

# From Explanation to Diagnosis: Next Generation Interactive Video Coach with Misstep Awareness

Xiao Jin, Rahul K. Dass, and Ashok K. Goel

Georgia Institute of Technology, Atlanta GA, USA  
{xjin96, rdass7, ag25}@gatech.edu

**Abstract.** Intelligent tutoring systems excel at generating explanations but rarely provide principled diagnosis of where and why a learner is wrong. We introduce a misstep-aware coaching capability for Ivy, a neurosymbolic AI coach, built on a two-model architecture that augments a Task-Method-Knowledge (TMK) model with a new Pedagogical Model (PM) in the context of an online graduate AI course at Georgia Tech. The PM makes instructor diagnostic knowledge explicit and machine-readable by encoding, for each quiz question and incorrect response, the learner’s underlying belief (a brief statement of the incorrect idea or missing knowledge), a TMK locus (the source of the misunderstanding), a misconception type and targeted scaffolding derived from the instructor’s Q&A key. Using quiz questions from the course, we demonstrate a proof-of-concept pipeline that detects and classifies learner errors and generates diagnosis-grounded scaffolding, moving Ivy beyond knowledge retrieval toward diagnostic misstep awareness, and enabling more precise, actionable feedback that supports conceptual change and advances adaptive learning systems in AI in education and the learning sciences.

**Keywords:** Intelligent tutoring systems · Pedagogical Model · Task-Method-Knowledge Model · misstep awareness.

## 1 Introduction

The rapid growth of online education has fundamentally changed how adult learners access higher education and workforce training [13, 1]. Unlike in-person classrooms, online learning is often a more passive experience [3]. Students need real-time, bidirectional interaction so instructors can observe their reasoning process, identify errors as they form, and intervene before misconceptions solidify. Intelligent tutoring systems (ITSS) have demonstrated the potential to improve student performance through timely and individualized feedback [14]. However, they remain limited in their ability to infer the underlying causes of learner errors. This diagnostic gap constrains the effectiveness of feedback, as meaningful instructional support depends not only on explaining what went wrong, but also on identifying why the error occurred [11].

The Interactive Video or Ivy [6] is an AI coach embedded alongside instructional videos deployed in a graduate hybrid AI course in Georgia Tech’s Master of Computer Science program. It combines a large language model with a

structured knowledge representation called a Task-Method-Knowledge (TMK) model [15, 9]. Ivy’s current architecture is a four-stage constrained generation pipeline that classifies learner queries, retrieves structured TMK entries, synthesizes constrained explanations, and optimizes coherence [12]. It effectively explains what a correct step looks like and why it serves the goal. This paper describes the design and proof-of-concept instantiation of a two-model architecture that enables Ivy to move from explanation to misstep-aware coaching.

**RQ:** Can a pedagogical model (PM) fill the diagnostic gap in intelligent tutoring systems by classifying learner missteps?

**RH:** PM-grounded Ivy will outperform TMK-only Ivy on feedback targeting, actionability, transferability, and scaffolding appropriateness, because error classification and underlying belief identification enable feedback that address the learner’s specific misstep rather than the skill topic in general.

## 2 Misconception Detection in Educational AI

Recent work has begun applying LLMs to misconception detection in specific domains. Fathi found that integrating LLM within a inquiry-based framework significantly improved first-year engineering students’ conceptual understanding of thermodynamics and reduced misconceptions [8]. Dimeren demonstrates that transformer-based and clustering methods can automatically detect and classify students’ physics misconceptions with near-expert accuracy, showing strong potential for AI to replace time-intensive human diagnosis of learner misunderstandings [7]. Kokver discovered that AI-based natural language processing models can effectively identify teacher candidates’ misconceptions about the greenhouse effect, offering a scalable alternative to expert human evaluation for diagnosing conceptual understanding [10].

While existing approaches in AI in education focus on content delivery or surface-level error detection, they often fail to identify learners’ misconceptions. Our approach offers a novel framework to address the learners’ specific misconceptions, enabling more targeted feedback and supporting deeper, sustained learning.

## 3 The Two-Model Architecture

### 3.1 TMK-only Ivy Architecture

Ivy’s architecture operates through four sequential stages [6, 5]. First, a question answerability module filters inappropriate or out-of-scope queries. Second, a knowledge retrieval module retrieves the top-k most relevant documents from a given TMK model. Third, a response generation module iteratively synthesizes an answer by incorporating information across the retrieved documents. Finally, a response optimizer refines the output for clarity and conciseness, tailoring verbosity to the nature of the query. This pipeline, powered by GPT-5

nano model via LangChain, allows Ivy to generate structured, accurate, and contextually aligned explanations that consistently outperform standard GPT and RAG-based baselines in system-level developer evaluations[6, 5].

A TMK model for a skill encodes three interlocking structures [15, 9, 2]. Task specifies the goal, its success criteria, and the conditions under which it is pursued. Method encodes a finite-state machine whose states represent world configurations and whose transitions represent executable steps with preconditions and postconditions. Knowledge defines the domain concepts. TMK encodes only the correct procedure with no representation of incorrect beliefs, and therefore cannot differentiate procedural errors.

### 3.2 Pedagogical Model: Schema and Structure

ITs that rely solely on knowledge retrieval risk generating responses that are factually correct yet pedagogically ineffective, as they lack an explicit model of why a student has erred. The pedagogical model (PM) is introduced to address this gap by equipping Ivy with a structured diagnostic layer that moves beyond answer retrieval toward reasoned, misconception-aware tutoring. The PM is a structured JSON artifact authored by instructors for each quiz questions. Each PM record contains the following fields showing in Table 1.

**Table 1.** Fields included in the PM record schema.

Field	Description
question_id	Unique identifier linking the PM record to the assessment item.
question_text	The full text of the question as presented to the learner.
options	A list of answer choices, each associated with a correctness flag.
error_type	One of the four taxonomy categories defined in Section 3.3.
tmk_locus	The specific TMK entry at which the learner’s error is localized.
underlying_belief	Statement of the incorrect belief or missing knowledge.
scaffolding	The recommended feedback form and content.

When a learner selects an incorrect answer, Ivy looks up the corresponding PM record, reads the `error_type` and `tmk_locus`, and selects the scaffolding entry as the seed for its constrained synthesis stage. The LLM then generates a response grounded in both the retrieved TMK entries and the PM scaffolding template, ensuring that the explanation is both procedurally faithful and targeted to the diagnosed misconception. More details about PM construction is described in section 4.

### 3.3 The Four-Category Misstep Taxonomy

The taxonomy draws on Norman’s slip-mistake distinction [17], Reason’s error classification [18], Chi and Roscoe’s conceptual-change framework [4], and Newell

and Simon’s task decomposition theory [16]. While these frameworks were developed in general cognitive contexts, we adapt and synthesize them specifically for AI education. Our taxonomy extends prior work by mapping these errors to specific TMK signal, enabling targeted feedback and instructional intervention. Four categories are defined in Table 2.

**Table 2.** Misstep taxonomy with definitions, TMK signals, and feedback focus.

<b>Error Type</b>	<b>Definition</b>	<b>TMK Signal</b>	<b>Feedback Focus</b>
Execution Slip	Local notation or syntax error with intact plan and reasoning.	Local mismatch; preconditions and goal alignment remain valid.	Corrective cue identifying the specific fix, preserving the plan.
Conceptual Gap	A concept or mechanism is missing	Missing knowledge, cannot explain preconditions	Explain missing concept or constraint; understanding check.
Causal Misunderstanding	Violates a required condition or the claimed effect.	Violated precondition or incorrect postcondition in the method transition.	Clarification of the violated causal link and its consequence.
Teleological Error	Jumps to an inapplicable technique; omits required subgoal.	Steps or subgoals are misordered, preventing progress toward the goal.	Restatement of the goal and guidance toward the correct subgoal.

### 3.4 Pairing TMK and PM: The Diagnosis Pipeline

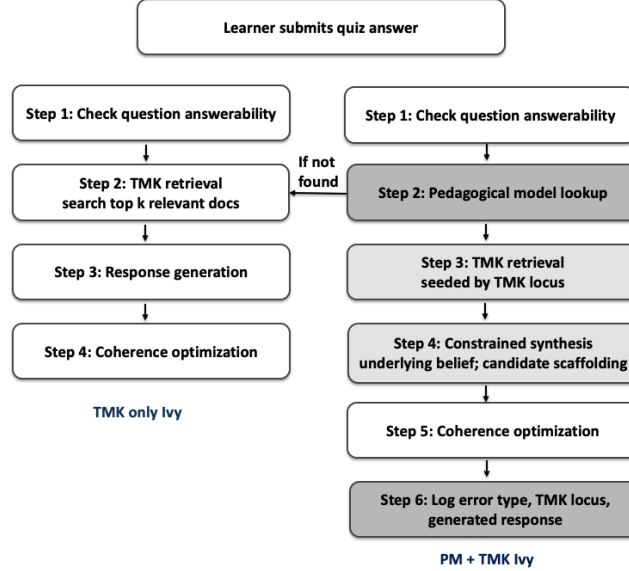
When a learner submits an answer or question, Ivy executes a five-step diagnosis pipeline. First, it classifies the query scope same as the existing TMK only Ivy. Second, if a quiz submission is detected, the system looks up the PM record. If the similarity score is below a threshold (0.5), indicating that no matching PM was found, the process falls back to the TMK-only Ivy pipeline for TMK retrieval. Otherwise, it proceeds to the third step. Third, it retrieves the relevant TMK entries specified by the `tmk_locus`. Fourth, it synthesizes a scaffolding response constrained by both the PM scaffolding template and the retrieved TMK entries. Fifth step is the same as the existing one. Finally, the system logs the identified error type, the associated TMK locus, and the generated response. The diagram (Figure 1) summarizes how the pedagogical model layer integrates with Ivy’s existing four-stage pipeline at runtime. Shaded boxes are new components introduced.

## 4 Proof-of-Concept Instantiation

### 4.1 Source Material

PMs were constructed from quiz questions in the Spring 2026 offering of CS 7637: Knowledge-Based AI(KBAI) at Georgia Institute of Technology. The source ma-

Fig. 1. Ivy Pedagogical Model Integration Pipeline



terial comprised: (1) the full text of each question and its answer choices; (2) the expert authored Q&A key specifying the correct answer and, for each incorrect answer, a brief rationale for why a learner might select it; and (3) the TMK models previously authored for the three targeted skills. In total, 9 questions spanning three skill areas were processed.

Three skill areas were instantiated, yielding 9 PM records and 22 incorrect answer pairs: MetaReasoning (1 question, 3 incorrect answers), Production Systems (5 questions, 13 incorrect answers), Scripts (3 questions, 6 incorrect answers).

## 4.2 PM Construction Process

Each PM file captures a diagnostic profile of a specific learning failure, including an error type, a TMK locus identifying where corresponding knowledge representation occurs, the underlying false belief driving the error, and scaffolding strategies for remediation. These files are stored in a dedicated repository and indexed at system initialization into a FAISS vector store, where each PM document is embedded in a OpenAI’s text embedding model. This one-time indexing step transforms the expert-curated misconception library into a retrievable semantic index, enabling efficient nearest-neighbor lookup at inference time without repeated embedding overhead.

Upon receiving a student query, the system computes its semantic embedding and performs a similarity search against the PM index, retrieving the single most

proximate misconception archetype subject to a minimum relevance threshold. If a candidate PM clears this threshold, it is passed alongside the student’s question to a large language model acting as a zero-shot classifier, which determines whether the retrieved archetype genuinely characterizes the student’s error and populates the remaining diagnostic fields. A confirmed match causes the system to bypass the TMK vector store entirely, instead injecting the structured pedagogical context directly into the generation prompt, thereby grounding the tutoring response in a theoretically motivated model of the student’s specific misconception rather than generic domain knowledge retrieval.

## 5 Preliminary Evaluation

### 5.1 Error Classification Analysis

Across all 22 incorrect answers, the distribution of error types was: causal misunderstanding 45% (n=10), teleological error 14% (n=3), conceptual gap 23% (n=5), and execution slip 18% (n=4). To obtain a preliminary estimate of PM-guided diagnosis coverage, we constructed an evaluation task using the 9 questions across three skills from the KBAI course quizzes. For each question, all candidate answer options were systematically enumerated and submitted independently to the Ivy PM-guided diagnosis pipeline. Each submission simulated a student selection, allowing exhaustive evaluation across all possible response trajectories within the domain.

The pipeline matched each response against the pedagogical model using similarity search, retrieving the most semantically relevant PM record and generating a predicted error-type diagnosis for each response. Across all question-answer pairs, the pipeline achieved 100% PM retrieval coverage. Every submitted response was successfully matched to a PM record and assigned a corresponding error-type label. No response fell below the similarity threshold or failed to retrieve a valid PM entry.

### 5.2 Scaffolding Quality Comparison

To assess whether PM adds diagnostic value beyond Ivy’s existing TMK-only pipeline, we conducted scaffolding quality comparison study. For each of the 38 answer options from 9 questions, we generated two candidate feedback responses: one using Ivy’s existing TMK-only pipeline and one using the PM-augmented pipeline. Raters evaluated each response on five dimensions using a 3-point scale: 0 (incorrect), 1 (partially correct), and 2 (correct). The dimensions as follows:

- **Accuracy:** Whether the feedback is correct.
- **Targeting:** Whether the feedback addresses the specific misconception.
- **Actionability:** Whether the feedback provides a clear path to attempt.
- **Transferability:** Does the feedback help the learner recognize when the skill applies in a new context.

- **Scaffolding Appropriateness:** Whether the form of feedback matches the nature of the learner’s error.

Table 3 reports mean scores (0–2 scale) and paired comparison p-values across five feedback quality dimensions for TMK-only and PM-augmented Ivy responses. PM-grounded responses outperformed TMK-only responses on all five dimensions, with statistically significant gains on four of the five.

**Table 3.** Comparison of average feedback quality(n = 38).

	Accuracy	Targeting	Action-ability	Transfer-ability	Scaffolding Appropriateness
TMK only	1.42	0.95	0.47	0.55	0.71
PM + TMK	1.58	1.50	1.32	1.32	1.42
p value	0.2712	0.0011	1.22E-06	1.33E-05	3.91E-05

The most significant improvements appeared on actionability ( $M = 1.32$  vs.  $M = 0.47$ ,  $p < .001$ ) and transferability ( $M = 1.32$  vs.  $M = 0.55$ ,  $p < .001$ ). This pattern is interpretable: without a PM record specifying the underlying incorrect belief and a recovery path, TMK-only responses tend to re-explain the correct procedure but leave the learner without a concrete next step or a sense of when the skill applies in a new context. Scaffolding appropriateness showed a similarly strong effect ( $M = 1.42$  vs.  $M = 0.71$ ,  $p < .001$ ), indicating that PM-grounded responses more reliably matched the feedback form to the diagnosed error type, consistent with the theoretical motivation for the taxonomy.

Targeting also improved significantly ( $M = 1.50$  vs.  $M = 0.95$ ,  $p = .001$ ), confirming that PM-grounded responses more consistently addressed the learner’s specific error rather than the general skill topic. Accuracy was the one dimension that did not reach statistical significance ( $M = 1.58$  vs.  $M = 1.42$ ,  $p = .271$ ), suggesting that both pipelines produce the correct characterizations of the skill. The TMK model alone is sufficient for factual correctness, although that correctness alone does not translate into targeted, actionable, or transferable feedback without the diagnostic layer the PM provides.

These results indicate that the diagnostic gap in TMK-only coaching is not a gap in factual accuracy but in pedagogical specificity. Knowing the correct procedure is not sufficient to tell a learner why their particular move was wrong, what to do next, or how to recognize a similar situation in the future. The PM layer addresses precisely these deficiencies

## 6 Discussion and Limitations

The strong effects on actionability and transferability are particularly notable. TMK-only responses score ( $M = 0.47$  and  $M = 0.55$  respectively), suggest that without the underlying belief and scaffolding template, Ivy defaults to

re-explaining the correct procedure. The response that may confirm what the learner should have done without helping them understand why they did not, or how to proceed. PM-grounded responses raised both dimensions to  $M = 1.32$ , a shift that raters described qualitatively as the difference between being told the answer and being shown a path forward. The improvement in scaffolding appropriateness ( $M = 0.71$  to  $M = 1.42$ ) further suggests that the four-category taxonomy provides appropriate feedback forms in a way that TMK signals alone cannot reliably support.

Nonetheless, several limitations bound the interpretation of these findings. The evaluation sample is small. Twenty-two responses spanning across three skills provide sufficient basis for a proof-of-concept but not for generalizable claims about PM performance across skill domains, question types, or learner populations. In addition, this study assesses only response quality as judged by expert evaluators. It does not measure whether PM-grounded scaffolding actually helps students correct their errors, reduces error recurrence within a session, improves post-test performance, or raises self-efficacy.

## 7 Conclusion and Next Steps

This paper has presented the design and proof-of-concept instantiation of a pedagogical model that extends Ivy’s TMK-grounded coaching architecture from explanation generation to misstep-aware diagnosis. The PM schema enables Ivy to classify learner errors into one of four theoretically grounded categories, localize the error within the skill’s goal hierarchy, infer the underlying incorrect belief, and deliver targeted scaffolding response.

We are currently working on scaling this next generation AI coaching agent from proof-of-concept to a fully deployed and validated system. We are scaling it to the full course and ultimately across multiple courses. In addition, we will expand the PM framework beyond multiple-choice questions to the full range of question types. Open-ended submissions require a different diagnostic approach in which Ivy must infer the error type from the learner’s free-form reasoning rather than look it up in a table. Extending the PM schema to encode error-type signatures combined with LLM-assisted classification constrained by TMK traces, is the natural next step. Furthermore, we would implement learner-model update across sessions. The current architecture diagnoses each submission independently, with no memory of the learner’s prior error patterns. A longitudinal learner model would track error-type recurrence for each individual. Learners who repeatedly exhibit the same error type would receive progressively more explicit intervention. Implementing this requires a persistent learner state representation linked to the PM taxonomy, session-level aggregation of error-type counts, and a scaffolding policy that conditions on both the current diagnosis and the learner’s accumulated error history. These extensions would help evolve Ivy into a dynamic longitudinal coaching system that supports learners throughout their learning journey.

**Acknowledgments.** This research has been supported by a US National Science Foundation grant to the National AI Research Institute of AI for Adult Learning and Online Education (AI-ALOE; aialoe.org). We thank members of the Ivy project team at Georgia Tech’s Design Intelligence Laboratory for their contributions to this work.

## References

1. Castro, M.D.B., Tumibay, G.M.: A literature review: efficacy of online learning courses for higher education institution using meta-analysis. *Education and Information Technologies* **26**(2), 1367–1385 (2021)
2. Chandrasekaran, B., Johnson, T.R., Smith, J.W.: Task-structure analysis for knowledge modeling. *Communications of the ACM* **35**(9), 124–137 (1992)
3. Chi, M.T.H., Wylie, R.: The icap framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist* **49**(4), 219–243 (2014). <https://doi.org/10.1080/00461520.2014.965823>, <https://doi.org/10.1080/00461520.2014.965823>
4. Chi, M.T., Roscoe, R.D.: The processes and challenges of conceptual change, pp. 3–27. Springer (2002)
5. Dass, R., Bowlin, T., Li, Z., Jin, X., Goel, A.: Improving procedural skill explanations via constrained generation: A symbolic-llm hybrid architecture (2025), <https://arxiv.org/abs/2511.20942>
6. Dass, R.K., Madhusudhana, R.H., Deye, E.C., Verma, S., Bydlon, T.A., Brazil, G., Goel, A.K.: Ivy: A hybrid knowledge-based and generative ai coach for explaining procedural skills. In: Cristea, A.I., Walker, E., Lu, Y., Santos, O.C., Isotani, S. (eds.) *Artificial Intelligence in Education*. pp. 233–246. Springer Nature Switzerland, Cham (2025)
7. Demirezen, M.U., Yilmaz, O., Ince, E.: New models developed for detection of misconceptions in physics with artificial intelligence. *Neural Computing and Applications* **35**(12), 9225–9251 (2023)
8. El Fathi, T., Saad, A., Larhzil, H., Lamri, D., Al Ibrahim, E.M.: Integrating generative ai into stem education: Enhancing conceptual understanding, addressing misconceptions, and assessing student acceptance. *Disciplinary and Interdisciplinary Science Education Research* **7**(1), 6 (2025)
9. Goel, A.K., Rugaber, S.: Gaia: A cad-like environment for designing game-playing agents. *IEEE Intelligent Systems* **32**(3), 60–67 (2017)
10. Kökver, Y., Pektaş, H.M., Çelik, H.: Artificial intelligence applications in education: Natural language processing in detecting misconceptions. *Education and Information Technologies* **30**(3), 3035–3066 (2025)
11. Kuhail, M.A., Alturki, N., Alramlawi, S., Alhejori, K.: Interacting with educational chatbots: A systematic review. *Education and Information Technologies* **28**(1), 973–1018 (2023)
12. Lum, C., Deye, E., Brazil, G., Bydlon, T., Verma, S., Madhusudhana, R., Dass, R., Goel, A.: Designing an ai coaching system for interactive video-based skill learning. In: Graf, S., Markos, A. (eds.) *Generative Systems and Intelligent Tutoring Systems*. pp. 281–291. Springer Nature Switzerland, Cham (2026)
13. Means, B., Neisler, J., et al.: Suddenly online: A national survey of undergraduates during the covid-19 pandemic. Tech. rep., Digital Promise (2020)
14. Mousavinasab, E., Zarifasanaiey, N., R. Niakan Kalhori, S., Rakhshan, M., Keikha, L., Ghazi Saedi, M.: Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interactive learning environments* **29**(1), 142–163 (2021)

15. Murdock, J.W., Goel, A.K.: Meta-case-based reasoning: self-improvement through self-understanding. *Journal of Experimental & Theoretical Artificial Intelligence* **20**(1), 1–36 (2008)
16. Newell, A., Simon, H.A.: *Human problem solving*. Prentice-hall (1972)
17. Norman Donald, A.: *The design of everyday things*. MIT Press (2013)
18. Reason, J.: *Human error*. Cambridge university press (1990)