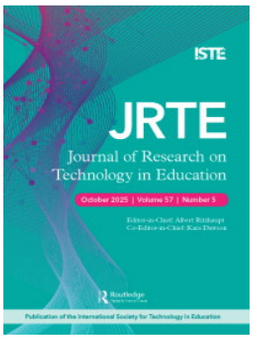

Routledge
Taylor & Francis Group

Journal of Research on Technology in Education

ISSN: 1539-1523 (Print) 1945-0818 (Online) Journal homepage: www.tandfonline.com/journals/ujrt20

Enhancing stealth assessment in game-based learning through goal recognition

Anisha Gupta, Wookhee Min, Dan Carpenter, Roger Azevedo & James Lester

To cite this article: Anisha Gupta, Wookhee Min, Dan Carpenter, Roger Azevedo & James Lester (23 Sep 2025): Enhancing stealth assessment in game-based learning through goal recognition, Journal of Research on Technology in Education, DOI: 10.1080/15391523.2025.2555246

To link to this article: https://doi.org/10.1080/15391523.2025.2555246

JOURNAL OF RESEARCH ON TECHNOLOGY IN EDUCATION
https://doi.org/10.1080/15391523.2025.2555246

Enhancing stealth assessment in game-based learning through goal recognition

Anisha Gupta[a], Wookhee Min[b], Dan Carpenter[b], Roger Azevedo[c] and James Lester[b]

[a]Workstation Software and Solutions Architecture, Lenovo, Morrisville, North Carolina, USA; [b]North Carolina State University, Raleigh, North Carolina, USA; [c]University of Central Florida, Orlando, Florida, USA

ABSTRACT

Stealth assessment in game-based learning analyzes gameplay behaviors to measure student competencies unobtrusively. Grounded in the theoretical premise that in-game objectives shape learning outcomes, this article investigates how goal recognition can enhance stealth assessment by using predictions about immediate gameplay objectives as evidence for modeling learning. We evaluated this approach with 119 middle school students using an educational microbiology game, examining both overall posttest performance and mastery of individual science concepts. Our deep learning architecture combines gameplay interactions, written reflections, and goal recognition to assess different learning dimensions. Results demonstrate that goal recognition significantly improves concept-level assessment across all four concepts, achieving higher accuracy and earlier prediction convergence. Incorporating goal recognition makes stealth assessment more responsive to students' intentions during gameplay, enabling concept-level scaffolding to enhance student learning.

ARTICLE HISTORY
Received 3 April 2025
Revised 17 June 2025
Accepted 26 August 2025

KEYWORDS
Multimodal learning analytics; natural language processing; reflection; game-based learning

Introduction

Digital games are increasingly recognized for their ability to create engaging educational experiences that support learning across different domains (Clark et al., 2016; Qian & Clark, 2016; Plass et al., 2020). However, effectively assessing student knowledge and providing adaptive support in these environments presents unique challenges. Traditional assessment methods, such as quizzes or direct questions, can disrupt the immersive gameplay experience and diminish student engagement. To address this assessment challenge, researchers have developed stealth assessment methods that evaluate student knowledge by analyzing their natural gameplay behaviors (Rahimi et al., 2023). This approach draws from evidence-centered design (ECD) (Mislevy et al., 2003), inferring student competencies based on their choices and problem-solving strategies used to accomplish the overarching objective of the game. Stealth assessment uses fine-grained, in-game interactions as evidence for determining students' competencies, eliminating the need for explicit quiz-like assessments (Rahimi & Shute, 2024).

Open-world educational games provide students with extensive freedom to navigate and explore (Alexander & Martens, 2017; Aung et al., 2019). While this freedom can enhance agency and engagement, it also creates the risk of students struggling to identify which activities are most relevant to their learning goals (Taub et al., 2019; Yew & Goh, 2016). Written reflection prompts have been integrated into these environments as a tool to promote self-regulated learning (Greene et al., 2011; Winne & Hadwin, 1998). These prompts encourage students to pause at

CONTACT Anisha Gupta agupta44@ncsu.edu Workstation Software and Solutions Architecture, Lenovo, Morrisville, NC, USA.
This article has been corrected with minor changes. These changes do not impact the academic content of the article.

key moments during gameplay to consider what they have learned and plan their future actions, helping them maintain focus on learning objectives while providing valuable natural language data about their thinking processes. Reflection in this context refers to students' deliberate examination of their learning experiences and strategic planning, typically expressed through written responses that capture their understanding, reasoning, and intended next steps in the game. These reflection responses provide evidence of student learning and planning in the game and could be used as supplementary evidence to inform stealth assessment models in addition to fine-grained game trace logs.

In addition to determining student competencies, it is also important to identify their in-game strategy to derive a more complete understanding of their learning and provide adaptive support that not only identifies knowledge gaps, but can also provide support that is tailored to the strategy that the student is pursuing. Prior work has explored plan and goal recognition to infer students' strategies based on in-game interactions. Goal recognition is especially challenging to achieve in open-world educational games where in-game goals may not be explicitly defined, some actions might be more exploratory and not necessarily goal-oriented, and students might also be pursuing multiple goals at the same time.

The trajectory taken by students to complete an open-world educational game can directly impact their knowledge competency and learning in the game (Shute et al., 2013). For example, if a student systematically learns from different in-game sources and then applies this learning to strategize toward solving the objective of the game, it could help solidify their knowledge and understanding. Alternatively, if a student employed a strategy that did not explore a certain area of the map, they would miss out on specific learning material placed in that part of the map that could potentially influence their knowledge acquisition for those specific concepts. To this extent, goal recognition could potentially inform stealth assessment models to make more robust predictions about students' knowledge competency.

Building on these foundations, this article investigates how goal recognition can enhance stealth assessment by leveraging predictions about students' immediate gameplay objectives as additional evidence for modeling their learning in reflection-enriched, game-based learning environments. We introduce a framework that integrates pretest performance of the student that represent students' content knowledge before interacting with the learning environment, game trace logs that capture students' moment-by-moment interactions, natural language written reflection responses that provide insight into their thinking and planning, and goal recognition predictions that identify their current objectives in the game environment. Our framework processes these multiple sources of information through a deep learning architecture designed to effectively combine behavioral, textual, and other predictive features. We evaluate stealth assessment at two distinct levels of granularity. Beyond predicting overall posttest performance, we introduce concept-level assessment that maps individual test questions to specific learning objectives covered in the game (disease examples, virus, bacteria, and disease types), enabling more precise identification of areas where students may need additional support. We evaluate the models using both standard metrics for predictive accuracy and specialized metrics for early prediction, which assess how quickly and robustly models can make reliable predictions during gameplay. This early prediction capability is crucial for providing timely adaptive support to students who may be struggling with particular concepts or losing focus on learning objectives.

The remainder of this article is organized as follows. In "Related work," we review related work on stealth assessment, goal recognition, and the use of written reflections in game-based learning. In "Crystal Island test bed," we describe the game-based learning environment and our dataset. We present our methodology in "Methods," including data labeling, feature extraction, and model architecture. Next, we present our experimental results in "Results," comparing different model variants and analyzing the impact of goal recognition on assessment performance. Finally, in "Discussion" and "Limitations" we discuss the implications of our findings and directions for future work.

Related work

Student modeling in game-based learning environments encompasses multiple research areas including stealth assessment, goal recognition, and analysis of written reflections (Geden et al., 2021). While these areas have largely developed independently, their integration presents opportunities for more comprehensive modeling of student learning.

Stealth assessment

Stealth assessment, grounded in evidence-centered design (Mislevy et al., 2003), aims to measure student competencies without disrupting gameplay engagement (Rahimi & Shute, 2024). By analyzing in-game behaviors and interactions, stealth assessment can make inferences about student knowledge without relying on explicit testing that could break game immersion (Shute, 2011; Shute & Ventura, 2013). Fang et al. (2023) demonstrated that game-based stealth assessment effectively predicted standardized literacy measures while maintaining the engaging qualities of gameplay over traditional testing formats. Poole et al. (2025) explored the benefits of leveraging stealth assessment for reading comprehension assessment. This highlights the potential of stealth assessment models for predicting learning outcomes from both trace log as well as textual features. Recent years have seen significant advances in stealth assessment approaches, particularly through the application of deep learning techniques. Min et al. (2020) demonstrated the effectiveness of deep neural networks for predicting learning outcomes based on gameplay traces, while Emerson et al. (2020) showed how multimodal data could enhance assessment accuracy for deep learning-based stealth assessment models. This prior work motivates our stealth assessment model architectures, using sequential deep learning-based models to process game trace logs and students' natural language reflections to predict posttest learning outcomes. A key challenge in stealth assessment is achieving sufficient granularity to identify specific conceptual gaps in student understanding. This has led to increased interest in concept-level assessment approaches that can map gameplay behaviors to particular learning objectives (Emerson et al., 2023; Henderson et al., 2022; Shute et al., 2021). Thus, we evaluate our stealth assessment framework for predicting overall posttest scores as well as concept-level posttest performance.

Goal recognition in games

Goal recognition in games is the process of identifying a player's immediate objectives or intentions based on their observed behavior within a game environment to provide targeted support and intervention. For example, in educational games, goal recognition can help detect when a student is attempting to brute-force their way through challenges without engaging with the learning content, allowing the system to intervene with appropriate guidance. It can also identify when students are pursuing valid strategies but could benefit from an optimized approach, such as recommending a sequence of actions that would better support their learning objectives. Additionally, goal recognition can determine when a student is struggling with a particular task and provide targeted resources to help them succeed, such as highlighting relevant reading materials that align with their current strategy.

Goal recognition presents particular challenges in open-world environments where goals may not be clearly defined and can be discovered through exploration. Early approaches relied primarily on probabilistic models (Ha et al., 2011), but recent work has demonstrated the potential of deep learning techniques. Min, Baikadi, et al. (2016) introduced the use of long short-term memory (LSTM) networks with action embeddings for recognizing player goals, demonstrating how neural approaches could outperform traditional methods. Recent work by Alshehri et al. (2023) introduced an explainable goal recognition framework using weights of evidence to provide human-interpretable explanations for goal predictions. Su et al. (2023) developed a process mining–based approach that learns skill models from historical observations to perform fast and

accurate goal recognition without requiring predefined domain models. Their work achieved accuracy comparable to state-of-the-art approaches while achieving faster recognition times. The integration of multiple data sources has also proven beneficial, with Min et al. (2017) demonstrating how gaze data could enhance recognition accuracy.

Students' in-game strategy and goals can influence what knowledge skills they master while interacting with a game environment. Shute et al. (2013) found correlation between certain in-game actions and accomplishments, and students' performance on assessments in a game-based learning environment designed to teach physics concepts. Similarly, Corredor (2006) related users' strategy and goals while navigating museum websites to their knowledge of the target content. This suggests that students' goals and learning are tightly coupled, and an understanding of in-game strategy and goals can help us get a better understanding of knowledge mastery. Moreover, goals achieved during gameplay also provide more concrete evidence of plans that students might write about in their natural language reflection responses. Thus, incorporating students' strategy and in-game trajectory as supplementary evidence in addition to their game trace logs and natural language reflection responses can help us better inform stealth assessment. To this extent, in this work, we evaluate the effectiveness of incorporating goal recognition as additional input to our stealth assessment models, in terms of improving predictive accuracy and early prediction for both overall posttest score prediction as well as concept-level stealth assessment.

Written reflections in game-based learning

Written reflection is a critical component of self-regulated learning in game-based environments (Greene et al., 2011). Reflection prompts encourage students to pause and consider their learning progress, supporting metacognitive processes that are especially important in open-world games where explicit guidance may be limited. Recent work has explored various approaches to structuring and implementing these reflections effectively. Baßeng and Budke (2024) demonstrated that reflection diaries integrated into gameplay sessions significantly improved students' ability to reflect on game content and connect it to real-world contexts compared to unguided reflection. Their research showed that combining lessons, play phases, and systematic written reflection created an effective learning arrangement for deeper engagement. This aligns with findings from Shaheen et al. (2023), who found that 86.5% of young adult participants responded positively to incorporating reflective design elements in educational games, with features like heads-up displays and progress tracking supporting meaningful reflection. Building on these implementation approaches, researchers have also investigated methods for analyzing reflection data. Gupta et al. (2024) explored automated assessment of reflection quality using pretrained language models, while Geden et al. (2021) demonstrated how reflection data could enhance prediction of learning outcomes when combined with behavioral traces. These findings suggest that written reflections can provide valuable insight into students' thinking processes and learning trajectories. Written reflections include students' natural language accounts of their learning in the game and their plan moving forward, which could complement goal recognition models and trace logs to evidence their gameplay trajectory, motivating us to include them as input features in our stealth assessment framework.

Knowledge modeling with multiple data sources

Effectively modeling student knowledge requires combining evidence from multiple sources to build a comprehensive understanding of learning progress. This is particularly challenging in game-based environments, where diverse data channels, such as behavioral traces, textual input, and assessment results, must be integrated. Recent work has explored various approaches to this challenge. Emerson et al. (2023) investigated how different fusion techniques could combine

multimodal data for knowledge modeling, while Henderson et al. (2022) demonstrated methods for early prediction of learning outcomes using multiple data streams. A key consideration is how different types of evidence complement each other in building accurate models of student understanding. In this work, we primarily use three different feature types as input to our models: textual reflection responses, temporal game trace logs, and students' dynamic in-game goals, as determined by pretrained goal recognition models. Drawing from prior work (Gupta et al., 2024), we present the data fusion technique that worked best for our stealth assessment framework, as well as ablation studies to evaluate the contribution of each feature type toward model performance.

The intersection of these research areas presents an opportunity to enhance stealth assessment by leveraging goal recognition. While previous work has established the value of both behavioral and reflection data for assessment, the potential contribution of goal recognition predictions remains largely unexplored. Understanding a student's immediate objectives could provide valuable context for interpreting their actions and assessing their learning progress, both overall as well as at the concept level. By comparing stealth assessment models with and without goal recognition across both assessment levels, we investigate three key research questions:

RQ1. How does incorporating goal recognition predictions impact the accuracy and early prediction performance of overall posttest score prediction compared to using only game trace logs and written reflections?

RQ2. Can goal recognition enhance concept-level stealth assessment by providing additional context for interpreting students' interactions with specific learning content?

RQ3. What combinations of input features (game traces, reflections, and goal predictions) are most effective for different types of assessment tasks?

These findings have important implications for designing more effective adaptive learning environments. More accurate predictions of student learning, particularly at the concept level, could enable more targeted and timely interventions. By demonstrating how goal recognition can enhance stealth assessment, we also provide new insights into the relationship between students' immediate objectives and their learning outcomes in open-world educational games.

Crystal Island test bed

This study analyzes data from two classroom studies involving middle school students engaging with Crystal Island game-based learning environment. Crystal Island is an open-world game-based learning environment focused on microbiology (Figure 1). In this virtual scenario, students undertake an investigation on a remote island, working to diagnose a disease outbreak and develop a comprehensive treatment strategy. The game provides an open-world environment where students can explore different locations, interact with non-player characters (NPCs) who

Figure 1. (Left) Crystal Island game-based learning environment; (Right) Example of an in-game reflection prompt encouraging metacognition.

assist in collecting evidence through virtual laboratory tests, and access various educational resources including books, articles, and posters that provide information about viruses, bacteria, disease transmission, and different types of illnesses.

Students encounter reflection prompts at key milestones during gameplay that ask them to write about what they had learned and their future plans. These prompts are triggered after students achieve specific in-game goals, such as reading a certain number of resources or obtaining a positive laboratory test result. Two concluding prompts encourage students to analyze their problem-solving strategies and reflect on how they would address similar future challenges.

The dataset includes 119 students who completed both pretest and posttest assessments during the study. Fifty-one percent of the students identified as female and 49% students identified as male. Their ages ranged between 13 and 14 years. Students completed a microbiology content knowledge pretest ($M = 6.78$, $SD = 2.75$, Min = 1, Max = 16) and posttest ($M = 7.36$, $SD = 3.36$, Min = 1, Max = 16) before and after gameplay, with an average gameplay duration of 78.63 min. The dataset comprised 579 written reflections from participants who had completed the posttest, with reflections averaging approximately 20 words in length. Students' reflection responses ranged from brief, nonspecific statements (e.g., "*i learned that the most people are getting sick*") to detailed explanations demonstrating content knowledge and strategic planning (e.g., "*Bacteria can come in many varieties and some are not harmful. E. coli is a type of bacteria. Many things do have bacteria, but some can be non pathogenic, and others can be pathogenic. I plan to test more objects like this*").

Stealth assessment is based on evidence-centered design (ECD), which utilizes three main models—task model, which defines the situations or activities designed to elicit behaviors of the target skills, knowledge, or attributes; evidence model, which specifies how to identify, extract, and interpret evidence from those behaviors; and competency model, which defines the knowledge, skills and attributes to be measured (Mislevy et al., 2003). In Crystal Island, the task model comprises tasks within the game, such as submitting a diagnosis and identifying the illness. The sequence of in-game actions taken by a student in the game, their written reflections, and the goals pursued by students in the game environment serve as evidence for their knowledge competency. In this work, we use this evidence as input to sequential deep-learning models to predict students' learning outcomes. Students' knowledge competency is measured by their score on the 17-item posttest attempted after completion of gameplay. These posttest labels serve as the target that the stealth assessment models aim to predict.

Methods

Students' knowledge competency and learning might be influenced by their gameplay trajectory and in-game goals. While students' game trace logs and reflection responses can hint at their plan and learning in the game, goal recognition models can help us accurately determine their dynamic goals during gameplay. Our framework thus enhanced stealth assessment by leveraging predictions from goal recognition models alongside game trace logs and written reflections to model student learning in game-based environments. We evaluated this approach at two levels of granularity: predicting overall posttest scores and assessing concept-level understanding. The framework processes three main types of input features: (1) fine-grained game trace logs that capture students' moment-by-moment interactions including actions taken, locations visited, and resource usage, (2) natural language written reflections that provide insight into students' thinking and planning processes, and (3) goal recognition predictions that identify students' immediate objectives in the game environment.

For both assessment tasks, we employed sequential deep learning architectures due to their ability to efficiently model the temporal nature of students' gameplay data, handle heterogeneous input types without requiring manual feature engineering, and automatically learn relevant patterns from sequential data. The models process each type of input through specialized subnetworks before combining them for final predictions (based on prior findings in Gupta et al.,

2024). For overall posttest score prediction, the framework outputs a single continuous value representing the predicted score. For concept-level assessment, it produces binary predictions for each of the four target concepts (disease examples, virus, bacteria, and disease types), indicating whether a student has achieved above-median competency in that area.

We compared variants of these models with and without goal recognition predictions to evaluate how this additional evidence impacts assessment performance. Through ablation studies, we also investigated the relative importance of different input types and their combinations. The following subsections detail our approach to processing each type of input feature and the specific architectures used for prediction.

Dataset labeling

Working with domain experts involved in developing Crystal Island and designing the assessment items, we identified four key concepts covered in the game: disease examples, virus, bacteria, and disease types. Disease examples cover specific diseases introduced in the game such as influenza and botulism. Disease types address broader understanding of different categories like pathogens, mutagens, and carcinogens. The virus and bacteria concepts focus on detailed understanding of these microorganisms, including their structure, reproduction, and transmission methods. Each test question was mapped to one or more of these concepts through consultation with the experts. Table 1 shows samples of questions from the knowledge test mapped to the target concepts.

Students' concept-level competence was determined on a scale of 0 to 1 by considering the fraction of questions mapped to the concept that they answered correctly. Since each concept had 2–4 questions mapped to it, this representation of concept-level competence resulted in fractional values that were clustered around discrete values (0, 0.5, 1, etc.), making the distribution more suited to modeling the concept-level stealth assessment task as a classification task than regression. Moreover, modeling concept-level stealth assessment as a classification task helps us leverage early prediction metrics (Min, Baikadi, et al., 2016) defined to evaluate how well our models converge on correct predictions. Finally, modeling concept-level stealth assessment as a classification task also reduces the complexity of the task for ease of prediction using our deep learning–based models, which is crucial given the limited dataset. Their competency on each concept was binarized by considering the median concept-level competence across all students for the target concept, resulting in labels corresponding to either low or high competency for each concept. The distribution of students' concept-level performance on the pretest and the posttest is shown in Figure 2.

In contrast, we modeled overall posttest score prediction as a regression task because posttest scores, determined as fractions over 17 questions, are more continuous in nature and better suited to regression modeling.

For goal recognition, we identified 10 key in-game milestones as target goals: speaking with various NPCs (camp cook, sick patient, lead scientist, virus expert, bacteria expert, lab technician), testing contaminated and uncontaminated samples, submitting a diagnosis, and solving the mystery. Students could accomplish these goals in any order, and completing the game by solving the mystery was possible without achieving all other goals. These milestones were selected

Table 1. Sample mapping from questions to target microbiology concepts.

Question	Target concepts
What role do vaccines play in your immune system?	Disease types
Which of the following statements about pathogens is FALSE?	Disease types
Viruses are known to take which of the following shapes?	Virus
Which of the following can have the largest size?	Virus bacteria
Which of the following diseases is caused by bacterial infection?	Bacteria disease examples
Which of the following statements about Salmonellosis and Influenza is TRUE?	Disease examples

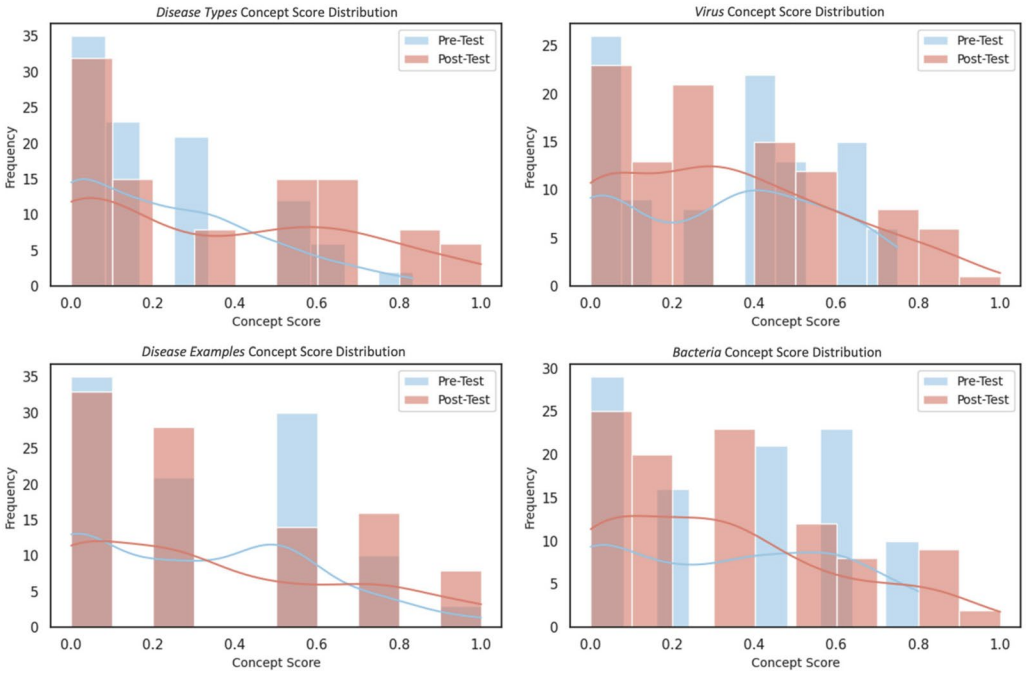

Figure 2. Distribution of student performance across different concepts on pretest and posttest.

based on their potential to inform adaptive gameplay support. Students were not explicitly aware of these goals until they achieved them, thus allowing them to discover the goals naturally through gameplay exploration rather than prescribing any of these goals at the start of the learning experience. For each timestamp of gameplay, the immediate next goal achieved by the student in the game was provided as ground truth label for goal recognition.

Data preprocessing

Our framework processes three distinct types of input features, including game trace logs, written reflections, and goal recognition predictions, along with students' pretest performance. Each requires specific preprocessing steps to transform the raw inputs for the deep learning models.

Game trace features

During gameplay, the system recorded detailed interaction logs including student movements across 24 distinct locations, conversations with NPCs, object interactions, and progress through various goals. We tracked nine different types of actions: movement, editing worksheet, accomplishing goal, NPC conversation, object scanning, reading resources, reflection prompt response, poster interaction, and worksheet submission. Additionally, we maintained records of students' interactions with in-game text resources, including books, articles, posters, and NPC conversations, as these serve as primary sources of content knowledge.

For each action a student performs in the game, we constructed a 43-dimensional feature vector that combined three types of information: (1) the specific action being performed (one-hot encoding based on nine possible actions including movement, editing worksheets, and NPC conversations), (2) the current location (one-hot encoding of the 24 possible locations on the game map), and (3) a binary vector indicating which of the 10 goals have been achieved up to that point. To complement this action-level representation, we also maintained 38-dimensional vector tracking of students' interactions with in-game text resources, including books, articles,

posters, and NPC conversations. This provides additional context about the learning materials students have accessed during gameplay. For timesteps early in gameplay with fewer than 20 previous actions, we used padding to maintain consistent input dimensionality. We maintained a sequence of feature vectors representing the previous 20 actions at each prediction timestep to capture recent gameplay patterns.

Written reflection processing

We represented students' written reflection responses using embeddings from language models (ELMo) pretrained on the 1 Billion Word Benchmark dataset (Chelba et al., 2013). Given our limited dataset of 579 unique reflections, we applied principal component analysis (PCA) to transform the 1,024-dimensional ELMo embeddings to 32-dimensional embeddings, enabling effective modeling while preserving the key variation in reflection content. The model takes as input the embeddings of all previously submitted reflections (maximum five), with padding for timesteps having fewer reflections.

Goal recognition features

We leveraged predictions from a pretrained goal recognition model that was previously validated on this dataset (Gupta et al., 2022). This model produces two sets of predictions through separate sub-models processing gameplay and reflection data. For each prediction timestep, we used both the students' gameplay sub-model output probabilities and their reflection-based sub-model output probabilities as features.

Pretest features

For overall posttest score prediction, we provided the normalized pretest score (0–17 scale) directly as a feature. However, for concept-level predictions, we created a binary vector representing correctness on each pretest item to capture more granular prior knowledge. This difference in pretest feature representations between concept-level and overall posttest score prediction was determined through preliminary analyses showing improved performance for concept-level predictions with the more detailed encoding of prior knowledge.

Model architecture

To investigate our research questions about the effectiveness leveraging goal recognition for informing stealth assessment, we developed model architectures for: (1) a goal recognition model whose predictions will be used as input features for stealth assessment models, (2) stealth assessment models for overall posttest score prediction with and without goal recognition, and (3) concept-level stealth assessment models with and without goal recognition.

Goal recognition model architecture

The goal recognition model uses separate sub-models that process game trace logs and written reflections separately to recognize students' real-time goals before combining their predictions using a shared model architecture (Figure 3), based on preliminary analyses (Gupta et al., 2022). The gameplay-based sub-model processes sequences of student actions through an LSTM layer (32 hidden units) with dropout (0.1) and L2 regularization (0.01) for the kernel, recurrent, and bias terms. This is followed by a dense layer (16 units) with ReLU activation, another dropout layer (0.2), and a final dense layer (10 units) with sigmoid activation.

The reflection-based sub-model processes ELMo embeddings of past reflection responses through a GRU layer (16 hid-den units) with similar dropout and regularization settings. The output passes through a dense layer (16 units), dropout (0.2), and a final dense layer (10 units) with sigmoid activation. The predictions from both sub-models are combined through decision-level

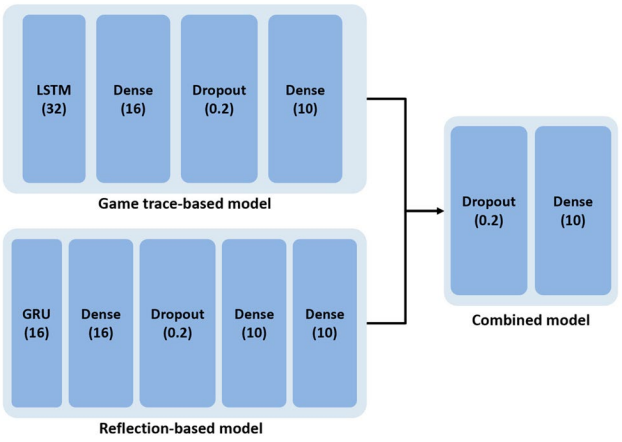

Figure 3. Goal recognition model architecture.

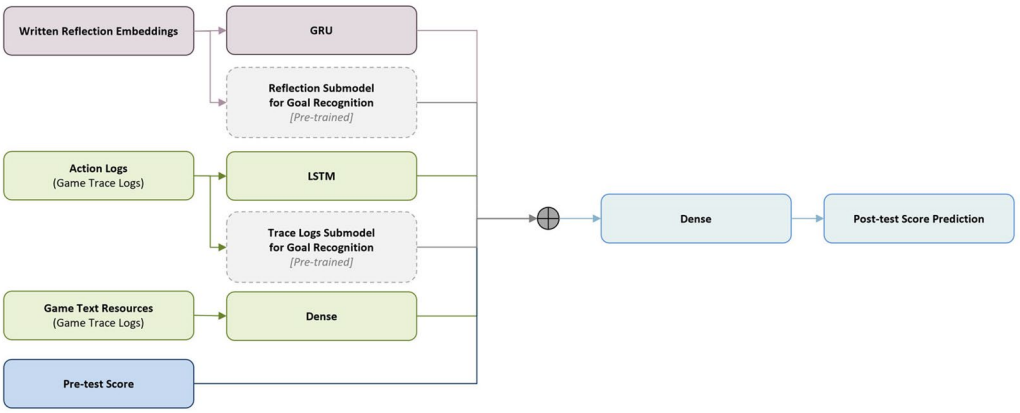

Figure 4. Overall posttest score prediction model architecture.

fusion, with a dense layer merging their outputs to produce final goal probabilities. These probabilities were post-processed to zero out already achieved goals, with the assumption that the same goal cannot be achieved multiple times during a single gameplay session. The gameplay and reflection-based sub-models were pretrained and their parameters frozen before using the predictions as input features for stealth assessment.

Overall posttest score prediction architecture

For predicting overall posttest scores, we processed each input type through specialized neural network layers (Figure 4). Written reflection embeddings are passed through a GRU-based subnetwork consisting of an GRU layer (8 hidden units), followed by dropout (0.2) and a linear transformation (1 unit). Game trace logs are processed by an LSTM-based subnetwork with a 32-unit LSTM layer, dropout (0.2), and a linear transformation (1 unit). Text resource interactions pass through a linear layer (2 units) and dropout (0.5).

For models incorporating goal recognition as input features, we concatenated these processed features with the normalized pretest score and both the gameplay and reflection sub-model predictions from the pretrained goal recognition model (Figure 4). The combined features pass through a final linear layer (1 unit) with sigmoid activation to produce the predicted posttest score. Comparing predictive performance of model architectures with and without goal

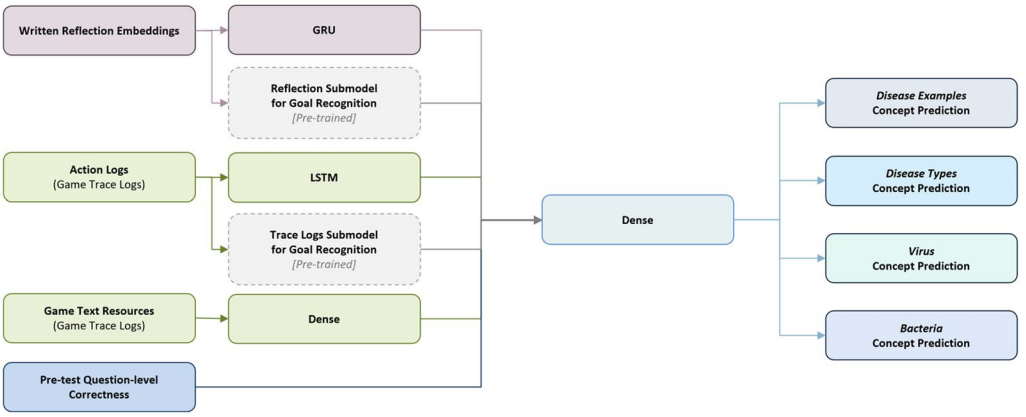

Figure 5. Concept-level stealth assessment model architecture.

recognition would allow us to evaluate how goal recognition predictions impact overall performance assessment (RQ1) and investigate the effectiveness of different feature combinations (RQ3).

Concept-level stealth assessment architecture

For concept-level stealth assessment, we modified the architecture to handle classification across multiple concepts (Figure 5). Game trace logs are processed through an LSTM layer (16 hidden units), written reflections through a GRU layer (4 units), and text resource interactions through a linear layer (4 units). These outputs are combined at an intermediate level with a binary representation of students' question-level correctness on the pretest.

In concept-level stealth assessment models leveraging goal recognition, we incorporated the pretrained model's predictions at this intermediate fusion stage, derived from the best-performing goal recognition model leveraging students' game trace logs and written reflections as input features. The combined representation passes through a shared linear layer (4 units) before branching into concept-specific prediction heads. Each concept has its own linear transformation (1 unit) followed by sigmoid activation to predict whether the student will achieve above-median performance. Comparison of predictive performance of these two concept-level model architecture variants would enable us to evaluate how goal recognition enhances concept-level assessment (RQ2) and examine optimal feature combinations for fine-grained prediction tasks (RQ3).

We trained the stealth assessment models using the Adam optimizer (Kingma & Ba, 2014) with different learning rates determined through preliminary experiments: 0.1 for overall score prediction and 0.001 with weight decay of 0.005 for concept-level prediction. For overall score prediction, we used mean squared error (MSE) as the loss function. For concept-level predictions, we used binary cross-entropy loss for each concept. During training, we employed early stopping with a patience of 10 based on validation set performance to avoid overfitting.

Results

We evaluated our framework using 10-fold cross-validation to ensure robust assessment of model performance. Within each fold of cross validation, we split the training set to use 80% of the data for training and the remaining 20% for validation. To maintain fair comparisons, we used consistent train/test splits across all model variants being compared. Hyperparameters were optimized using grid search across learning rates (0.001, 0.01, 0.1), hidden layer units (8, 16, 32), and dropout rates (0.1, 0.2, 0.5). The optimal configuration achieved through validation was a learning rate of 0.1 for overall posttest score prediction and 0.001 learning rate with weight decay of 0.005 for concept-level prediction.

We assessed model performance using multiple evaluation metrics. For overall posttest score prediction, we used R^2 and MSE, considering the average prediction across all gameplay sequences for each student to account for varying gameplay lengths. For concept-level stealth assessment, which we modeled as a binary classification task, we reported accuracy and F1 scores for each concept. We defined "above-median competence" based on the median concept-level performance across all students for each specific concept. Additionally, we examined early prediction capability through standardized convergence point and convergence rate (CR) to measure how quickly models can make reliable predictions during gameplay. Standardized convergence point (SCP) is computed as $\sum_{i=1}^{n}(k_i / m_i)/n$, where n represents the total number of action sequences corresponding to same target label, m_i is the number of actions in the i^{th} action sequence and k_i depends on the model's convergence for that action sequence; if the model predictions converge, k_i is the action at which the model successfully converges on the correct prediction, while for predictions that do not converge, k_i is computed as $(m_i + p)/m_i$, where p is a constant penalty parameter (Min, Mott, et al., 2016). (In this work, the penalty parameter was set to 1). The CR measures the percentage of action sequences where the model's prediction converges on the correct outcome before the final timestamp (Blaylock & Allen, 2003). Models with lower SCP and higher convergence rates demonstrate better predictive performance.

To benchmark our approach and evaluate the effectiveness of deep learning-based models, we compared against random forest baselines for both regression (overall posttest score prediction) and classification (concept-level stealth assessment) tasks. For classification tasks, we also report majority class baselines across all evaluation metrics. We used one-sided Wilcoxon signed-rank tests to determine if models incorporating goal recognition show statistically significant improvement over their baseline variants.

Through ablation studies, we evaluated different combinations of input features (pretest scores, game traces, and reflections) both with and without goal recognition predictions. This allows us to isolate the specific impact of goal recognition on stealth assessment performance and identify which feature combinations are most effective for different prediction tasks. The following sections present detailed results for overall posttest score prediction and concept-level assessment, examining both predictive accuracy and early prediction capabilities.

Goal recognition model performance

First, we evaluated the performance of our goal recognition model which would later provide predictions as input features for stealth assessment (Table 2). The model predicts students' immediate next goals at each timestep of gameplay from among 10 possible in-game milestones. Using game trace logs as input features, the model achieved 57.29% accuracy and an F1 score of 0.55, significantly outperforming both the random forest baseline (46.40% accuracy, 0.39 F1 score) and majority class baseline (15.00% accuracy, 0.39 F1 score) ($p < 0.05$). The model demonstrated strong early prediction capabilities with an SCP of 0.69 and CR of 62.84%.

Table 2. Comparison of predictive performance for goal recognition, including random forest (RF) and majority class baseline.

Models	Accuracy	F1 score	SCP	CR
Majority baseline	15.00	0.39	0.95	15.00
RF	46.40	0.40	0.97	32.39
Pretest	46.48	0.35	0.77*	35.19
Game trace logs	57.29*	0.55*	0.69*	**62.84***
Reflections	51.57	0.41	0.76*	38.93
Pretest, game trace logs	55.02*	0.50	0.71*	59.23*
Pretest, reflections	50.10	0.38	0.78*	36.96
Game trace logs, reflections	**57.89***	**0.56***	**0.68***	61.55*
Pretest, Reflections, game trace logs	51.60	0.48	0.74*	51.41*

* indicates that the model has significantly better metric results as compared to the RF baseline ($p < 0.05$). The best values for each metric are highlighted in bold.

The model variant using both game trace logs and written reflections showed comparable performance with 57.89% accuracy and 0.56 F1 score (SCP = 0.68, CR = 61.55%). Both model variants significantly outperformed the random forest baseline across all metrics ($p < 0.05$). Since the performance was comparable between these variants, we used the model combining both game trace logs and reflection features to generate goal recognition predictions for our stealth assessment models, maintaining consistency with prior work.

Other feature combinations were also evaluated but showed lower performance. Using only reflection features achieved 51.57% accuracy and 0.41 F1 score, while combining pretest scores with both game traces and reflections achieved 51.60% accuracy and 0.48 F1 score.

Overall posttest score prediction with and without goal recognition

We evaluated whether incorporating goal recognition predictions could improve stealth assessment models' ability to predict students' overall posttest scores (Table 3). The baseline model using pretest scores and written reflections achieved the best performance with an R^2 score of 0.33 and MSE of 0.025, significantly outperforming the random forest baseline (R^2 = 0.23, MSE = 0.029, $p < 0.05$). Adding game trace logs to this model showed similar performance (R^2 = 0.32, MSE = 0.026).

When goal recognition predictions were incorporated alongside all other features (pretest scores, reflections, and game trace logs), the model achieved an R^2 score of 0.28 and MSE of 0.027. The best performing model variant using goal recognition predictions combined pretest scores, reflections, and goal recognition features, achieving an R^2 score of 0.32 and MSE of 0.025. While this significantly outperformed the random forest baseline ($p < 0.05$), it did not show improvement over the best performing model without goal recognition.

Additional feature combinations were evaluated but showed lower performance. Using only game trace logs resulted in poor performance (R^2 = −0.10, MSE = 0.040), while written reflections alone showed modest predictive power (R^2 = 0.06, MSE = 0.035).

Concept-level stealth assessment with and without goal recognition

We evaluated how goal recognition predictions impact stealth assessment models' ability to predict students' competence across four key concepts: disease examples, virus, bacteria, and disease types (Tables 4–7).

For disease examples concept prediction (Table 4), the best model without goal recognition achieved 63.00% accuracy and 0.70 F1 score using game trace logs and reflections. Adding goal recognition predictions improved performance to 66.54% accuracy and 0.72 F1 score. Both models significantly outperformed the random forest baseline (55.82% accuracy, 0.5 F1 score, $p < 0.05$), with the goal recognition-enhanced model also showing better early prediction performance (SCP = 0.61, CR = 64.18%).

Table 3. Comparison of posttest score predictive performance with and without goal recognition as input, including a random forest (RF) baseline.

	Models	R^2 score	MSE
Without goal recognition	RF	0.23	0.029
	Pretest	0.26	0.028
	Game trace logs	−0.10	0.040
	Reflections	0.06	0.035
	Pretest, Game trace logs	0.26	0.027
	Pretest, Reflections	**0.33***	**0.025***
	Game trace logs, Reflections	0.06	0.034
	Pretest, Reflections, Game trace logs	0.32*	0.026*
With goal recognition	Pretest, Reflections, Game trace logs, Goal recognition	0.28	0.027
	Pretest, Reflections, Goal recognition	0.32*	0.025*

* indicates that the model has significantly better metric results as compared to the RF baseline ($p < 0.05$). The best values for each metric are highlighted in bold.

Table 4. Comparison of predictive performance for disease examples concept with and without goal recognition as input, including random forest (RF) and majority class baseline.

	Models	Accuracy	F1 score	SCP	CR
Without goal recognition	Majority baseline	50.6	0.49	0.51	49.20
	RF	55.82	0.50	0.51	54.19
	Pretest	60.74	0.60	**0.40**	59.82
	Game trace logs	59.10	0.67	0.63	61.88
	Reflections	61.78	0.70	0.53	**68.10**
	Pretest, game trace logs	60.89	0.67	0.57	61.72
	Pretest, Reflections	61.40	0.68	0.66	61.06
	Game trace logs, reflections	63.00	0.70	0.62	56.75
	Pretest, reflections, game trace logs	61.57	0.69	0.63	59.03
With goal recognition	Game trace logs, reflections, goal recognition	64.91	0.71	57.83	65.05
	Pretest, reflections, game trace logs, goal recognition	**66.54***	**0.72**	0.61	64.18

* indicates that the model using goal recognition as input performs significantly better than the best-performing model without goal recognition input ($p < 0.05$). The best values for each metric are highlighted in bold.

Table 5. Comparison of predictive performance for virus concept with and without goal recognition as input, including random forest (RF) and majority class baseline.

	Models	Accuracy	F1 score	SCP	CR
Without goal recognition	Majority baseline	55.46	0.54	0.46	54.36
	RF	56.43	0.57	**0.44**	55.91
	Pretest	55.62	0.53	0.47	53.40
	Game trace logs	54.78	0.56	0.57	54.32
	Reflections	60.05	0.66	0.57	59.36
	Pretest, game trace logs	58.26	0.60	0.46	57.01
	Pretest, reflections	59.25	0.63	0.60	59.75
	Game trace logs, reflections	59.19	0.64	0.63	57.75
	Pretest, reflections, game trace logs	59.66	0.61	0.52	55.88
With goal recognition	Reflections, goal recognition	58.76	0.65	0.61	53.08
	Pretest, reflections, game trace logs, goal recognition	**65.14***	**0.67**	0.46	**64.37***

* indicates that the model using goal recognition as input performs significantly better than the best-performing model without goal recognition input ($p < 0.05$). The best values for each metric are highlighted in bold.

For the virus concept (Table 5), the best performance without goal recognition was achieved using only reflections (60.05% accuracy, 0.66 F1 score). Incorporating goal recognition alongside all features improved performance to 65.14% accuracy and 0.67 F1 score, with notably better early prediction metrics (SCP = 0.46, CR = 64.37% compared to SCP = 0.57, CR = 59.36%).

The bacteria concept showed the highest overall improvement with goal recognition (Table 6). The best model without goal recognition achieved 71.12% accuracy and 0.75 F1 score using all features. Adding goal recognition improved this to 75.63% accuracy and 0.77 F1 score, while also enhancing early prediction capabilities (SCP from 0.40 to 0.38, CR from 68.58% to 74.67%).

For disease types concept prediction, the best model without goal recognition achieved 68.63% accuracy and 0.71 F1 score using all features (Table 7). Including goal recognition improved performance to 72.60% accuracy and 0.73 F1 score, with improved early prediction metrics (SCP from 0.40 to 0.37, CR from 70.50% to 72.17%).

Across all four concepts, models incorporating goal recognition predictions consistently showed improved performance over their counterparts without goal recognition. The improvements were

Table 6. Comparison of predictive performance for bacteria concept with and without goal recognition as input, including random forest (RF) and majority class baseline.

	Models	Accuracy	F1 score	SCP	CR
Without goal recognition	Majority baseline	59.65	0.40	0.60	39.99
	RF	59.62	0.50	0.45	45.32
	Pretest	59.88	0.60	0.40	60.36
	Game trace logs	59.78	0.60	0.40	60.10
	Reflections	65.06	0.71	0.51	66.76
	Pretest, game trace logs	66.47	0.66	**0.38**	62.71
	Pretest, reflections	65.97	0.71	0.51	66.19
	Game trace logs, reflections	66.99	0.72	0.50	62.95
	Pretest, reflections, game trace logs	71.12	0.75	0.40	68.58
With goal recognition	Pretest, reflections, game trace logs, goal recognition	**75.63***	**0.77**	**0.38**	**74.67***

* indicates that the model using goal recognition as input performs significantly better than the best-performing model without goal recognition input ($p < 0.05$). The best values for each metric are highlighted in bold.

Table 7. Comparison of predictive performance for disease types concept with and without goal recognition as input, including random forest (RF) and majority class baseline.

	Models	Accuracy	F1 score	SCP	CR
Without goal recognition	Majority baseline	58.10	0.57	0.44	56.86
	RF	60.22	0.64	0.36	60.95
	Pretest	67.20	0.65	**0.35**	65.19
	Game trace logs	58.53	0.57	0.43	56.97
	Reflections	61.16	0.69	0.54	62.02
	Pretest, game trace logs	64.74	0.66	0.36	66.41
	Pretest, reflections	57.66	0.65	0.57	62.89
	Game trace logs, reflections	59.31	0.67	0.59	61.28
	Pretest, reflections, game trace logs	68.63	0.71	0.40	70.50
With goal recognition	Pretest, reflections, game trace logs, goal recognition	**72.60***	**0.73**	0.37	**72.17**

* indicates that the model using goal recognition as input performs significantly better than the best-performing model without goal recognition input ($p < 0.05$). The best values for each metric are highlighted in bold.

most pronounced for the bacteria concept (accuracy improvement of 4.51%) and virus concept (accuracy improvement of 5.09%). All concept-level predictions benefited from goal recognition in terms of both predictive accuracy and early prediction metrics, with consistent improvements in SCP and CR values across concepts.

Discussion

Our results demonstrate how goal recognition differentially impacts stealth assessment at varying levels of granularity, providing empirical support for our theoretical framework that students' gameplay trajectories and in-game goals directly influence their concept-specific learning outcomes. At the overall performance level, incorporating goal recognition predictions did not improve posttest score prediction beyond what could be achieved using pretest scores and written reflections (R^2 = 0.33) (RQ1). However, at the concept level, goal recognition consistently enhanced predictive performance across all four concepts, with absolute accuracy improvements ranging from 2.9% to 5.1%. This pattern aligns with our theoretical framework by demonstrating that students' immediate in-game goals, which reflect their chosen learning trajectory, provide stronger signals about their understanding of specific concepts than their overall knowledge acquisition. Students who systematically pursue goals related to learning about viruses (e.g.,

consulting the virus expert, reading virus-related materials) are more likely to master virus-specific concepts, whereas overall performance depends on aggregated learning across all trajectories.

We acknowledge that our comparison between concept-level and overall performance assessment involves different modeling approaches: classification for concepts and regression for overall scores. However, several factors support our interpretation that goal recognition provides greater value for concept-level assessment. First, within each modeling framework, we observe consistent patterns: goal recognition improved all four concept-level classification tasks (with accuracy gains of 2.9–5.1%), while showing limited benefit for overall score regression (R^2 decreased from 0.33 to 0.28 when adding goal recognition). Second, this differential impact aligns with our theoretical framework—students' immediate in-game goals (e.g., testing contaminated samples, consulting the virus expert) have direct conceptual relevance to specific learning objectives, whereas overall performance aggregates across all concepts and may be more strongly influenced by general factors like prior knowledge.

The impact of goal recognition varied across different concepts, with the most substantial improvements observed for the bacteria concept (accuracy increase of 4.51%) and virus concept (accuracy increase of 5.09%) (RQ2). These differences support our trajectory-based learning theory by suggesting that certain concepts benefit more from goal-oriented exploration patterns. For instance, understanding bacteria concepts may require a more structured sequence of goals (e.g., first learning about microorganisms, then testing samples, and consulting the bacteria expert), making goal recognition particularly valuable for predicting mastery of this concept. More importantly, incorporating goal recognition improved early prediction metrics (SCP and CR) across all concepts. These consistent improvements in early prediction capabilities suggest that goal recognition helps identify concept-level learning patterns sooner in gameplay and can potentially help in early identification of suboptimal learning trajectories in gameplay, enabling timely interventions before students miss critical learning opportunities.

Analysis of feature effectiveness revealed that different combinations of features yielded optimal performance for different prediction tasks ([RQ3). For overall posttest prediction, combining pretest scores with written reflections proved most effective. However, pretest scores showed varying importance for concept-level predictions, indicating that while overall performance correlates strongly with initial knowledge, concept-specific understanding develops more through gameplay interactions. Interestingly, adding pretest scores to game traces and reflection responses actually reduced goal recognition accuracy (from 57.89% to 51.60%). This finding supports our theoretical framework by highlighting that goal recognition is fundamentally about students' dynamic, context-dependent choices within the current game state rather than their static prior knowledge. Students with similar prior knowledge may pursue entirely different goals based on their current location, recent discoveries, or emerging hypotheses about the game's mystery. Given their less important nature, the limited size of our dataset, and the complexity of our deep learning–based models, pretest scores may introduce noise that obscures these immediate contextual factors that drive students' in-game goals.

Written reflections emerged as a consistently beneficial feature across all concepts, though their optimal combination with other features varied by concept. Disease examples showed best performance with game traces and reflections, while virus concept prediction benefited most from reflections alone. Both bacteria and disease types achieved optimal results using all available features including goal recognition. This variation aligns with our trajectory theory in that differential concepts are embedded differently within the game's learning pathways, requiring different types of evidence to assess mastery.

These findings have several implications for game-based assessment design. Goal recognition appears most valuable for fine-grained, concept-level assessment rather than overall performance prediction, suggesting that students' learning trajectories and in-game goals can be leveraged by stealth assessment models to identify concept-specific knowledge patterns. The improvements in early prediction metrics indicate potential for indicate potential for identifying and redirecting suboptimal trajectories before they result in conceptual gaps in learning.

Moreover, the variation in optimal feature combinations across concepts suggests that assessment systems may need to dynamically adjust their feature utilization based on both assessment granularity and how specific concepts are structured within the game's learning environment.

Our results demonstrate key patterns in how different aspects of student modeling relate to each other. The stronger relationship between immediate goals and concept-specific understanding validates our theoretical premise that the path students take through an open-world educational game directly impacts what they learn. The consistent benefits of written reflections across different concepts indicates their value as a complementary data source that captures students' metacognitive awareness of their learning trajectory. Additionally, the improvement in early prediction capabilities shows that goal recognition can help identify learning patterns earlier in gameplay, supporting the theoretical importance of trajectory-based interventions in open-world educational games.

Limitations

Our analysis is based on data collected from students interacting with the Crystal Island game-based learning environment, focusing specifically on middle-grades microbiology content. While our results demonstrate the value of goal recognition for concept-level assessment in this context, future work should examine whether these benefits generalize across different subject domains and game-based learning environments.

Our goal recognition framework currently operates on a set of 10 predefined milestone goals that students can achieve during gameplay. This discrete representation may not fully capture the continuous nature of learning progress or situations where students pursue multiple goals simultaneously. Additionally, the sequential processing of game interactions and goals may not account for more complex patterns of goal-oriented behavior that emerge during gameplay.

While our findings suggest goal recognition provides more value for concept-level than overall performance assessment, we recognize that different evaluation metrics (classification vs. regression) make absolute comparisons challenging. Future work could explore consistent modeling approaches across both assessment levels to strengthen these comparisons. Including more questions per concept on the knowledge tests could help us achieve a more continuous distribution for concept-level stealth assessment labels to support regression analyses.

Finally, while our current analysis demonstrates the potential of using goal recognition to enhance stealth assessment, we performed all evaluations offline using collected data. Implementation in live classroom settings would require careful consideration of computational requirements and processing latency to ensure timely delivery of predictions. Future work should investigate the practical aspects of deploying such a system in real educational environments, including how teachers and students would interact with trajectory-based interventions informed by stealth assessment models enhanced with goal recognition.

Conclusion

Game-based learning environments face a critical challenge in assessing student knowledge without disrupting the immersive gameplay experience. While stealth assessment offers a promising solution by analyzing gameplay behaviors, existing approaches may not fully capture how students' immediate objectives and goals influence their learning trajectories. This work introduces a framework for enhancing stealth assessment by leveraging goal recognition predictions alongside game trace logs and written reflections, aiming to build more comprehensive and theoretically grounded models of student learning in open-world game-based environments.

Our findings demonstrate that the value of leveraging goal recognition for stealth assessment varies significantly with assessment granularity, aligning with our theoretical framework that

students' chosen gameplay trajectories directly influence their concept-specific learning outcomes. While incorporating goal recognition predictions showed limited benefit for overall posttest score prediction, likely due to the aggregated nature of overall performance, it consistently improved concept-level assessment across all four target concepts, with absolute accuracy improvements ranging from 2.9 to 5.1%. This highlights that students' immediate, context-specific goals provide stronger evidence about their understanding of individual concepts compared to their overall knowledge acquisition. Moreover, goal recognition consistently enhanced early prediction capabilities across all concepts, indicating potential for identifying and proactively addressing suboptimal learning trajectories during gameplay.

Analysis of feature effectiveness revealed varying optimal combinations for different assessment tasks. Overall posttest prediction performed best using pretest scores and written reflections. However, at the concept level, goal recognition and gameplay interactions played more significant roles, whereas pretest scores sometimes introduced noise by masking immediate, context-dependent goal-driven behaviors. This underscores the dynamic and trajectory-based nature of concept-specific understanding, reinforcing the importance of capturing students' immediate, goal-oriented decisions within the game. Written reflections consistently emerged as valuable predictors across all tasks, although their optimal combination with other features varied by concept, supporting the need for flexible and adaptive assessment frameworks.

Future work should explore the generalizability of these findings across different subject domains and diverse game-based learning environments. Investigating how varying levels of game complexity influence the relationship between goal recognition and concept-level assessment will further validate and refine our theoretical framework. Additionally, future research should examine how adaptive scaffolding informed by goal-enhanced concept assessment impacts learning outcomes compared to traditional assessment approaches. Additionally, researchers should pursue real-time implementation strategies in classroom settings, exploring practical considerations for deploying these trajectory-based intervention models to provide timely and effective support during gameplay.

Acknowledgement

This manuscript was refined with language editing and rephrasing support from Anthropic's Claude 3.5 Sonnet (October 2024 version).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by funding from the National Science Foundation (NSF) under Grant DRL-1661202. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

Notes on contributors

Anisha Gupta is a Workstation Software and Solutions Engineer at Lenovo. Her PhD research focused on reflection-enriched student modeling in game-based learning environments. Her current work is focused on building predictive models for improving the performance of workstations.

Dan Carpenter is a Research Scientist in the Department of Computer Science at North Carolina State University, as well as Assistant Director for the NSF AI Institute for Engaged Learning. His research interests focus on the application of artificial intelligence in educational settings.

Wookhee Min is a Senior Research Scientist in the Center for Educational Informatics at North Carolina State University and Managing Director of the NSF AI Institute for Engaged Learning. His research focuses on artificial intelligence, particularly the design, development, and evaluation of AI-driven adaptive learning technologies.

Roger Azevedo is a Pegasus Professor in the School of Modeling Simulation and Training at the University of Central Florida. He is also an affiliated faculty in the Departments of Computer Science and Internal Medicine at the University of Central Florida and the lead scientist for the Learning Sciences Faculty Cluster Initiative. His main research area includes examining the role of cognitive, metacognitive, affective, and motivational self-regulatory processes during learning with advanced learning technologies.

James Lester is the Goodnight Distinguished University Professor in Artificial Intelligence and Machine Learning at North Carolina State University. He is the Director of the Center for Educational Informatics and the Director of the National Science Foundation AI Institute for Engaged Learning. His research centers on transforming education with artificial intelligence.

References

Alexander, R., & Martens, C. (2017, August). Deriving quests from open world mechanics. *Proceedings of the 12th International Conference on the Foundations of Digital Games* (pp. 1–7).

Alshehri, A., Miller, T., & Vered, M. (2023, July). Explainable goal recognition: A framework based on weight of evidence. In *Proceedings of the International Conference on Automated Planning and Scheduling* (Vol. *33*, pp. 7–16).

Aung, M., Demediuk, S., Sun, Y., Tu, Y., Ang, Y., Nekkanti, S., Raghav, S., Klabjan, D., Sifa, R., & Drachen, A. (2019, August). The trails of Just Cause 2: Spatio-temporal player profiling in open-world games. *Proceedings of the 14th International Conference on the Foundations of Digital Games* (pp. 1–11).

Baßeng, G., & Budke, A. (2024). Game on, reflection on: Reflection diaries as a tool for promoting reflection skills in geography lessons. *Education Sciences*, *14*(3), 316. https://doi.org/10.3390/educsci14030316

Blaylock, N., & Allen, J. (2003, August). Corpus-based, statistical goal recognition. *International Joint Conference on Artificial Intelligence* (Vol. *18*, pp. 1303–1308). Lawrence Erlbaum Associates Ltd.

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2013). One billion word benchmark for measuring progress in statistical language modeling. arXiv preprint arXiv:1312.3005.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research*, *86*(1), 79–122. https://doi.org/10.3102/0034654315582065

Corredor, J. (2006). General and domain-specific influence of prior knowledge on setting of goals and content use in museum websites. *Computers & Education*, *47*(2), 207–221. https://doi.org/10.1016/j.compedu.2004.10.010

Emerson, A., Cloude, E. B., Azevedo, R., & Lester, J. (2020). Multimodal learning analytics for game-based learning. *British Journal of Educational Technology*, *51*(5), 1505–1526. https://doi.org/10.1111/bjet.12992

Emerson, A., Min, W., Azevedo, R., & Lester, J. (2023). Early prediction of student knowledge in game-based learning with distributed representations of assessment questions. *British Journal of Educational Technology*, *54*(1), 40–57. https://doi.org/10.1111/bjet.13281

Fang, Y., Li, T., Huynh, L., Christhilf, K., Roscoe, R. D., & McNamara, D. S. (2023). Stealth literacy assessments via educational games. *Computers*, *12*(7), 130. https://doi.org/10.3390/computers12070130

Geden, M., Emerson, A., Carpenter, D., Rowe, J., Azevedo, R., & Lester, J. (2021). Predictive student modeling in game-based learning environments with word embedding representations of reflection. *International Journal of Artificial Intelligence in Education*, *31*(1), 1–23. https://doi.org/10.1007/s40593-020-00220-4

Greene, J., Moos, D., & Azevedo, R. (2011). Self-regulation of learning with computer- based learning environments. *New Directions for Teaching and Learning*, *2011*(126), 107–115. https://doi.org/10.1002/tl.449

Gupta, A., Carpenter, D., Min, W., Rowe, J., Azevedo, R., & Lester, J. (2022, October). Enhancing multimodal goal recognition in open-world games with natural language player reflections. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *18*(1), 37–44.

Gupta, A., Carpenter, D., Min, W., Rowe, J., Azevedo, R., & Lester, J. (2024). Detecting and mitigating encoded bias in deep learning-based stealth assessment models for reflection-enriched game-based learning environments. *International Journal of A rtificial Intelligence in Education*, *34*(3), 1138–1165.

Ha, E. Y., Rowe, J. P., Mott, B. W., & Lester, J. C. (2011, October). Goal recognition with Markov logic networks for player-adaptive games. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *7*(1), 32–39. (https://doi.org/10.1609/aiide.v7i1.12434

Henderson, N., Acosta, H., Min, W., Mott, B., Lord, T., Reichsman, F., Dorsey, C., Wiebe, E., & Lester, J. (2022). *Enhancing stealth assessment in game-based learning environments with generative zero-shot learning*. International Educational Data Mining Society.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Min, W., Baikadi, A., Mott, B., Rowe, J., Liu, B., Ha, E. Y., & Lester, J. (2016). A generalized multidimensional evaluation framework for player goal recognition. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *12*(1), pp. 197–203.

Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Smith, A., Wiebe, E., Boyer, K. E., & Lester, J. C. (2020). DeepStealth: Game-based learning stealth assessment with deep neural networks. *IEEE Transactions on Learning Technologies*, *13*(2), 312–325. https://doi.org/10.1109/TLT.2019.2922356

Min, W., Mott, B. W., Rowe, J. P., Liu, B., & Lester, J. C. (2016, July). Player goal recognition in open-world digital games with long short-term memory networks. *International Joint Conferences on Artificial Intelligence* (pp. 2590–2596).

Min, W., Mott, B., Rowe, J., Taylor, R., Wiebe, E., Boyer, K., & Lester, J. (2017). Multimodal goal recognition in open-world digital games. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *13*(1), 80–86.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i–29. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Plass, J. L., Mayer, R. E., & Homer, B. D. (Eds.) (2020). *Handbook of game-based learning*. MIT Press.

Poole, F. J., Coss, M. D., & Clarke-Midura, J. (2025). Developing stealth assessments to assess young Chinese learners' L2 reading comprehension. *Language Learning and Technology*, *29*(2), 104–131.

Qian, M., & Clark, K. R. (2016). Game-based learning and 21st century skills: A review of recent research. *Computers in Human Behavior*, *63*, 50–58. https://doi.org/10.1016/j.chb.2016.05.023

Rahimi, S., & Shute, V. J. (2024). Stealth assessment: A theoretically grounded and psychometrically sound method to assess, support, and investigate learning in technology-rich environments. *Educational Technology Research and Development*, *72*(5), 2417–2441. https://doi.org/10.1007/s11423-023-10232-1

Rahimi, S., Shute, V., Khodabandelou, R., Kuba, R., Babaee, M., Esmaeiligoujar, S. (2023). Stealth assessment: A systematic review of the literature. *Proceedings of the 17th International Conference of the Learning Sciences-ICLS 2023* (pp. 1977–1978). International Society of the Learning Sciences.

Shaheen, A., Ali, S., & Fotaris, P. (2023). Assessing the efficacy of reflective game design: A design-based study in digital game-based learning. *Education Sciences*, *13*(12), 1204. https://doi.org/10.3390/educsci13121204

Shute, V. J. (2011). Stealth assessment in computer-based games to support leaning. *Computer Games and Instruction*, *55*(2), 503–524.

Shute, V. J., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games* (p. 102). The MIT Press.

Shute, V. J., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C. P., Kuba, R., Liu, Z., Yang, X., & Sun, C. (2021). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games. *Journal of Computer Assisted Learning*, *37*(1), 127–141. https://doi.org/10.1111/jcal.12473

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in Newton's playground. *The Journal of Educational Research*, *106*(6), 423–430. https://doi.org/10.1080/00220671.2013.832970

Su, Z., Polyvyanyy, A., Lipovetzky, N., Sardiña, S., & van Beest, N. (2023). Fast and accurate data-driven goal recognition using process mining techniques. *Artificial Intelligence*, *323*, 103973. https://doi.org/10.1016/j.artint.2023.103973

Taub, M., Mudrick, N. B., Roger, A., Plass, J. L., Mayer, R. E., & Homer, B. D. (2019). Self-regulation, self-explanation, and reflection in game-based learning. In *The handbook of game- based learning*.

Winne, P., & Hadwin, A. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Motivation and self-regulated learning: Theory, research, and applications* (pp. 297–314). Erlbaum.

Yew, E. H., & Goh, K. (2016). Problem-based learning: An overview of its process and impact on learning. *Health Professions Education*, *2*(2), 75–79. https://doi.org/10.1016/j.hpe.2016.01.004