Multimodal LLM-based approach for detecting Disengagement from Task Goal (DTG) in Math problem solving

Nidhi Nasiar[1], Ryan S. Baker[2], Jaclyn Ocumpaugh[3], Zeyneb S. Sarioglan[1], Caitlin Mills[4], and Samuel Rhodes[5]
[1]University of Pennsylvania
[2]Adelaide University
[3]University of Houston
[4]University of Minnesota
[5]Virginia Commonwealth University
{nasiar, zeysario}@upenn.edu, {ryanshaunbaker, jlocumpaugh}@gmail.com, cmills@umn.edu, rhodessr2@vcu.edu

ABSTRACT
Learner disengagement has been widely studied and modeled in math learning environments to enable timely supports and promote engagement. Recent advances in multimodal large language models (MLLMs) offer new opportunities for building such detectors. This study investigates the potential of an MLLM-based approach to detect disengagement from the task goal (DTG) in screen-recording videos of students solving math problems in CueThink. DTG is operationalized as actions unrelated to the intended learning task. We develop and validate a DTG detector using LLaVA-Video-7B-Qwen2(Large Language and Vision Assistant) using various prompting strategies, and analyze themes across the few model errors. We also examine associations between DTG behavior and learning gains, as well as students' math beliefs, anxiety, and metacognition. Results indicate that the model, using role assignment and chain-of-thought prompting, identifies DTG behavior with good performance ($\kappa = 0.71$, precision = 77.6%, recall = 76.4%). Performance limitations that emerged related to low-fidelity data, limited contextual knowledge, and following instructions. DTG behavior was marginally negatively correlated with normalized learning gains later in the academic year (winter to spring), but not during the earlier fall-to-winter period. There were no significant correlations to beliefs or anxiety.

Keywords
Multimodal LLM, LLaVA, disengagement, math, problem-solving, detector, video-analysis.

1. INTRODUCTION
The field of Educational Data Mining (EDM) has long-standing interest in various types of disengagement within and across learning platforms [34, 21]. Student disengagement is a multifaceted meta-construct [43] that includes behavioral (e.g. gaming the system), emotional (e.g. boredom), and cognitive components (e.g. mind wandering) [19]. Identifying and understanding student disengagement is important to efforts to reduce it [1], as engagement is a necessary condition for learning and a key indicator of long-term achievement and academic success [36, 39]. Students can only benefit from the virtual learning environments if they are engaged with the system [10]. Engagement also plays a role in students' academic resilience and the development of resources for coping adaptively with stressors, which in turn may affect the development of long-term academic mindsets [37].

Within online learning contexts, disengagement has been modeled using various Machine Learning (ML) models. Specific examples of disengaged behavior include DTG (Disengaged from Task-Goal) [22], gaming the system [5], off-task behavior [4], and mind wandering [32]. There has been increased interest in identifying the various ways students can deviate from expected patterns while using educational software, including not using features of the platform and failure to engage with content [21]. Detection of these behaviors has helped researchers better understand them, allowing researchers and teachers to develop interventions that can remediate them and reduce their negative impacts on user outcomes [2].

For this paper, we focus specifically on Disengagement from Task Goal (DTG). This behavior has been observed in online learning but given a variety of names in the published literature [22, 35, 8, 42]. Examples include students who, instead of plotting points from a mathematical function in a learning platform for high school mathematics, plotted a smiley face [42]. In other learning systems, learners have avoided diagnosing virtual patients and instead placed virtual unrelated objects on them [35], tinkered with science simulations in ways unrelated to the stated learning goals [8], or otherwise engaged in actions that appear to have no relationship to the intended learning task [42]. To date, the majority of studies of DTG have been in science learning environments [22, 35, 42], with very limited work in math contexts.

In the context of math problem solving, DTG may take several forms, including using math tools and text to draw silly drawings, making smiley faces when plotting points, creating repetitive designs and patterns, typing curse words instead of answers, or writing funny terms for friends. It is worth noting that DTG differs from off-task behavior as that is typically understood: off-task behavior typically involves disengaging completely from the learning task, whereas in DTG, the student is engaging with the task, but in a fashion unrelated to the learning task's design goals or incentive structure.

Disengagement detectors have often been built using classical machine learning (ML). However, ML-based models are time and resource-intensive, requiring extensive feature engineering and labeled datasets. With the advancements in generative AI and large

language and multimodal models, new opportunities have emerged for building detectors. Large language models like ChatGPT have shown potential in facilitating education research through supporting qualitative data analysis tasks such as coding transcripts and analyzing open-ended responses [45, 48, 30]. Recent advances in multimodal large language models (MLLMs) have extended these capabilities to process both visual and textual information [47]. MLLMs have demonstrated promising results in tasks such as image captioning, video understanding, and visual question answering [44, 48].

One potential application of MLLMs in educational research is facilitating qualitative video analysis, which has yielded important results about students' learning processes, but previously required time-intensive coding of behavioral patterns and interactions [16, 33]. Emerging educational research has explored Machine Learning classification to detect teaching and learning moves [18], but these efforts also relied on manually coded data for training. Early work with MLLMs in online learning has been promising [51], demonstrating their potential for streamlining this process by supporting video selection and behavior coding—tasks that traditionally require significant manual effort from researchers. To date, however, their use for modeling DTG has not yet been studied.

Accordingly, the focus of this paper is to build and validate a DTG detector based on the multimodal LLM, LLaVA (Large Language and Vision Assistant) [27], and to systematically analyze its errors. Additionally, we examine how DTG behavior correlates with learning gains and measures of math beliefs, anxiety, and metacognition in math problem-solving—constructs that research has shown are increasingly important for math learning [3, 13, 15].

2. METHODS

2.1 Learning Platform and Data Collected

CueThink (2022) is a digital learning application designed to enhance middle-school students' mathematics problem-solving skills by encouraging self-regulated learning (SRL) and supporting students in sharing their problem-solving processes. The platform asks students both to solve a mathematics problem and to create a shareable screencast video that presents their solution and demonstrates their reasoning. CueThink structures each problem into a Thinklet, which consists of four phases—Explore, Plan, Solve, and Review—that closely align with Winne and Hadwin's [41] SRL model. A full description of the CueThink phases is provided in Zhang et al. [49].

The study included 180 students in Grades 6–8 from suburban schools in the Southwestern United States who used CueThink during the 2021–2022 school year. Self-reported gender demographics were collected (66 boys, 98 girls, 1 non-binary, 3 chose not to report, and 12 had missing values). Across the year, students completed 746 Thinklets (4.09 Thinklets per student, on average). Pre- and post-test surveys were administered at the beginning and end of the school year; the survey scales are described in detail below (section 2.4).

This study primarily makes use of data from the Solve phase. During this phase, CueThink collects fine-grained interaction log and trace data capturing students' whiteboard actions, including position coordinates for each pen stroke (e.g., initiating a path and rendering it point-by-point). Students can use a variety of mathematical tools and objects on the whiteboard—such as number lines, protractors, tables, highlighters, and text/equation entry (Figure 1). These interaction traces can be rendered into a screen-recording that visualizes students' step-by-step work as they solve the problem. These videos were used both as inputs to the model and to establish ground truth labels (see section 2.2).

All data were collected and analyzed under a protocol approved by the University of Pennsylvania Institutional Review Board (IRB protocol number 856358). The LLaVA model was run locally on our lab server for analysis, so no data was transmitted to any external server, minimizing data privacy risks.

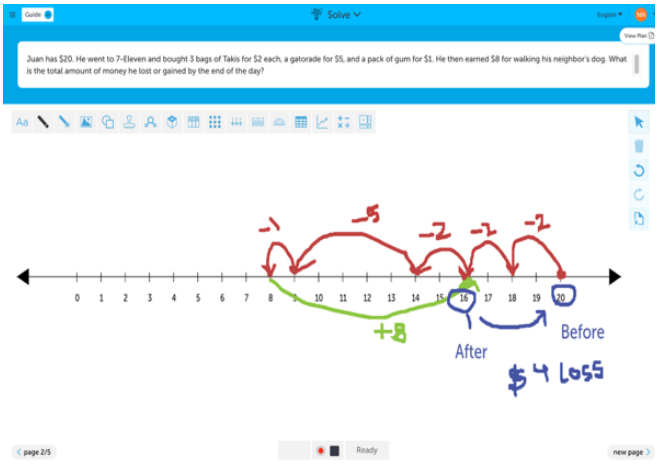

Figure 1. Solve Phase in CueThink

2.2 Operationalizing DTG

DTG (Disengaged from Task Goal) was operationalized as actions that have no relation to the intended learning task. The intended learning task here is to use system-internal tools to write and solve an equation on the digital whiteboard, and tools included a number line, table, protractor, ruler, or mathematical symbols. Each video was binary coded for DTG behavior by 2 researchers to establish the ground truth. Each researcher watched the full video, which showed all the students' actions in chronological order as they solved the math problem in the context of the learning environment, to then make a decision about whether DTG behavior was ever observed or not.

The operationalization was developed through an iterative process to ensure that the behaviors classified as DTG aligned with previous conceptualizations [22] and were salient in the dataset. Through this process, indicators of engagement (non-DGT behavior) were readily agreed upon, including using a number line to add or subtract or organizing given values in the problem statement into a table to solve the problem, or any written text or images drawn for finding a solution. Meanwhile, several specific criteria for DGT were identified, including: (1) text unrelated to given math problems, (2) usage of given math tools for purposes other than solving the math problem, and (3) other task-unrelated actions (silly drawings, smileys).

Some clarifications were made to the DGT criteria. For example, behaviors that might have indicated DTG were ignored if they occurred after the problem was completed. As such, if the video included a student who finished solving the problem on the whiteboard with a solution before creating an unrelated drawing, the video was not labeled as DTG. If the video is blank and no actions take place on the whiteboard, it is also not treated as DTG because there was no action.

Using this operationalization, two coders first independently coded 120 videos from a separate dataset from another year on the same learning system to establish interrater reliability ($\kappa = 0.79$). Once consensus was reached, the coders proceeded to code a total of 746 videos from 180 students. These coded videos constituted our "ground truth" that was used to compare against model labels to validate the model's performance. Of these 746 videos, 546 videos from 129 students were designated as the development set and used for iterative prompt refinement and evaluating strategies, while the remaining 200 videos from 51 students (none of whom appeared in the development set) served as a held-out test set for evaluating the final model's performance and further analysis.

2.3 LLaVA as a DTG detector

2.3.1 LLaVA Model Details

We used LLaVA-Video-7B-Qwen2 [50], a state-of-the-art video MLLM was selected because it reflects recent technological advances in video MLLMs, is open source, and has demonstrated suitable performance on video understanding benchmarks [20,50] and in educational applications [51]. We used the pre-trained model with temperature set to 0 to ensure reproducibility. We did not fine-tune the model because our aim was to assess the potential of existing pre-trained models for detecting disengagement behaviors, and fine-tuning requires substantial computational resources and large annotated datasets.

2.3.2 Video Analysis and Script Setup

To process the video data, we selected sampling and segmentation settings that balance temporal coverage and processing capacity. Each video was first decomposed into individual frames. These frames, together with the text prompts—were tokenized as inputs to the model. Because the MLLM has a fixed context window that limits the total number of tokens it can process at once, analyzing longer videos involves a trade-off between the number of tokens allocated per frame (increasing visual detail) and the total number of frames included (increasing temporal coverage).

Each video was segmented into 10-second clips and sampled at three frames per clip. This conservative choice was derived empirically by testing multiple segmentation lengths (i.e., 10, 5, and 1-second clips) and sampling values (3 vs. 5 frames per clip), in order to capture sufficient student actions and visual context to support DTG coding while remaining within token limits. Different segment lengths and sampling rates showed no meaningful differences in output quality, and therefore, we adopted a 10-second segmentation with three frames per clip (approximately one frame every 3.3 seconds, or ~0.3 fps) to maintain computational efficiency. All clips were then processed through LLaVA using the refined prompt. To obtain a single video-level prediction per student, clip-level predictions were aggregated using an inclusive decision rule, such that, if DTG behavior was identified in one or more clips within a video, the student was labeled as exhibiting DTG behavior for that problem.

2.3.3 Prompting strategies

We employed an iterative approach to refine the text prompts given to LLaVA using the development set (546 videos), focusing on maximizing the accuracy of labeling for DTG behavior. Once prompt development was complete, we evaluated the final prompt configuration on a held-out test set of 200 videos to assess the model's performance on unseen data. Given that LLaVA—like other generative models—is known to be sensitive to prompt framing and task specification [40, 29], we tested several prompting strategies, including (1) a baseline prompt, (2) a role-assignment strategy (which assigns a specific role or perspective to the model in order to facilitate a more contextualized interpretation e.g., Liu et al. [29]), (3) a chain-of-thought strategy (which facilitates step-by-step reasoning processes), (4) a combination of both approaches. These prompting strategies were selected based on prior work showing that prompting choice can affect multimodal LLM output. Chain-of-thought prompting has been shown to improve visual reasoning in multimodal LLMs, including LLaVA [48], while role-assignment has produced mixed results across settings [29]. We tested both techniques and their combination to identify the configuration best suited to DTG behavior labeling for our data. We also tested variations by changing key instructional verbs (e.g., "analyze" vs. "examine") while maintaining the role assignment component. Additionally, we provided contextual knowledge about math problem-solving in the prompt to help the model identify regular math-solving actions (task-related actions), such as strikethroughs to numbers and text that indicate problem-solving processes.

Baseline Prompt. The baseline prompt was as follows:

"Assign a binary code for DTG (disengagement from task goal) behavior defined as 'actions that have no relation to the math problem."

Role assignment and chain-of-thought strategy. The baseline combined-approach prompt was as follows:

"You are a [role assignment: qualitative coding expert] and your task is to [analyze/examine] and assign a binary coding for the construct of DTG (disengagement from task goal), which is defined as 'actions that have no relation to the math problem'. The videos show the whiteboard with student actions. The intended learning task for students was to solve math problems using the given tools (like a number line, table, protractor, or ruler, and other mathematical symbols). [chain-of-thought reasoning: insert step-by-step rules for operationalization criteria]."

Throughout this iterative process, we maintained concise prompts to optimize token allocation, as longer prompts can lead to attention dilution [23, 28] and degrade model performance even before reaching technical token limits [26].

2.3.4 Validation of LLaVA's DTG labels

During the prompt-refinement process, prompts given to LLaVA used a development set (546 videos), focusing on maximizing the performance of the model for correctly labeling DTG behavior. Once prompt development was complete, we evaluated the final prompt configuration on a held-out test set of 200 videos to assess the model's performance on unseen data across three metrics—kappa, precision, and recall.

2.3.5 Error Analysis

After finalizing the prompt, the test dataset was sent to the model for binary coding the videos for DTG behavior. For any clips where the MLLM label contradicted the human ground-truth label, further analyses were conducted. Namely, the 1st and 3rd authors re-watched the associated videos and analyzed these clips to identify potential causes of disagreement and possible instances of model hallucinations. We then conducted a thematic analysis [7] to distill recurring patterns across these errors, iteratively grouping these observations into broader categories.

2.4 External Survey Measures

Validated external measures were used to assess students' content knowledge [14], math anxiety [9], metacognitive awareness [38], and beliefs on problem-solving [25]. Each of these measures was

used first as a pre-test and then as a post-test. Once the pre-test was complete, students were provided with access to the learning platform. Their interaction with the platform was integrated into their regular classroom instruction, with their teacher assigning problems for them to complete.

Content Knowledge. Partnering districts administered i-Ready diagnostic assessments [14] as a proxy for mathematics content knowledge. The i-Ready [14] instrument is an adaptive assessment tool used to identify math topics students are struggling with; it examines students' understanding of mathematical sub-domains, including numbers and operations, algebra, geometry, and measurement. This assessment was administered three times, once each at the beginning (September), middle (December to January), and end (May) of the academic year.

Normalized learning gains [31] were calculated for those periods for the i-Ready tests from Fall to Winter; and Winter to Spring as NormLG_F-W and Norm_LG_W-S, respectively:

$$\text{NormLG} = \begin{cases} (\text{post} - \text{pre})/(\text{max} - \text{pre}) & \text{if post} \geq \text{pre} \\ (\text{post} - \text{pre})/\text{pre} & \text{if post} < \text{pre} \end{cases}$$

The normalized gains for two semesters were used for correlation analysis, and not the raw i-Ready scores.

The Junior Metacognitive Awareness Inventory (JrMAI) is a measure of metacognitive and cognitive strategies applied by learners [38] that was developed for students in grades 6 through 12. The current study uses a 9-item abbreviated version of the JrMAI that was validated with students in grades 6 -8 [24]. The measure includes 5 items identified as regulation of cognition and 4 items identified as knowledge of cognition.

The modified Abbreviated Math Anxiety Scale (mAMAS; [9]) uses a two-factor structure, which results in two subscales learning math anxiety (Learning subscale), and math evaluation anxiety (Evaluation subscale) [9]. The scale has shown good internal consistency, with an overall Cronbach's α of 0.85, a Cronbach's α of 0.77 for the Learning subscale and a Cronbach's α of 0.79 for the Evaluation subscale [9, 11]. The scale was developed for math learners between 8 and 13 years old (i.e., overlapping our research sample) and includes a 9-item self-report of math anxiety that are averaged to produce one final scale for analysis. The mAMAS was slightly modified to change adapted words to American English (e.g., "maths" to "math").

The Indiana Mathematics Belief Scales (IMBS; [25]) prompts students regarding their beliefs about mathematics and mathematical problem-solving over a 36-item survey. The measure is divided into six subscales. In this work, we used 2 of the 6 subscales (1 and 6). Specifically, we administered shortened versions of three subscales, which measure (a) student beliefs that they can solve time-consuming math problems (5 items), and (b) student beliefs on useful of math in their real lives, respectively (3 items). The shortened versions were validated using middle school students.

2.4.1 Analysis

In addition to the analysis of LLaVA's performance to accurately label DTG behaviors, we also examined the association between DTG labels and the external measures in this study. For analysis, first we aggregated DTG behavior at the student level by averaging DTG across each student's Thinklets. Then, for calculating correlations with survey measures, we used Spearman rank correlation because the variables were non-normally distributed. Afterwards, we applied Benjamini–Hochberg post-hoc corrections to control the false discovery rate for multiple comparisons [6].

3. RESULTS

3.1 Performance across prompts

Several metrics were used for evaluation including Cohen's kappa (κ), which was calculated for how well specific prompts produced outputs in agreement with human labels. Precision indicated the proportion of positive predictions that were correct. Recall measured the proportion of actual positive instances that the model successfully identified.

Table 1 reports the 3 metrics across the baseline prompt and prompt variations that introduced role assignment, chain-of-thought–inspired prompting, and their combination, evaluated under two phrasings ("analyze and assign" vs. "examine and assign").

Table 1. Kappa (κ), precision, and recall across prompts

Prompt Strategy	Analyze and Assign			Examine and Assign		
	κ	Prec.	Recall	κ	Prec.	Recall
Baseline	0.40	50.4%	59.0%	0.34	48.7%	59.2%
Role-assignment only	0.68	75.0%	78.3%	0.57	62.7%	75.5%
CoT only	0.60	66.7%	74.5%	0.53	59.2%	71.0%
Role-assignment + CoT	0.71	77.6%	76.4%	0.65	71.4%	75.3%

Across different prompting strategies, the 'role-assignment and chain-of-thought (CoT)' prompt with 'analyze and assign' achieved the highest kappa (κ = 0.71) and precision (77.6%), though role-assignment only had marginally higher recall (78.3% vs. 76.4%). This was followed by 'role-assignment only' with 'analyze and assign' (κ = 0.68, precision = 75%, and recall = 78.3%). Overall, 'analyze and assign' had higher values across all metrics and prompting strategies in comparison to 'examine and assign'. The baseline prompt, without any context or role, had the lowest scores for both 'analyze and assign' (κ = 0.40, precision = 50.4%, recall = 59%) and 'examine and assign' (κ = 0.34, precision = 48.7%, recall = 59.2%). We selected 'role-assignment and CoT' with 'analyze and assign' as the final prompting strategy due to higher metrics overall, and also balanced precision and recall, This prompt was used by the model to code the test set that was used for further analyses.

3.2 DTG Rates

The descriptive statistics of Disengaged-from-Task-Goal (DTG) behavior coded by the LLaVA model show that, across all Thinklets (N = 746), 79 were coded as DTG (10.59%), consistent with prior reports of DTG rates (e.g., Gobert et al., 2015; Sabourin et al., 2013). At the student level (N = 180), 67 students (37.22%) exhibited DTG behavior at least once across Thinklets. For each student, the DTG average was computed as the proportion of that student's Thinklets coded as DTG (DTG-coded Thinklets ÷ total Thinklets), with a score ranging from 0 (no DTG behavior) to 1 (DTG behavior on all Thinklets completed by that student). By self-identified gender, 32 of 66 boys (48.5%) and 33 of 98 girls (33.7%) exhibited DTG behavior. One non-binary student showed no DTG; among the three students who chose not to report gender, one showed DTG. A small number of students had missing gender values and were excluded from gender-specific percentages.

3.3 Error Analysis

The model showed good performance, comparable to or better than previous ML-based approaches for DTG behavior [22, 35].

However, to understand the few cases where it failed and enable future improvements, we conducted an analysis of cases incorrectly labeled as DTG behavior by the LLaVA model. We identified distinctive themes that can be categorized as follows:

3.3.1 Visual processing difficulties (low-fidelity data)

The LLaVA model exhibited systematic mislabeling stemming from low-fidelity visual interpretation challenges. Small-scale features posed particular difficulties: miniature smiley faces were incorrectly flagged as DTG when they were constructed from text characters like semicolons and parentheses (e.g., ";)" or ":)"), and small font sizes substantially increased OCR (optical character recognition) error rates. When handwritten objects appeared visually similar to numeric values, the model struggled to distinguish between task-relevant content and off-task elements—a differentiation that proved especially challenging given the handwritten nature of student work.

The model demonstrated poor recognition of cursive script and unclear handwriting. Pen scratches used by students to correct or cross out content were systematically mislabeled as DTG behaviors, as were unclear handwriting, numerical scribbles, and ambiguous pen-drawn elements that human annotators identified as legitimate task-related work. Additionally, when small objects overlaid background elements, human raters easily recognized the highlighted foreground from its color contrast, but the model often did not, indicating limitations in using color and spatial reasoning for labeling by the model. LLaVA connects a vision encoder (e.g., CLIP ViT) to a language model via visual instruction tuning [27], and not a conventional OCR pipeline, which might explain its sensitivity to tiny/cursive text and overlaps.

3.3.2 Lack of contextual knowledge

The model demonstrated insufficient understanding of mathematical tools, their appropriate usage contexts, and math problem-solving conventions. It mislabeled legitimate mathematical notation and work-in-progress indicators as DTG. For instance, tally marks, which human annotators recognized as counting tools based on their implicit contextual knowledge, were incorrectly flagged as disengagement. Similarly, incomplete solutions where students had drawn circles or shown intermediate steps without final answers were coded as DTG, whereas humans interpreted these as unfinished work rather than disengagement.

The model also struggled with distinguishing purposeful tool use from disjointed, unrelated behavior. Cases involving overlays of multiple mathematical objects (e.g., protractors, circles of different radii, rows of circles, colorful blocks, empty tables) were sometimes coded as 'not DTG' by the model when they clearly represented disengagement rather than actual problem-solving. Similarly, the rapid use of multiple mathematical tools to create unrelated imagery (e.g., multiple consecutive circles and lines forming an eye shape) was sometimes also difficult for LLaVA to recognize as DTG. Furthermore, the model did not interpret the written text correctly in the math problem-solving context. When students wrote text clearly unrelated to the problem, human annotators could read and infer the DTG behavior of the content, but LLaVA lacked this interpretive capability.

3.3.3 Instruction-following limitations

Consistent with previous findings [51], LLaVA model sometimes was unable to follow all of the instructions it was given, particularly in following conditional rules specified in the prompt. For example, it was unsuccessful at following instructions stating that activities that occurred after the student had completed the problem should not be classified as DTG.. Such errors are instructive about the limitations of LLaVA, which nonetheless achieved an overall model performance high enough to justify its use in aggregate correlational analyses.

3.4 Correlating DTG and Survey Measures

After applying the Benjamini–Hochberg post-hoc correction [6], none of the results of the Spearman correlations between students' disengagement from task goal (DTG) behavior and various outcome measures. However, marginally significant results (defined as $p < 2 \times$ adjusted α) were observed for one measure: normalized learning gains from winter to spring ($\rho = -0.278$, $p = 0.02$, adjusted $\alpha = 0.013$). In other words, students who more frequently engaged in DTG behavior demonstrated lower learning gains during this period. However, the relationship did not hold for normalized learning gains as calculated from fall to winter ($\rho = -0.155$, $p = 0.118$, adjusted $\alpha = 0.02$), suggesting that the negative effects of DTG may become more pronounced in the later part of the academic year.

The positive correlation between DTG behavior and post-test scores on the Indiana Mathematics Belief Scale (IMBS) subscale measuring beliefs about solving time-consuming math problem was not significant after post-hoc correction ($\rho = 0.157$, $p = 0.009$, bh-adjusted $\alpha = 0.004$), as the p-value was not below the critical (adjusted α) value. This correlation was also absent at pre-test ($\rho = -0.031$, $p = 0.687$, bh-adjusted $\alpha = 0.038$). No significant correlations were observed for the belief subscale on the usefulness of mathematics at either of the tests.

The negative correlation between DTG behavior and pre-test scores on the Math Anxiety Scale learning subscale was not significant after the Benjamini–Hochberg post-hoc correction ($\rho = -0.146$, $p = 0.018$, bh-adjusted $\alpha = 0.008$), as it did not meet the critical value criterion. This relationship was not significant at post-test ($\rho = -0.026$, $p = 0.736$, bh-adjusted $\alpha = 0.042$). The evaluation subscale of mathematics anxiety showed no significant correlations at either time point, which may reflect the fact that students were engaged in regular mathematics instruction rather than high-stakes testing contexts.

No other significant correlations were observed between DTG behavior and the metacognitive awareness scale.

4. DISCUSSION & CONCLUSION

4.1 Main Findings

This study demonstrates the potential of using LLaVA (Large Language and Vision Assistant), a multimodal large language model (MLLM) for automated detection of disengagement from task goal (DTG) behavior in screen recordings of virtual educational software. The pre-trained LLaVA model achieved acceptable performance ($\kappa = 0.71$, precision of 77.6%, and recall of 76.4%) in detecting DTG behavior during math problem-solving tasks. This shows that the model had high agreement with human raters while effectively minimizing both false positives and false negatives, and maintained balanced performance across the two metrics, suggesting robust and reliable detection capability across different types of DTG instances.

Rates of DTG labels by LLaVA model were consistent with prior reports of rates of this behavior across learning platform [22, 35]. Different prompting strategies were evaluated to improve the model performance, since the models are known to be sensitive to prompts. The results showed that framing the task to "analyze and

assign" and including assigning a role and chain-of-thought reasoning produced the highest kappa scores for coding DTG.

Analyzing the few error cases, the visual processing issues align with known weaknesses in vision-language models that rely on visual instruction tuning rather than dedicated OCR pipelines, particularly when processing low-resolution or visually ambiguous inputs [27]. Such challenges can be harder in educational settings where student-generated content is inherently messy, handwritten, and varies in quality. A lack of in-depth contextual knowledge of math, specifically understanding the tools and problem-solving processes required for particular tasks, represented another category with instances of incorrect DTG labels. A similar effect has been documented previously by Huang et al. [23], where they show that when multimodal LLMs lack sufficient contextual grounding, they can move toward speculative interpretation rather than evidence-based analysis. Lastly, the instruction-following limitations, including failures to apply conditional logic, are consistent with past findings [51], suggesting that sometimes gaps in sequential reasoning exist for the model.

While no correlations were statistically significant after the B&H correction, the marginally significant negative correlation between DTG behavior and normalized learning gains from winter to spring reflects an interesting relationship between DTG behavior and learning outcomes that warrants further investigation with larger samples. Notably, this association was absent in the fall-to-winter period, which may indicate that the cumulative effects of DTG behavior can become more harmful as the academic year progresses and mathematical content increases in complexity, consistent with prior research linking off-task behavior to reduced learning over time [12].

4.2 Future Work

This work has demonstrated the ability of MLLMs to considerably speed the labeling of DTG behaviors. Future research should investigate the individual differences and situational factors that lead students to engage in DTG behavior. For example, it is possible that students who engage in DTG are exhibiting a work avoidance goal orientation [17]. If researchers could distinguish between students who engage in DTG because they lack the necessary knowledge or skills and students who engage in DTG because the task is too easy, for example, we might be able to develop stronger scaffolding and adaptivity.

One approach to exploring this would be to examine DTG behavior using a broader range of survey metrics, including those that might be used for studies of students across learning domains (i.e., to determine if the same kinds of behaviors that emerge in a students' math classes also show up in science). However, additional qualitative data might also be informative. For example, now that there are functional DTG detectors for CueThink, it would be possible to use them to trigger classroom interviews, which allow researchers to approach, observe, and interview students at the moment a construct of interest occurs. In the context of DGTs, this could lead to a deeper understanding of this behavior and the needs of the students who exhibit it.

In addition to deeper investigations of DTG behaviors, future work should also explore ways to improve model performance and to validate it across different learning domains and student populations. These approaches should likely include:

Improving data size and quality: Researchers should provide high-fidelity data when using MLLMs for automated coding. Since large language models cannot reliably predict uncertainty or provide confidence intervals for their labeling [46], they cannot effectively mark boundary cases where human judgment would naturally acknowledge ambiguity.

Domain knowledge integration: The model would benefit from a lightweight domain ontology for common math tools and their typical usage patterns (e.g., protractor → angle labels; tally marks → counting). Additionally, implementing consistency checks to ensure tool presence co-occurs with task-relevant structure (labels, measurements, aligned marks), enhancing text understanding for semantic relevance to flag unrelated content as DTG, and training with negative examples showing decorative or purposeless tool use would improve contextual understanding.

Experimenting with fine-tuning: OCR capabilities can be enhanced through fine-tuning via visual instruction tuning specifically for handwritten mathematical content [27]. This targeted fine-tuning could address the model's current sensitivity to cursive text, small-scale features, and overlapping elements. However, this would require a much larger dataset, and is computationally expensive.

4.3 Conclusion

This study developed and validated a multimodal large language model (MLLM)–based detector for Disengagement from Task Goal (DTG) behavior using LLaVA (Large Language and Vision Assistant) on screen recordings of middle school students solving math problems on the CueThink platform. Among the prompting strategies evaluated, the combination of role assignment and chain-of-thought reasoning with an "analyze and assign" framing achieved the highest performance ($\kappa = 0.71$, precision = 77.6%, recall = 76.4%). These results suggest that an off-the-shelf, pretrained MLLM can detect DTG behavior in visual data with acceptable performance, without the extensive feature engineering and large annotated datasets traditionally required by classical machine learning approaches.

Analysis of the misclassified instances helped us identify key model limitations and inform areas for future improvement. DTG was negatively correlated with learning gains from winter to spring, with a marginally significant association, suggesting that the negative effects of DTG may become more apparent later in the academic year.

Overall, the ability to detect DTG in screen recordings of learners' interactions with a math platform opens new possibilities for analyzing behavior and delivering timely support in digital learning environments. This could help ensure that students engage meaningfully with assigned tasks and ultimately achieve better learning outcomes. As MLLMs continue to improve, they hold considerable potential to make large-scale educational video analysis more reliable and scalable.

5. ACKNOWLEDGMENTS

The authors would like to thank the CueThink team for their partnership and collaboration on this work. The authors also acknowledge Yiqiu Zhou and Maciej Pankiewicz for their support in setting up the model.

This research was supported by NSF Award #DRL-2300829. Any opinions, findings, and conclusions expressed in this paper are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

6. REFERENCES

[1] Appleton, J. J., Christenson, S. L., & Furlong, M. J. (2008). Student engagement with school: Critical conceptual and methodological issues of the construct. Psychology in the Schools, 45(5), 369-386.

[2] Arroyo, I., Ferguson, K., Johns, J., Dragon, T., Meheranian, H., Fisher, D., Barto, A., Mahadevan, S., & Woolf, B. P. (2007). Repairing disengagement with non-invasive interventions. Proceedings of the 13th International Conference on Artificial Intelligence in Education (pp. 195–202)

[3] Ashcraft, M. H., & Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance. Journal of Experimental Psychology: General, 130(2), 224–237.

[4] Baker, R. S. (2007). Modeling and understanding students' off-task behavior in intelligent tutoring systems. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 1059-1068).

[5] Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z.: Off-Task Behavior in the Cognitive tutor classroom: when students "Game The System". In: Proceedings of ACM CHI 2004: Computer-Human Interaction, pp. 383–390. Vienna, Austria (2004)

[6] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1), 289-300.

[7] Braun, V., & Clarke, V. (2022). Conceptual and design thinking for thematic analysis. Qualitative psychology, 9(1), 3.

[8] Buckley, B., Gobert, J., Horwitz, P., & O'Dwyer, I. (2010) Looking inside the black box:assessments and decision-making in BioLogica. International Journal of Learning Technology, 5 (2), 166190.

[9] Carey, E., Hill, F., Devine, A., & Szűcs, D. (2017). The modified abbreviated math anxiety scale: A valid and reliable instrument for use with children. Frontiers in Psychology, 8, 11.

[10] Christenson, S. L., Reschly, A. L., & Wylie, C. (2012). Handbook of research on student engagement. Springer.

[11] Cipora, K., Szczygieł, M., Willmes, K., & Nuerk, H. C. (2015). Math anxiety assessment with the AMAS: Applicability and usefulness: Insights from the Polish adaptation. Frontiers in Psychology, 6, 1833

[12] Cocea, M., Hershkovitz, A., Baker, R.S.J.d.: The Impact of Off-task and Gaming Behaviors on Learning: Immediate or Aggregate? In: Proceedings of the 14th International Conference on Artificial Intelligence in Education, 507—514 (2009).

[13] Craig, S. D., Xie, J., Huang, X., Graesser, C., & Hu, X. (2014). The impact of epistemological beliefs on student interactions with an intelligent tutoring system. Proceedings of the International Conference on Intelligent Tutoring Systems (pp. 660–661). Springer.

[14] Curriculum Associates. (2017). i-Ready diagnostic & instruction: Reading and mathematics. North Billerica, MA: Author

[15] Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. Educational Psychology Review, 28, 425–474.

[16] Derry, S. J., Pea, R. D., Barron, B., Engle, R. A., Erickson, F., Goldman, R., Hall, R., Koschmann, T., Lemke, J. L., Sherin, M. G., & Sherin, B. L. (2010). Conducting Video Research in the Learning Sciences:Guidance on Selection, Analysis, Technology, and Ethics. Journal of the Learning Sciences, 19(1), 3–53.

[17] Duda, J. L., & Nicholls, J. G. (1992). Dimensions of achievement motivation in schoolwork and sport. Journal of educational psychology, 84(3), 290.

[18] Foster, J. K., Korban, M., Youngs, P., Watson, G. S., & Acton, S. T. (2024). Automatic classification ofactivities in classroom videos.Computers and Education: Artificial Intelligence,6, 1-13, Article 100207.https://doi.org/10.1016/j.caeai.2024.100207

[19] Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. Review of educational research, 74(1), 59-109.

[20] Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., ... & Sun, X. (2025). Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 24108-24118).

[21] Glazewski, K. D., Hmelo-Silver, C. E., & Lester[1], J. C. (2020, July). Detecting Off-Task Behavior from Student. In Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I (Vol. 12163, p. 55). Springer Nature.

[22] Gobert, J. D., Baker, R. S., & Wixon, M. B. (2015). Operationalizing and detecting disengagement within online science microworlds. Educational Psychologist, 50(1), 43-57.)

[23] Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., ... & Yu, N. (2024). Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13418-13427).

[24] Kim, B., Zyromski, B., Mariani, M., Lee, S. M., & Carey, J. C. (2017). Establishing the factor structure of the 18-item version of the junior metacognitive awareness inventory. Measurement and Evaluation in Counseling and Development, 50(1-2), 48-57.

[25] Kloosterman, P., & Stage, F. K. (1992). Measuring beliefs about mathematical problem solving. School Science and Mathematics, 92(3), 109–115. https://doi.org/10.1111/j.1949-8594.1992.tb12154.x

[26] Levy, M., Jacoby, A., & Goldberg, Y. (2024). Same task, more tokens: the impact of input length on the reasoning performance of large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 15339-15353).

[27] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. Advances in neural information processing systems, 36, 34892-34916.

[28] Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., & Wang, L. (2024). Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning (arXiv:2306.14565).

[29] Liu, X, Zhou, Y., Wei, Z., Baker, R.S., Barany, A., Ocumpaugh, J. (2026). Supporting Qualitative Video Analysis with Multimodal Large Language Models. In Journal of Computer Assisted Learning (in press).

[30] Liu, X., Wei, Z., Baker, R. S., Metcalf, S. J., Zhang, J., Barany, A., Slater, S., Swanson, L. & Gagnon, D. J. (2025). Integrating large language models and machine learning to detect struggle in educational games. In International Conference on Artificial Intelligence in Education (pp. 398-405). Cham: Springer Nature Switzerland.

[31] Marx, J. D., & Cummings, K. (2007). Normalized change. American Journal of Physics, 75(1), 87-91.

[32] Mills, C., D'Mello, S., Bosch, N., & Olney, A. M. (2015, June). Mind wandering during learning with an intelligent tutoring system. In International conference on artificial intelligence in education (pp. 267-276). Cham: Springer International Publishing.

[33] Ramey, K. E., Champion, D. N., Dyer, E. B., Keifert, D. T., Krist, C., Meyerhoff, P., Villanosa, K., & Hilppö, J. (2016). Qualitative Analysis of Video Data: Standards and Heuristics.

[34] Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley interdisciplinary reviews: Data mining and knowledge discovery, 10(3), e1355.

[35] Sabourin, J. L., Rowe, J. P., Mott, B. W., & Lester, J. C. (2013). Considering Alternate Futures to Classify Off-Task Behavior as Emotion Self-Regulation: A Supervised Learning Approach. Journal of Educational Data Mining, 5(1), 9-38.

[36] San Pedro, M. O., Baker, R., Bowers, A., & Heffernan, N. (2013). Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In Educational Data Mining 2013.

[37] Skinner, E. A., & Pitzer, J. R. (2012). Developmental dynamics of student engagement, coping, and everyday resilience. In Handbook of research on student engagement, 21-44. Boston, MA: Springer US.

[38] Sperling, R. A., Howard, B. C., Miller, L. A., & Murphy, C. (2002). Measures of children's knowledge and regulation of cognition. Contemporary Educational Psychology, 27, 51–79.

[39] Tobin, T. J., & Sugai, G. M. (1999). Using sixth-grade school records to predict school violence, chronic discipline problems, and high school outcomes. Journal of emotional and Behavioral Disorders, 7(1), 40-53.

[40] Trad, F., & Chehab, A. (2025). Evaluating the efficacy of prompt-engineered large multimodal models versus fine-tuned vision transformers in image-based security applications. ACM Transactions on Intelligent Systems and Technology, 16(4), 1-22.

[41] Winne PH and Hadwin AF. Hacker DJ, Dunlosky J, and Graesser A. (1998). Studying as self-regulated learning. Metacognition in educational theory and practice. Hillsdale, NJ Lawrence Erlbaum, 277-304.

[42] Wixon, M., Baker, R. S. D., Gobert, J. D., Ocumpaugh, J., & Bachmann, M. (2012). WTF? detecting students who are conducting inquiry without thinking fastidiously. In International Conference on User Modeling, Adaptation, and Personalization (pp. 286-296). Berlin, Heidelberg: Springer Berlin Heidelberg.

[43] Wong, Z. Y., & Liem, G. A. D. (2022). Student engagement: Current state of the construct, conceptual refinement, and future research directions. Educational Psychology Review, 34(1), 107-138.

[44] Wu, J., Gan, W., Chen, Z., Wan, S., & Yu, P. S. (2023). Multimodal Large Language Models: A Survey. 2023 IEEE International Conference on Big Data (BigData), 2247–2256.

[45] Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P. Y. (2023). Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In Companion proceedings of the 28th international conference on intelligent user interfaces (pp. 75-78).

[46] Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2023). Can llms express their uncertainty? An empirical evaluation of confidence elicitation in llms. arXiv preprint arXiv:2306.13063.

[47] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A Survey on Multimodal Large Language Models (arXiv:2306.13549). arXiv. http://arxiv.org/abs/2306.13549

[48] Zhang, D., Yu, Y., Dong, J., Li, C., Su, D., Chu, C., & Yu, D. (2024). MM-LLMs: Recent Advances in MultiModa Large Language Models (arXiv:2401.13601). arXiv. http://arxiv.org/abs/2401.13601

[49] Zhang, J., Andres, J. M. A. L., Hutt, S., Baker, R. S., Ocumpaugh, J., Nasiar, N., ... & Young, T. (2022). Using machine learning to detect SMART model cognitive operations in mathematical problem-solving process. Journal of Educational Data Mining, 14(3), 76-108.

[50] Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., & Li, C. (2024a). Video instruction tuning with synthetic data (arXiv Preprint No. arXiv:2410.02713).

[51] Zhou, Y., Kang, J., & Nguyen, H. (2025). Can We Trust Large Language Models for Video Analysis: An Exploration of Hallucination in Multimodal LLMs. In Proceedings of the 19th International Conference of the Learning Sciences-ICLS 2025, pp. 110-118. International Society of the Learning Sciences.

Columns on Last Page Should Be Made as Close As Possible to Equal Length