 and ChatGPT4o/temperature=1), all significant at $p < .001$. Persistence correlations showed more variation, with ChatGPT4o/temperature=1 producing the strongest association ($r = –0.26$, $p < .001$) compared to ChatGPT4o-mini/temperature=0.3 ($r = –0.16$, $p < .001$). Notably, Interest showed the most temperature-dependent patterns: while ChatGPT4o/temperature=0.3 maintained a negative correlation ($r = –0.16$, $p < .001$), the ChatGPT4omini models at temperatures 0.3, 0.7, and 1.0 all produced positive correlations ($r = 0.06$, $r = 0.04$, and $r = 0.02$, respectively), though with reduced significance. Self-efficacy correlations were relatively stable across temperatures, ranging from $r = –0.14$ to $r = –0.19$, suggesting this construct's relationship with performance was less sensitive to temperature variations than Interest or Persistence.

We subsequently examined the coding of persistence in greater detail to understand how construct ambiguity might affect human-LLM agreement. Table 3 presents the distribution of persistence codes by coder type. Human coders differed substantially in how frequently they identified persistence, suggesting different thresholds for what constitutes evidence of the student exerting effort in the face of obstacles. Notably, all coders found negative correlations between Persistence codes and task performance (as shown in Table S10 in Supplemental Materials), potentially highlighting a connection between task difficulty experienced and the demonstration of persistence. This indicates that despite disagreement on specific instances, all coders captured a meaningful construct related to task difficulty. For example, Human 1

conveyed that they tended to treat ambiguous statements such as those beginning with "I think…" as evidence of persistence interpreting uncertainty as continued engagement. In contrast, Human 2 expressed requiring explicit expressions of effort such as "let me try again" as evidence of persistence. Meanwhile, we noted that the LLMs appeared to consistently code questions posed to agents as indicating persistence, thus interpreting help-seeking as evidence of persistence. These patterns may help to explain the low overall average agreement across all coders ($\alpha = 0.498$) as each coder applied a coherent but distinct interpretation of the construct's boundaries.

Table 3: Coding of persistence by coder type.

Coder / LLM Configuration		Total Persistence Codes	% of All Codes	*Mean* Persistence Codes per Student	*SD* Persistence Codes per Student	Correlations between Count of Persistence Codes and Performance
Model Type	**Temperature**					
ChatGPT4o	0	1191	17.5%	11.34	7.71	$r = –0.25$, $p < .001$
ChatGPT4o	0.3	1251	18.5%	11.91	8.26	$r = –0.25$, $p < .001$
ChatGPT4o	0.7	1205	18.1%	11.48	7.46	$r = –0.25$, $p < .001$
ChatGPT4o	1	1174	17.4%	11.18	7.71	$r = –0.26$, $p < .001$
ChatGPT4o-mini	0	355	5.9%	3.38	2.32	$r = –0.17$, $p < .001$
ChatGPT4o-mini	0.3	407	6.5%	3.88	2.25	$r = –0.16$, $p < .001$
ChatGPT4o-mini	1	400	6.3%	3.81	2.33	$r = –0.18$, $p < .001$
ChatGPT4o-mini	0.7	471	7.4%	4.49	2.66	$r = –0.18$, $p < .001$
Human 1	–	1498	30.5%	14.27	6.72	$r = –0.32$, $p < .001$
Human 2	–	164	10.4%	1.56	2.38	$r = –0.08$, $p = .002$

Note: Human 2 only coded part of the transcript (see Section 2.2) and thus total counts may be misleading.

4. DISCUSSION

This study investigated the potential of LLMs to automatically identify key psychological and behavioral constructs, particularly persistence, from conversational data generated in a CBA with middle school English learners. By systematically varying LLM configurations, including model type and temperature settings, and comparing outputs to human-coded benchmarks, we sought to evaluate the reliability, accuracy, and construct-specific patterns of LLM-generated codes. Our findings contribute to the development of scalable methods for leveraging LLMs to detect constructs associated with learning, offering a promising framework for refining CBAs and advancing real-time educational analytics. Furthermore, this work is among the first to map construct characteristics (see Table S1 in Supplemental Materials; see also Liu et al., 2025) to LLM configurations, providing actionable guidance for researchers for interpreting LLM outputs based on construct characteristics.

Preliminary analyses revealed several important trends. Higher temperature settings, which introduce greater randomness in LLM output, were associated with a larger number of generated codes but often at the expense of reliability. More abstract or context-dependent constructs, such as Prior KSAs, were less frequently identified at lower temperature settings, which tend to produce more deterministic outputs. In contrast, more concrete, behaviorally anchored constructs like Persistence demonstrated relative stability across temperature variations. These patterns align with several of the characteristics of the constructs: those with moderate-to-high clarity and specificity (i.e., Persistence, Self-efficacy, Interest, and Perceived Challenge) generally

achieved overall higher reliability, though the temperature settings that produced the most reliable set of codes varied depending on the construct.

Though many of our hypotheses were supported, several results yielded unexpected findings. In Supplemental Materials, Table S11 provides a summary of the key findings and Table S12 reports key findings with respect to each hypothesis. First, we found that temperature alone had minimal effects (*H1a partially supported*). We predicted differences based on model temperature settings, though observed little variation in agreement within LLM as a function of temperature setting alone (Tables S6 and S7 in Supplemental Materials). With that said, we noted that the mini models varying by temperature appeared to produce a greater range of percent agreement across all constructs (89.55% to 97.15% agreement) than did the regular models (91.72% to 96.05% agreement; see Table S7 in Supplemental Materials). For structured coding tasks with explicit rules, prompt constraints may override temperature settings and perhaps the coding framework provided sufficient guidance that even high-temperature randomness could not substantially alter outputs. In addition, we found that human-LLM agreement exceeded human-human agreement for some constructs (*H3a not supported*). Perceived Challenge showed zero human-human agreement ($\kappa = 0$) but high LLM-LLM agreement ($\kappa = 0.831$-0.893). Furthermore, all correlations with performance were negative (*H5b not supported*) though we had predicted positive correlations for Self-efficacy and Prior KSAs based on established theory (Bandura, 1977; Shapiro, 2004). These negative correlations shown in Table S10 in Supplemental Materials instead may indicate that in conversational assessments, expressions of these constructs are more likely to occur when students are struggling. Students appeared to use expressions of Self-efficacy or reference Prior KSAs when they were uncertain, a function of dialogue that differs from traditional self-report measures.

Despite strong agreement between different LLM configurations, alignment between LLM-generated codes and those generated by the human coders was consistently low, with Cohen's kappa values near zero across most constructs and configurations. This finding underscores the broader challenge of measuring complex, context-sensitive constructs in educational settings and suggests that LLM-based coding may inherit or even amplify existing problems in construct operationalization and measurement consistency. Notably, constructs with generally lower ratings for clarity, specificity, and granularity of the operational definition (i.e., Persistence, Interest, Perceived Challenge, and Prior KSAs) showed particularly low human-LLM alignment (κ ranging from 0 to 0.152), whereas Self-efficacy, with higher coders, exhibited the highest alignment ($\kappa = 0.419$). This finding may reflect the fact that humans tend to apply rule-based systems for coding, in which the rationale of coding in a certain way is explicit, while in contrast, the LLMs apply codes based on statistical probabilities (Bauer & Zapata-Rivera, 2020). Such fundamental underlying differences may reflect how different constructs are identified and what constitutes evidence of the constructs. Scenarios where LLMs tend to achieve high internal consistency but low human alignment in qualitative analysis should not necessarily be interpreted as evidence of inadequacy of one approach over another. Instead, such patterns may serve as signals that the construct operationalization itself requires further refinement.

Temperature sensitivity also appeared related to theoretical coherence and clarity; constructs with moderate theoretical coherence but lower clarity (Persistence, Prior KSAs) benefited from higher temperature settings, while those with moderate clarity (Interest, Perceived Challenge) performed better at lower temperatures. Correlations between LLM-generated codes and task performance further complicated interpretation. The context-dependent nature of these correlations suggests that Persistence, as detected by LLMs, may not always indicate productive engagement but could instead reflect ongoing student difficulty or confusion. This raises critical

questions about the construct validity of automated coding and highlights the need to distinguish adaptive from maladaptive forms of persistence in future analyses.

The low consistency in measuring our chosen concepts, as noted in existing literature (Alexander, 2003; Reninger & Hidi, 2011), was reflected in our low human-LLM agreement. This supports recent research indicating that LLMs, while promising, struggle with nuanced, culturally specific, or abstract concepts that require deep contextual understanding (Kuzman & Ljubešić, 2025; Törnberg, 2025). For example, Kuzman and Ljubešić (2025) found that while LLMs can perform well on some coding tasks, they show significant limitations when dealing with more abstract concepts. Similarly, Törnberg (2025) highlighted challenges in using LLMs for coding content that requires a deep knowledge of the social, cultural, and political aspects of a given context. In contrast, human coders tended to have greater discernment in applying codes for constructs based on the task context, yet even between the two human coders there was still considerable inconsistency as the human coders occasionally interpreted coding rules in different ways. The human coders were also able to provide a rationale for their coding decisions, even when not applied consistently, which in turn could be used to clarify and refine the coding scheme. Meanwhile, though the LLMs were prompted to provide a justification for applying certain codes, it remains unclear whether these actually reflect the underlying process behind the statistical model (Turpin et al., 2023).

Furthermore, the "black-box" nature of LLMs raises concerns about reproducibility and transparency, especially considering research pointing to the variability in their outputs based on different prompts and parameter settings. In particular, previous research has found substantial variability in LLM performance across different prompting strategies (Liu et al., 2025), with some research even finding contradictory results with identical datasets as a result of different prompting strategies (Anagnostidis & Bulian, 2024). Our observation that temperature settings affect coding reliability further confirms that model tuning is crucial for balancing interpretive richness and predictive accuracy.

While persistence is key to self-regulated learning (Skinner et al., 2020; Zimmerman, 2002), frequent expressions of these concepts in student conversations might actually signal unproductive struggle rather than positive engagement (Eccles & Wigfield, 2020; Pintrich & De Groot, 1990). This distinction is vital for designing adaptive learning systems. Our results suggest that LLM-based coding of persistence and related factors is quite nuanced and may require greater involvement of human raters in order to be carefully interpreted within the specific task and learner context. Each task's section-specific patterns also warrant investigation of how construct expressions evolve throughout a task or during problem-solving steps and whether different coding approaches might better capture adaptive versus maladaptive manifestations of these constructs over different timespans.

4.1. Implications for Human-LLM Collaboration

This study highlights the potential promise and necessary cautions when using LLMs for automated coding of constructs from real-world conversational data. In particular, it illustrates how important clear construct definitions, consistent measurement, and context sensitivity are for both human and AI coding. Furthermore, our findings may emphasize that training humans and machines for qualitative coding are fundamentally different, requiring careful optimization of LLM settings and prompts to improve reliability and validity. Human coding differs fundamentally from LLM coding in several ways (see Bauer & Zapata-Rivera, 2020). Humans engage in an iterative, reflective process, asking clarifying questions, discussing ambiguous cases, and adjusting their understanding of constructs dynamically. The human coders in the present study

expressed that they benefited from collaborative dialogue and shared understanding, which helps resolve disagreements and refine coding schemes over time. This interactive calibration enables human coders to incorporate contextual nuances, cultural knowledge, and subtle linguistic cues that are difficult for LLMs to capture. This social and discursive dimension of human coding is absent, or at least not easily captured, in the LLM coding process, where each output is generated independently without iterative feedback or negotiation. In contrast, the LLM coding procedure used in the present study involved a one-pass, prompt-driven process that may not have included the necessary steps for clarification or consensus-building. Consequently, human coding tends to be more flexible and context-sensitive, whereas LLM coding offers scalability and speed but may struggle with ambiguous or context-dependent constructs. While LLMs can process large volumes of data rapidly and consistently, they rely heavily on their training data, the quality of the prompt, and model parameters, and they may produce variable outputs depending on temperature settings and prompt phrasing.

Despite these differences, at the onset our study's design sought to bridge these differences by aligning the prompts and coding framework as closely as possible, enabling a comparison of human and LLM coding performance. Our findings may also challenge an obvious conclusion that the human-LLM disagreement is evidence of inadequacy of using LLMs for qualitative coding, with three patterns of findings supporting an alternative explanation. First, LLM consistency exceeded human consistency for relatively poorly-defined constructs. Persistence (with a low operational clarity rating) showed human-human $\kappa = 0.514$ but LLM-LLM $\kappa = 0.701$–0.818. This suggests LLMs apply coding rules more uniformly than humans, even when those rules are ambiguous. Second, the clarity of the construct's operational definition appeared to be related to overall agreement. Codes for Self-efficacy ($\alpha = 0.580$) and Perceived Challenge ($\alpha = 0.732$) showed higher overall reliability than Persistence ($\alpha = 0.498$) and Prior KSAs ($\alpha = 0.544$), aligning with their operational clarity ratings (see Table S1 in Supplemental Materials). This pattern would not likely emerge if LLM errors were random or entirely construct-independent. Third, prompt refinement improved human-LLM alignment differentially by construct. Between Version 1 and Version 2 of coding instructions (see Appendix C in Supplemental Materials), we added specific behavioral indicators for Persistence ("re-attempting, asking for clarification, or expressing a desire to keep working"). However, agreement remained moderate, suggesting that prompt engineering cannot fully compensate for fundamental construct ambiguity. These findings have important implications for educational measurement. Rather than viewing LLMs as inadequate substitutes for human coders, researchers could recognize their use as diagnostic tools that reveal construct operationalization problems. When LLMs show high self-consistency but diverge from humans, this signals the need for better construct definition rather than better prompts.

Our findings also point to several areas for improvement that could enhance the consistency, transparency, and interpretability of LLM-generated codes. These include refining prompt specificity, clarifying coding definitions, and leveraging examples to reduce ambiguity in category boundaries. Furthermore, incorporating iterative prompt testing and feedback loops may further align LLM outputs with human judgment. Such refinements are critical for advancing the reliability of human–LLM collaboration in qualitative data analysis, especially when applied at scale or in high-stakes educational contexts. By using predefined codes as a starting point for coding as well as including human review and validation, educational researchers and practitioners can achieve more efficient, consistent, and transparent qualitative analyses. This collaboration may advance theoretical models of persistence and self-regulated learning by identifying linguistic signs of unproductive struggle and clarifying how these concepts appear in authentic student dialogue (see Graesser & McNamara, 2010).

Our findings point to several potential opportunities for humans and LLMs to collaborate and further identify constructs such as persistence in complex conversational data. LLMs can quickly process large amounts of conversation data, generate initial codes, and flag potential instances of specific concepts. Human experts, with their deep understanding of context, culture, and subtle meanings, can then review, refine, and interpret these initial outputs, focusing their expertise where it's most needed. Such collaboration could make manual coding less labor-intensive, improve consistency by giving humans a starting point for coding, and make the coding process more transparent by documenting the human-AI interaction (Chew et al., 2023; Dunivin, 2025). Human coders can also guide LLM tuning, adjusting settings like temperature and prompt design to get better coding quality for specific concepts. Our findings on temperature sensitivity and how different concepts are coded offer practical advice for setting up human-LLM workflows. For example, more deterministic LLM settings might be better for concepts where precise prediction is key, while more exploratory settings could help human coders discover new themes. Training humans versus training machines for qualitative coding is fundamentally different, and decisions concerning the prompt as well as the model configuration are non-trivial. Future research should explore interactive coding platforms that allow humans and AI to work together in real-time, clarifying codes and reaching consensus, much like human coding teams do. These findings may also point to the need for an iterative refinement process for improving both human calibration and automated classification performance.

With respect to the identification of complex and operationally ambiguous constructs such as persistence, our study may point to several actionable steps. For the constructs of persistence specifically, developing task-specific taxonomies that distinguish types of persistent behavior (e.g., strategic re-attempting vs. repetitive trial-and-error) rather than treating persistence as a unitary construct. For example, in Appendix D (Supplemental Materials), we present one possible taxonomy based on the findings and through discussions with coders from the present study.

4.2. LIMITATIONS

Several factors limit how broadly our findings can be applied. The qualitative nature of this study and the specific dataset mean our results might not apply to all situations. Future research should include larger, more diverse groups of students and different subject areas. The short and often minimal student responses may have made coding challenging for both humans and LLMs. Furthermore, given the population the tool was designed to serve includes English language learners, the students may have been more likely to use non-conventional language. Additionally, the data used in this study was derived from a conversational system reflecting less dynamic conversational exchanges than between humans or between humans and current LLM-powered conversational systems, which may facilitate richer exchanges.

Nearly all constructs in the present study received relatively low consistency estimates across coders. This issue appears to be particularly reflected in overall low human-LLM alignment. For example, kappa values for human-LLM comparisons were consistently zero or near zero across all constructs and configurations. This finding highlights the broader challenge of construct measurement in educational contexts and suggests that LLM-based coding applications may inherit and even amplify existing challenges in measuring such constructs. For example, it remains difficult to separate highly related concepts, especially without very clear definitions provided to coders. Future work should focus on developing clearer operational definitions and better training for coders to improve how well different concepts can be distinguished (Miles et al., 2014). High percent agreement in such cases may reflect agreement on the dominant (often

negative) category, rather than true shared understanding of the construct (Feinstein & Cicchetti, 1990). This phenomenon, often referred to as the "prevalence paradox," occurs when the majority of coded instances fall into a single category, typically the absence of a behavior or construct. In these situations, both raters (or human or LLM) are likely to agree simply because most responses are negative, not because they are applying the construct in the same way when it is present. As a result, percent agreement becomes inflated and does not accurately reflect the reliability of identifying the less frequent, but theoretically meaningful, positive cases. The literature demonstrates that Kappa is sensitive to this imbalance: it discounts agreement on the dominant category and places greater weight on agreement for the less common, positive cases, which are more informative for construct validity (Gwet, 2008). Thus, a high percent agreement alongside a low Kappa value signals that the observed consistency is superficial and may mask substantial differences in how raters interpret and apply the construct, especially in the critical instances where its presence is most relevant.

We would also like to note that the factors which contributed to the study's primary limitations that affected the consistency of codes may also serve as a contribution: we deliberately examined constructs with varying operational clarity, including Persistence, which has well documented definitional variability (Sparks et al., 2025). While this decision likely had a negative impact on our agreement and reliability statistics, producing values lower than studies focusing solely on well-defined constructs, our findings may also provide somewhat realistic estimates of LLM performance across the range of constructs educational researchers may actually be interested in identifying in naturalistic learning contexts. Rather, these findings demonstrate that ambiguity in the operationalization of constructs affects all coders, and that LLMs make this ambiguity more visible through their consistent application of whatever interpretative framework being derived and applied from the prompt.

4.3. Future Directions

Future research should address biases and context-specific challenges to improve the reliability and validity of LLM-driven qualitative analyses. One potential direction is to separately explore how the codes derived based on expressions from pedagogical agents relate to students' own expression of certain constructs, as well as performance and learning, and how such factors vary throughout the task as evidenced by multimodal data sources including process data (see Bernacki, 2018). With regard to methodological improvements, processing smaller data segments in batches (Lin et al., 2023), using multiple LLM agents for coding as in a multi-agent system (Rasheed et al., 2024), or only coding instances where the LLM is highly confident (Li et al., 2024), could potentially help mitigate errors due to hallucinations and biases, and ultimately improve accuracy. Integrating different types of data, including response time and physiological measures, could also enrich how concepts are measured and interpreted (D'Mello & Graesser, 2012). Further exploration is needed into how concept expressions change during problem-solving, with a focus on distinguishing between helpful and unhelpful forms of persistence. Finally, developing transparent and understandable frameworks for AI-assisted qualitative coding will be crucial for wider adoption and trust in these new methods.

4.4. Conclusion

This study contributes to a growing understanding of qualitative approaches in educational research by exploring the capabilities and limitations of LLMs in identifying complex psychological constructs from complex conversational data. While our findings highlight challenges in achieving direct human-LLM alignment, they underscore the important role of human

discernment in human-LLM collaboration. Further research in this area may help to identify new avenues for analyzing rich educational data, refining assessment tools, and gaining deeper, more actionable insights into student learning processes that ultimately support learners.

5. Acknowledgments

This work is supported by the National Science Foundation and the Institute of Education Sciences under Grant #2229612. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the U.S. Department of Education. We would like to acknowledge the time, effort, and thoughtful discussions contributed by Cheryl Van Ness and James Rice during the coding process and in shaping the prompt revision strategy.

6. Supporting Materials

The supplemental materials referenced in this document are available in the Open Science Framework repository here: https://doi.org/10.17605/osf.io/s85ck.

7. Use of Generative AI

During the preparation of this work, the author(s) used ChatGPT4o by OpenAI as described in the Methods section in order to code for key constructs and was also used to refine the descriptions of the construct dimension and sub-dimension labels. In addition ChatGPT4o was used to render an initial version of Table S11 in Supplemental Materials, which presents a summary of key findings. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Ainley, M., and Hidi, S. 2014. Interest and enjoyment. In *International handbook of emotions in Education*, L., Linnenbrink-Garcia and R. Pekrun, Eds., Routledge, 205–227. https://doi.org/10.4324/9780203148211

Ainley, M., Hidi, S., and Berndorff, D. 2002. Interest, learning, and the psychological processes that mediate their relationship. *Journal of Educational Psychology, 94*(3), 545–561. https://doi.org/10.1037/0022-0663.94.3.545

Alexander, P. A. 2003. The development of expertise: The journey from acclimation to proficiency. *Educational Researcher, 32*(8), 10–14. https://doi.org/10.3102/0013189X032008010

Anagnostidis, S., and Bulian, J. 2024. How susceptible are LLMs to influence in prompts? *arXiv preprint* arXiv:2408.11865. https://doi.org/10.48550/arXiv.2408.11865

Bandura, A. 1977. Self-efficacy: toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191–215. https://psycnet.apa.org/doi/10.1037/0033-295X.84.2.191

Bandura, A. 2006. Toward a psychology of human agency. *Perspectives on Psychological Science, 1*(2), 164–180. https://doi.org/10.1111/j.1745-6916.2006.00011.x

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. AIED 2024*, A. M. Olney, I. A. Chounta, Z. Liu, O. C. Santos, and I. I. Bittencourt, Eds., Lecture Notes in Computer Science, vol 14830, Springer, Cham, 134–149. https://doi.org/10.1007/978-3-031-64299-9_10

BATTLE, E. S. 1965. Motivational determinants of academic task persistence. *Journal of Personality and Social Psychology, 2*(2), 209–218. https://doi.org/10.1037/h0022442

BAUER, M. I., AND ZAPATA-RIVERA, D. 2020. Cognitive foundations of automated scoring. In *Handbook of automated scoring: Theory into Practice*, D. Yan, A. A. Rupp, and P. W. Foltz, Eds., CRC Press, 13–28.

BERNACKI, M. L. 2018. Examining the cyclical, loosely sequenced, and contingent features of self-regulated learning: Trace data and their analysis. In *Handbook of self-regulation of learning and performance* (2nd ed.), B. J. Zimmerman and D. H. Schunk, Eds., Routledge/Taylor & Francis Group, 370–387. https://psycnet.apa.org/doi/10.4324/9781315697048-24

BERNACKI, M. L., NOKES-MALACH, T. J., AND ALEVEN, V. 2015. Examining self-efficacy during learning: Variability and relations to behavior, performance, and learning. *Metacognition and Learning* 10, 99–117. https://doi.org/10.1007/s11409-014-9127-x

BORSBOOM, D., MELLENBERGH, G. J., AND VAN HEERDEN, J. 2004. The concept of validity. *Psychological Review, 111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

BOTELHO, A., BARAL, S., ERICKSON, J. A., BENACHAMARDI, P., AND HEFFERNAN, N. T. 2023. Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. *Journal of Computer Assisted Learning*, *39*(3), 823–840. https://doi.org/10.1111/jcal.12793

CHARMAZ, K. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. Sage.

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 preprint* arXiv:2306.14924. https://doi.org/10.48550/arXiv.2306.14924

CLARK, R. E., AND SAXBERG, B. 2018. Engineering motivation using the belief-expectancy-control framework. *Interdisciplinary Education and Psychology, 2*(1), 1–26. https://riverapublications.com/assets/files/pdf_files/engineering-motivation-using-the-belief-expectancy-control-framework.pdf

COHEN J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

CRONBACH, L. J., AND MEEHL, P. E. 1955. Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

CROSSLEY, S., MCNAMARA, D., BAKER, R. S., WANG, Y., PAQUETTE, L., BARNES, T., AND BERGNER, Y. 2015. Language to Completion: Success in an Educational Data Mining Massive Open Online Course. In *Proceedings of the 8th International Conference on Educational Data Mining*, O. C. Santos, J. B. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, and

M. Desmarais, Eds., 388–391. https://www.educationaldatamining.org/EDM2015/proceedings/short388-391.pdf

D'MELLO, S., AND GRAESSER, A. 2012. Dynamics of affective states during complex learning. *Learning and Instruction, 22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001

DICERBO, K. E. 2014. Game-based assessment of persistence. *Journal of Educational Technology & Society, 17*(1), 17–28. https://www.jstor.org/stable/jeductechsoci.17.1.17

DOCHY, F. J., AND ALEXANDER, P. A. 1995. Mapping prior knowledge: A framework for discussion among researchers. *European Journal of Psychology of Education, 10*(3), 225–242. https://doi.org/10.1007/BF03172918

DOWELL, N., AND KOVANOVIĆ, V. 2022. Modeling educational discourse with natural language processing. In *The handbook of learning analytics* (2nd ed.), C. Lang, G. Siemens, A. F. Wise, D. Gašević, and A. Merceron, Eds., Society for Learning Analytics Research (SoLAR), 105–119. https://www.solaresearch.org/publications/hla-22/hla22-chapter11/

DU, J., HEW, K. F., AND LIU, L. 2023. What can online traces tell us about students' self-regulated learning? A systematic review of online trace data analysis. *Computers & Education, 201,* 104828. https://doi.org/10.1016/j.compedu.2023.104828

DUNIVIN, Z. O. 2025. Scaling hermeneutics: A guide to qualitative coding with LLMs for reflexive content analysis. *EPJ Data Science, 14*(28). https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-025-00548-8

ECCLES, J. S., AND WIGFIELD, A. 2020. From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology, 61*, 101859. https://doi.org/10.1016/j.cedpsych.2020.101859

EFKLIDES, A. 2011. Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist, 46*(1), 6–25. https://doi.org/10.1080/00461520.2011.538645

FEINSTEIN, A. R., AND CICCHETTI, D. V. 1990. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–549. https://doi.org/10.1016/0895-4356(90)90158-L

GIGNAC, G. E. 2021. People who consider themselves smart do not consider themselves interpersonally challenged: Convergent validity evidence for subjectively measured IQ and EI. *Personality and Individual Differences, 174*, 110664. https://doi.org/10.1016/j.paid.2021.110664

GRAESSER, A., AND MCNAMARA, D. 2010. Self-regulated learning in learning environments with pedagogical agents that interact in natural language. *Educational Psychologist, 45*(4) 234–244. https://doi.org/10.1080/00461520.2010.515933

GWET, K. L. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61*(2), 297–308. https://doi.org/10.1348/000711006X126600

HARACKIEWICZ, J. M., BARRON, K. E., TAUER, J. M., CARTER, S. M., AND ELLIOT, A. J. 2000. Short-term and long-term consequences of achievement goals: predicting interest and performance over time. *Journal of Educational Psychology, 92*(2), 316–330. https://doi.org/10.1037/0022-0663.92.2.316

HESELTINE, M., AND VON HOHENBERG, B. C. 2024. Large language models as a substitute for human experts in annotating political text. *Research & Politics, 11*(1), 20531680241236239. https://doi.org/10.1177/20531680241236239

HULLEMAN, C. S., SCHRAGER, S. M., BODMANN, S. M., AND HARACKIEWICZ, J. M. 2010. A meta-analytic review of achievement goal measures: Different labels for the same constructs or different constructs with similar labels? *Psychological Bulletin, 136*(3), 422–449. https://psycnet.apa.org/buy/2010-07936-008

JOZSA, K., WANG, J., BARRETT, K. C., AND MORGAN, G. A. 2014. Age and Cultural Differences in Self-Perceptions of Mastery Motivation and Competence in American, Chinese, and Hungarian School Age Children. *Child Development Research 2014*, 1, 803061. https://doi.org/10.1155/2014/803061

KAI, S., ALMEDA, M.V., BAKER, R.S., HEFFERNAN, C., AND HEFFERNAN, N. 2018. Decision tree modeling of wheel-spinning and productive persistence in skill builders. *Journal of Educational Data Mining*, *10*(1), 36–71. https://doi.org/10.5281/zenodo.3344810

KANE, M. T. 2013. Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

KARABENICK, S. A., AND GONIDA, E. N. 2017. Academic help seeking as a self-regulated learning strategy: Current issues, future directions. In *Handbook of self-regulation of learning and performance*, D. H. Schun and J. A. Greene, Eds., Routledge, 421–433.

KASNECI, E., SESSLER, K., KÜCHEMANN, S., BANNERT, M., DEMENTIEVA, D., FISCHER, F., GASSER, U., GROH, G., GÜNNEMANN, S., HÜLLERMEIER, E., KRUSCHE, S., KUTYNIOK, G., MICHAELI, T., NERDEL, C., PFEFFER, J., POQUET, O., SAILER, M., SCHMIDT, A., SEIDEL, T., … KASNECI, G. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

KLASSEN, R. M., AND USHER, E. L. 2010. Self-efficacy in educational settings: Recent research and emerging directions. In *The decade ahead: Theoretical perspectives on motivation and achievement*, vol. 16, T. Urdan and S. A. Karabenick, Eds., Emerald, 1–33.

KRAPP, A. 2002. Structural and dynamic aspects of interest development: Theoretical considerations from an ontogenetic perspective. *Learning and Instruction, 12*(4), 383–409. https://doi.org/10.1016/S0959-4752(01)00011-1

KRIPPENDORFF, K. 2004. Measuring the reliability of qualitative text analysis data. *Quality and Quantity, 38*, 787–800. https://doi.org/10.1007/s11135-004-8107-7

KRIPPENDORFF, K. 2011. Agreement and information in the reliability of coding. *Communication Methods and Measures, 5*(2), 93–112. https://doi.org/10.1080/19312458.2011.568376

KUZMAN, T., AND LJUBEŠIĆ, N. 2025. LLM teacher-student framework for text classification with no manually annotated data: a case study in IPTC news topic classification. *IEEE Access*, *13*. https://ieeexplore.ieee.org/abstract/document/10900365

LANDIS, J. R., AND KOCH, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

LECHNER, C. M., DANNER, D., AND RAMMSTEDT, B. 2019. Grit (effortful persistence) can be measured with a short scale, shows little variation across socio-demographic subgroups, and is associated with career success and career engagement. *PLoS One, 14*(11), e0224814. https://doi.org/10.1371/journal.pone.0224814

LENT, R. W., BROWN, S. D., AND LARKIN, K. C. 1984. Relation of self-efficacy expectations to academic achievement and persistence. *Journal of Counseling Psychology, 31*(3,) 356–362. https://doi.org/10.1037/0022-0167.31.3.356

LI, J., ZHU, Y., LI, Y., LI, G., AND JIN, Z. 2024. Showing LLM-Generated Code Selectively Based on Confidence of LLMs. *arXiv preprint* arXiv:2410.03234. https://arxiv.org/abs/2410.03234

LIN, J., DIESENDRUCK, M., DU, L., AND ABRAHAM, R. 2023. BatchPrompt: Accomplish more with less. *arXiv preprint* arXiv:2309.00384. https://arxiv.org/abs/2309.00384

LIU, X., ZAMBRANO, A. F., BAKER, R. S., BARANY, A., OCUMPAUGH, J. ZHANG, J., PANKIEWICZ, M., NASIAR, N., AND WEI, Z. 2025. Qualitative coding with GPT-4: Where it works better. *Journal of Learning Analytics, 12*(1), 169–185. https://doi.org/10.18608/jla.2025.8575

LOEVINGER, J. 1957. Objective tests as instruments of psychological theory. *Psychological Reports, 3*(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

LOPEZ, A. A., GUZMAN-ORTH, D., ZAPATA-RIVERA, D., FORSYTH, C. M., AND LUCE, C. 2021. *Examining the accuracy of a conversation-based assessment in interpreting English learners' written responses.* (Research Report No. RR-21-03). Educational Testing Service. https://doi.org/10.1002/ets2.12315

MARTIN, A., RYAN, R. M., AND BROOKS-GUNN, J. 2013. Longitudinal associations among interest, persistence, supportive parenting, and achievement in early childhood. *Early Childhood Research Quarterly, 28*(4), 658–667. https://doi.org/10.1016/j.ecresq.2013.05.003

MCCAFFREY, D. F., CASABIANCA, J. M., RICKER-PEDLEY, K. L., LAWLESS, R. R., AND WENDLER, C. 2021. Best practices for constructed-response scoring. *ETS Research Report Series* 2021, 1, 1–58. https://www.ets.org/pdfs/about/cr_best_practices.pdf

MCCLURE, C., SMYSLOVA, O., HALL, A., AND JIANG, Y. 2024. Deductive coding's role in AI vs. human performance. In *Proceedings of the 17th International Conference on Educational Data Mining (EDM 2024),* C., Demmans Epp, B. Paaßen, and D., Joyner, Eds. https://educationaldatamining.org/edm2024/proceedings/2024.EDM-posters.91/

MEINDL, P., IYER, R., AND GRAHAM, J. 2019. Distributive justice beliefs are guided by whether people think the ultimate goal of society is well-being or power. *Basic and Applied Social Psychology, 41*(6), 359–385. https://doi.org/10.1080/01973533.2019.1663524

MELLON, J., BAILEY, J., SCOTT, R., BRECKWOLDT, J., MIORI, M., AND SCHMEDEMAN, P. 2024. Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Research & Politics, 11*(1), 20531680241231468. https://doi.org/10.1177/20531680241231468

MEREDITH, W. 1993. Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. https://doi.org/10.1007/BF02294825

MESSICK, S. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

MILES, M. B., HUBERMAN, A. M., AND SALDAÑA, J. 2014. *Qualitative data analysis: A methods sourcebook* (3rd ed.). SAGE Publications.

MINBASHIAN, A., WOOD, R. E., AND BECKMANN, N. 2010. Task-contingent conscientiousness as a unit of personality at work. *Journal of Applied Psychology, 95*(5), 793–806. https://doi.org/10.1037/a0020016

MULTON, K. D., BROWN, S. D., AND LENT, R. W. 1991. Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology, 38*(1), 30–38. https://psycnet.apa.org/buy/1991-16867-001

NEWTON, P. E., AND SHAW, S. D. 2014. Validity in educational and psychological assessment. http://digital.casalini.it/9781473904064

NUUTILA, K., TAPOLA, A., TUOMINEN, H., MOLNÁR, G., AND NIEMIVIRTA, M. 2021. Mutual relationships between the levels of and changes in interest, self-efficacy, and perceived difficulty during task engagement. *Learning and Individual Differences, 92*, 102090. https://doi.org/10.1016/j.lindif.2021.102090

OBER, T. M., COUREY, K. A., AND FLOR, M. 2026. Integrating topic modeling and LLM prompt engineering into a human-driven approach to analyze interview transcripts. *Journal of Educational Data Mining, 18*(1).

O'REILLY, T., WANG, Z., AND SABATINI, J. 2019. How much knowledge is too little? When a lack of knowledge becomes a barrier to comprehension. *Psychological Science, 30*(9), 1344–1351. https://doi.org/10.1177/0956797619862276

OUYANG, S., ZHANG, J. M., HARMAN, M., AND WANG, M. 2024. An empirical study of the non-determinism of ChatGPT in code generation. *ACM Transactions on Software Engineering and Methodology, 34*(2), 1–28. https://doi.org/10.1145/3697010

PAJARES, F. 1996. Self-efficacy beliefs in academic settings. *Review of Educational Research, 66*(4), 543–578. https://doi.org/10.3102/00346543066004543

PEEPERKORN, M., KOUWENHOVEN, T., BROWN, D., AND JORDANOUS, A. (2024). Is temperature the creativity parameter of large language models?. *arXiv preprint* arXiv:2405.00492. https://arxiv.org/abs/2405.00492

PINTRICH, P. R., AND DE GROOT, E. V. 1990. Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33–40. https://psycnet.apa.org/buy/1990-21075-001

PORTER, T., MOLINA, D. C., BLACKWELL, L., ROBERTS, S., QUIRK, A., DUCKWORTH, A. L., AND TRZESNIEWSKI, K. 2020. Measuring mastery behaviours at scale: The Persistence, Effort, Resilience, and Challenge-Seeking (PERC) Task. *Journal of Learning Analytics, 7*(1), 5–18. https://doi.org/10.18608/jla.2020.71.2

QIAO, S., FANG, X., GARRETT, C., ZHANG, R., LI, X., AND KANG, Y. 2024. Generative AI for qualitative analysis in a maternal health study: Coding in-depth interviews using large language models (LLMs). *medRxiv*, 2024-09. https://doi.org/10.1101/2024.09.16.24313707

RASHEED, Z., WASEEM, M., AHMAD, A., KEMELL, K. K., XIAOFENG, W., DUC, A. N., AND ABRAHAMSSON, P. 2024. Can large language models serve as data analysts? A multi-agent assisted approach for qualitative data analysis. *arXiv preprint* arXiv:2402.01386. https://doi.org/10.48550/arXiv.2402.01386

RAZAVI, A., SOLTANGHEIS, M., ARABZADEH, N., SALAMAT, S., ZIHAYAT, M., AND BAGHERI, E. 2025. Benchmarking prompt sensitivity in large language models. In *European Conference on Information Retrieval*, 303–313. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-88714-7_29

RENINGER, K. A., AND HIDI, S. 2011. Revisiting the conceptualization, measurement, and generation of interest. *Educational Psychologist, 46*(3), 168–184. https://doi.org/10.1080/00461520.2011.587723

SANSONE, C., AND THOMAN, D. B. 2005. Interest as the missing motivator in self-regulation. *European Psychologist, 10*(3), 175–186. https://doi.org/10.1027/1016-9040.10.3.175

SCHUNK, D. H. 1985. Self-efficacy and classroom learning. *Psychology in the Schools* 22, 2, 208–223. https://doi.org/10.1002/1520-6807(198504)22:2%3C208::AID-PITS2310220215%3E3.0.CO;2-7

SCHUNK, D. H., AND DIBENEDETTO, M. K. 2016. Self-efficacy theory in education. In *Handbook of motivation at school*, K. R. Wentzel, Ed., Routledge, 34–54.

SHAH, S. T. U., HUSSEIN, M., BARCOMB, A., AND MOSHIRPOUR, M. 2025. From inductive to deductive: LLMs-based qualitative data analysis in requirements engineering. *arXiv preprint* arXiv:2504.19384. https://doi.org/10.48550/arXiv.2504.19384

SHAPIRO, A. M. 2004. How including prior knowledge as a subject variable may change outcomes of learning research. *American Educational Research Journal, 41*(1), 159–189. https://doi.org/10.3102/00028312041001159

SHEPARD, L. A. 2016. Evaluating test validity: Reprise and progress. *Assessment in Education: Principles, Policy & Practice, 23*(2), 268–280. https://doi.org/10.1080/0969594X.2016.1141168

SILVIA, P. J. 2005. What is interesting? Exploring the appraisal structure of interest. *Emotion, 5*(1), 89–102. https://psycnet.apa.org/buy/2005-02259-008

SIMONSMEIER, B. A., FLAIG, M., DEIGLMAYR, A., SCHALK, L., AND SCHNEIDER, M. 2022. Domain-specific prior knowledge and learning: A meta-analysis. *Educational Psychologist, 57*(1), 31–54. https://doi.org/10.1080/00461520.2021.1939700

SKINNER, E. A., AND PITZER, J. R. 2012. Developmental dynamics of student engagement, coping, and everyday resilience. In *Handbook of research on student engagement*, S. Christenson, A. Reschly, and C. Wylie, Eds., Springer US, 21–44. https://doi.org/10.1007/978-1-4614-2018-7_2

SKINNER, E. A., GRAHAM, J. P., BRULE, H., RICKERT, N., AND KINDERMANN, T. A. 2020. "I get knocked down but I get up again": Integrative frameworks for studying the development of motivational resilience in school. *International Journal of Behavioral Development, 44*(4), 290–300. https://doi.org/10.1177/0165025420924122

SPARKS, J. R., LEHMAN, B., GLADSTONE, J., ZHANG, S., SCHROEDER, N., AND ISRAEL, M. 2025. Measuring persistence and academic resilience of K-12 students: Systematic review and operational definitions. *Frontiers in Education, 10*, 1673500. https://doi.org/10.3389/feduc.2025.1673500

STEWART, S., LIM, D. H., AND KIM, J. 2015. Factors influencing college persistence for first-time students. *Journal of Developmental Education*, *38*(30, 12–20. https://www.jstor.org/stable/24614019

TINTO, V. 2017. Reflections on student persistence. *Student Success, 8*(2), 1–8. https://search.informit.org/doi/abs/10.3316/INFORMIT.593199291602507

TOBIAS, S. 1994. Interest, prior knowledge, and learning. *Review of Educational Research, 64*(1), 37–54. https://doi.org/10.3102/00346543064001037

TÖRNBERG, P. 2025. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review, 43*(6), 1181–1195. https://doi.org/10.1177/08944393241286471

TULIS, M., AND FULMER, S. M. 2013. Students' motivational and emotional experiences and their relationship to persistence during academic challenge in mathematics and reading. *Learning and Individual Differences, 27*, 35–46. https://doi.org/10.1016/j.lindif.2013.06.003

TURPIN, M., MICHAEL, J., PEREZ, E., AND BOWMAN, S. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems, 36*, 74952–74965. https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html

WIGFIELD, A., AND ECCLES, J. S. 2000. Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

WIGFIELD, A., MUENKS, K., AND ECCLES, J. S. 2021. Achievement motivation: What we know and where we are going. *Annual Review of Developmental Psychology, 3*(1), 87–111. https://doi.org/10.1146/annurev-devpsych-050720-103500

YOSHIDA, L. 2025. Do we need a detailed rubric for automated essay scoring using large language models?. In *Artificial Intelligence in Education. AIED 2025. Lecture Notes in Computer Science, vol 15882*, A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, and S. Isotani, Eds., Cham: Springer Nature Switzerland, 60–67.

ZAPATA-RIVERA, D., AND FORSYTH, C. M. 2022, June. Learner modeling in conversation-based assessment. In *International Conference on Human-Computer Interaction*. Cham: Springer International Publishing, 73–83. https://doi.org/10.1007/978-3-031-05887-5_6

ZAPATA-RIVERA, D., JACKSON, T., AND KATZ, I. R. 2015. Authoring conversation-based assessment scenarios. In *Design Recommendations for Intelligent Tutoring Systems Volume 3: Authoring Tools and Expert* Modeling Techniques, R. A. Sottilare, A. C. Graesser, X. Hu, and K. Brawner Eds., U.S. Army Research Laboratory, 169–178.

ZAPATA-RIVERA, D., SPARKS, J. R., FORSYTH, C. M., AND LEHMAN, B. 2023. Conversation-based assessment: current findings and future work. In *International Encyclopedia of Education (Fourth Edition)* R. J. Tierney, F. Rizvi, and K. Ercikan, Eds., Elsevier, 504–518). https://doi.org/10.1016/B978-0-12-818630-5.10063-6

ZHANG, S., MESHRAM, P. S., GANAPATHY PRASAD, P., ISRAEL, M., AND BHAT, S. 2025. An LLM-based framework for simulating, classifying, and correcting students' programming knowledge with the SOLO taxonomy. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 2*, J. A. Stone, T. Yuen, L. Shoop, S. A. Rebelsky, and J. Prather, Eds., 1681–1682. https://doi.org/10.1145/3641555.3705125

ZHOU, M., AND KAM, C. C. S. 2017. Trait procrastination, self-efficacy and achievement goals: the mediation role of boredom coping strategies. *Educational Psychology, 37*(7), 854–872. https://doi.org/10.1080/01443410.2017.1293801

ZIEMS, C., CHEN, J., ZHANG, A., AND YANG, D. 2023. Can large language models transform computational social science? *arXiv preprint*. https://doi.org/10.48550/arXiv.2305.03514

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

ZIMMERMAN, B. J., AND MOYLAN, A. R. 2009. Self-regulation: Where metacognition and motivation intersect. In *Handbook of metacognition in education*, D. J. Hacker, J. Dunlosky, and A. C. Graesser, Eds., Routledge, 299–315.

ZUMBO, B. D. 2009. Validity as contextualized and pragmatic explanation, and its implications for validation practice. In *The concept of validity: Revisions, new directions, and applications*, R. W. Lissitz, Ed., Information Age Publishing, 65–82. https://psycnet.apa.org/record/2009-23060-004