

Detecting Knowledge Gaps from Conversational AI Interactions Using Curriculum Prerequisite Graphs

Youssef Medhat^{1,*}, Junsoo Park¹, Ploy Thajchayapong¹ and Ashok K. Goel¹

¹Georgia Institute of Technology, Atlanta, GA, USA

Abstract

Large online courses generate thousands of student questions directed at conversational AI teaching assistants, yet these interaction logs remain largely untapped as diagnostic signals. We present a pipeline that maps student questions from a conversational AI teaching assistant to curriculum topics using a few-shot text classifier, grounded in a GPT-4-extracted prerequisite knowledge graph of course concepts. Evaluated on 1,340 question events from 164 students in a graduate-level AI course, our classifier achieves 80.0% accuracy across 43 labels (42 curriculum topics plus an “unknown” abstention class). Topic-level question volume correlates significantly with student self-reported difficulty from an independent mid-semester survey ($\rho = 0.491$, $p = 0.008$, $n = 28$ topics), providing convergent evidence that the classified question stream reflects genuine topic difficulty. These results demonstrate that conversational AI interaction logs, mapped onto curriculum structure, carry actionable signals about topic-level knowledge gaps and provide instructors with a curriculum-grounded view of which topics warrant attention.

Keywords

knowledge gap detection, conversational AI teaching assistants, curriculum knowledge graph, prerequisite structure, few-shot text classification, educational data mining, learning analytics

1. Introduction

Conversational AI teaching assistants have been deployed in large-scale online courses to answer student questions around the clock [1]. Each submitted question is a behavioral trace: it reveals which topic a student is thinking about, when, and, indirectly, whether they have understood prior material [2]. Despite the scale of these interaction logs, relatively little prior work has analyzed them for individualized diagnostic purposes beyond aggregate dashboards [3, 4].

Existing approaches to knowledge gap detection rely on structured item responses (correct or incorrect answers to assessments), which require carefully designed item banks aligned to a skill taxonomy [5, 6]. Free-form conversational questions are harder to analyze: they must first be mapped to curriculum topics, and the diagnostic signal is implicit. A student asking many questions about a topic, or asking about it weeks after it was taught, is exhibiting a behavioral signature of difficulty, but extracting this signal requires both a text classifier and a curriculum model.

This paper presents a pipeline that bridges these two requirements. We use a GPT-4-extracted prerequisite knowledge graph of course topics and a few-shot text classifier (FastFit [7]) to transform raw student questions from a conversational AI teaching assistant into curriculum-grounded topic-level signals. The design choice of treating *topics* (not individual questions) as the analytic unit is motivated by how student struggle actually manifests in coursework. Students frequently show difficulty with a particular lesson or topic, but the underlying cause is often not local to that topic: it lies in an upstream prerequisite that was not fully consolidated. Treating questions as topic-level signals, and overlaying them on a prerequisite knowledge graph, is what makes these structurally grounded knowledge gaps visible at all. Topic-level aggregation additionally matches the granularity at which an instructor can

CSEDM'26: 10th Educational Data Mining in Computer Science Education (CSEDM) Workshop, June 28, 2026, Seoul, South Korea

*Corresponding author.

✉ ymedhat3@gatech.edu (Y. Medhat); jpark3232@gatech.edu (J. Park); ploy@gatech.edu (P. Thajchayapong); ashok.goel@cc.gatech.edu (A. K. Goel)

ORCID 0009-0002-3343-0149 (Y. Medhat); 0009-0003-4450-0103 (J. Park); 0009-0000-5993-1094 (P. Thajchayapong); 0000-0003-4043-0614 (A. K. Goel)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

act and is empirically more robust to classifier noise than per-question correctness assessment would be. This work-in-progress is organized around a single research question:

RQ. *Can naturally occurring student questions to a conversational AI teaching assistant, mapped onto a curriculum prerequisite knowledge graph, surface topic-level knowledge gaps that match an independent measure of topic difficulty?*

We decompose the RQ into two hypotheses:

- **H1** (Classification): A few-shot classifier maps questions to curriculum topics with sufficient accuracy for downstream topic-level analysis.
- **H2** (Survey Alignment): Topic-level question volume from the classified stream correlates with self-reported difficulty from an independent survey.

Support for H1 establishes that the upstream measurement step is reliable, and support for H2 then constitutes evidence that the resulting topic-level signal reflects genuine course-level knowledge gaps: those topics on which the cohort, in aggregate, struggles.

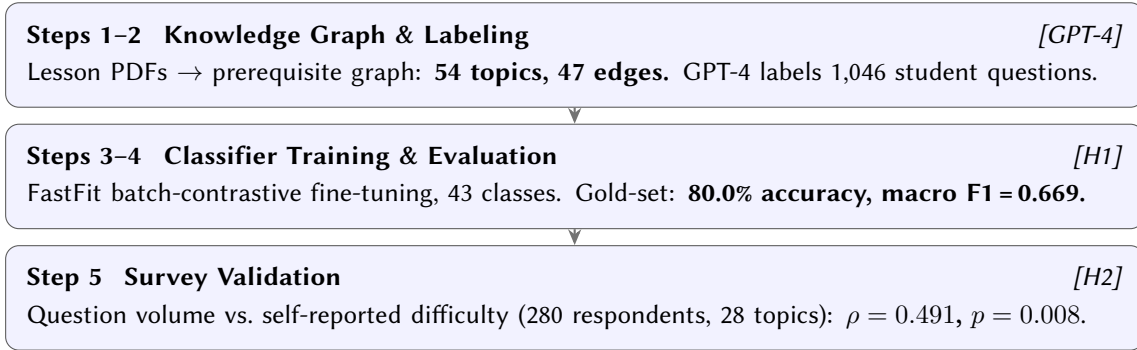


Figure 1: Pipeline overview. H-labels indicate where each hypothesis is evaluated.

2. Related Work

Knowledge tracing. Bayesian Knowledge Tracing [5] and Deep Knowledge Tracing [6] estimate per-skill mastery from structured item-response sequences. Extensions incorporate prerequisite structure [8]. These approaches require aligned item banks and cannot be applied to the free-form conversational questions we study here.

Prerequisite structure in curricula. Prerequisite relationships between topics have been extracted from course syllabi [9], textbooks [10], and knowledge graphs. Prior work uses these structures for curriculum design and concept recommendation, but not for gap detection from conversational logs.

Conversational AI in CS education. Our work differs from prior conversational-AI deployments in CS education in two specific respects: (i) we pair the chat-log signal with an LLM-extracted curriculum prerequisite knowledge graph, and (ii) we validate the resulting topic-level signal against an independent self-reported difficulty survey rather than against engagement or retention proxies. Prior work in this space has largely deployed intelligent tutoring systems and conversational agents in programming courses and STEM contexts [11]. Recent deployments of conversational AI teaching assistants in large online courses field thousands of student queries per semester, and recent work has begun to analyze these chat logs for automated knowledge gap detection [3] and for learning-analytics dashboarding [4]. Complementary lines of work in our group examine the representational and behavioral aspects of these deployments: learner representations derived from chat-log interactions and their evaluation for downstream differentiation [12], the effect of memory-based versus context-only conditioning on the personalization behavior of LLM-based agents [13], and a real-world classroom assessment of bidirectional feedback in multi-tier AI deployments [14].

Few-shot text classification. Sentence-level embedding models such as Sentence-BERT [15] have enabled accurate text classification from limited labeled data by reusing semantic similarity between pre-trained embeddings. SetFit [16] and FastFit [7] build on this by adapting the embedding space with contrastive objectives, achieving competitive accuracy with as few as a single labeled example per class. Our use of FastFit is motivated by this property: several curriculum topics in our setting have highly imbalanced training coverage.

LLM-assisted curriculum mining. Large language models have recently been used to extract structured educational artifacts from unstructured course content, including concept lists, definitions, and prerequisite relations [10]. We use GPT-4 in this role to bootstrap both the knowledge graph and an initial pool of weakly labeled questions; a human-annotated gold set is used to evaluate the downstream classifier.

3. Method

3.1. Knowledge Graph Construction

Role in the pipeline. The prerequisite knowledge graph serves two distinct roles in this work-in-progress. First, the *topic nodes* provide the closed label space the rest of the pipeline depends on: H1 measures classifier accuracy against a gold set whose labels are drawn from this set, and H2 correlates topic-level question volume with the survey’s self-reported difficulty by aligning topic names across the two instruments. Both analyses require a fixed, named topic set; the knowledge-graph extraction is what produces it. Second, the *directed prerequisite edges* encode the curriculum structure that future per-student gap analyses can propagate over: when a student is observed to have difficulty on a downstream topic, the upstream prerequisites in the graph identify candidate root-cause topics that merit review. The current paper focuses on the topic-level signal (H1 and H2); we treat the propagation of per-student gap signals along prerequisite edges as the immediate next analytical step on this data, described in Section 5.

A course instructor with domain expertise used GPT-4 to extract topics and prerequisite relationships from 26 lesson PDFs of a graduate-level AI course at a US R1 university, following a three-stage protocol. **(i) Topic extraction:** each PDF was processed individually and GPT-4 was prompted to list the distinct teachable units introduced in the lesson along with a one-sentence operational definition. **(ii) Topic boundary consolidation:** near-identical topics across lessons were merged; topics the textbook treats as distinct (e.g., Case-Based Reasoning vs. Analogical Reasoning) were kept separate even when vocabulary overlapped. **(iii) Edge extraction:** for each candidate pair, GPT-4 was given the two definitions and asked whether one is a direct prerequisite for the other; ambiguous or bidirectional responses were discarded. The resulting graph contains **54 topic nodes** and **47 directed prerequisite edges** (e.g., *Semantic Networks* $\rightarrow$ *Frames*; *Case-Based Reasoning* $\rightarrow$ *Analogical Reasoning*; *Logic* $\rightarrow$ *Explanation-Based Learning*). The 26 lesson PDFs and the resulting node/edge counts are summarized in Figure 2.

Graph validation. Two pieces of evidence support treating the resulting graph as defensible for the downstream analyses rather than as raw LLM output. First, the edge-extraction stage *discards* every GPT-4 prerequisite judgment that is bidirectional or ambiguous, so the graph only contains the subset of relations the LLM was unambiguous about. Second, representative edges (*Semantic Networks* $\rightarrow$ *Frames*; *Case-Based Reasoning* $\rightarrow$ *Analogical Reasoning*) match the lesson ordering in the course textbook and prior pedagogical work on the same curriculum. We do not claim the resulting graph is the unique correct structure; the protocol is a low-cost, instructor-auditable way to obtain a topic-edge artifact that supports the downstream analyses without the manual effort of writing the graph from scratch.

3.2. Question Classification

LLM-assisted labeling. Training labels for the 1,046-question pool were produced by GPT-4 under a closed-label protocol: the prompt lists all 54 topics with their one-sentence definitions from the

Lesson block	Representative topics	Sample prerequisite edges
Foundations	Foundations of AI; Cognitive Architectures	Cognitive Architectures → Frames
Knowledge rep.	Semantic Networks; Frames; Logic	Semantic Networks → Frames
Problem solving	Means-Ends Analysis; Planning; Sub-Goals	Planning → Sub-Goals
Learning	Generalization; Version Spaces; EBL	Logic → Explanation-Based Learning
Reasoning	CBR; Analogical; Commonsense	Case-Based → Analogical Reasoning
Agents/ethics	Properties of AI Agents; Ethics	Foundations of AI → AI Agents

Figure 2: Representative slice of the 54-topic prerequisite knowledge graph extracted by GPT-4 from 26 lesson PDFs. Each row groups topics taught in the same lesson block with a sample prerequisite edge; the full graph has 47 directed edges.

knowledge-graph stage, followed by the student question, and asks for the single best topic label or the literal string unknown when no topic fits. Because the label space is fixed and enumerated in the prompt, GPT-4 cannot introduce new topics, and the downstream classifier inherits the same bounded label space.

Why not use the LLM directly as the classifier? GPT-4 produces the training labels in the previous step, so it is natural to ask why we add a downstream classifier rather than calling GPT-4 at inference time. We retain the FastFit classifier for two reasons. First, GPT-4 at inference time is non-deterministic and externally versioned; FastFit gives reproducible predictions across runs. Second, on the same 70-question gold set with the same 43-class label space, GPT-4 used directly under the same closed-label prompt scores 0.786 accuracy / 0.654 macro F1, slightly under FastFit’s 0.800/0.669 (Table 1). The downstream classifier therefore matches the LLM’s direct accuracy on this task while removing per-query LLM dependence.

Model. We trained a FastFit [7] classifier on the 1,046 LLM-labeled student questions. FastFit uses batch contrastive learning over sentence embeddings to achieve high accuracy with limited per-class examples, a critical property given that some topics have as few as one training sample. The classifier assigns each question to one of the 54 topic classes or an “unknown” label when confidence is insufficient.

Backbone and preprocessing. We use `all-mpnet-base-v2` [15] as the frozen sentence-embedding backbone. Each raw Caliper event is parsed to extract the student-authored question string; system-generated prompts and duplicate questions within a single session are removed prior to classification.

Training configuration. FastFit is fine-tuned with a batch size of 32, a learning rate of 1×10^{-5} , 40 epochs, and 5 training repeats per class. The token-level similarity head is enabled; the sentence-transformer backbone weights are held frozen. Training runs in under 20 minutes on a single GPU.

Class coverage. Of the 54 topic nodes in the knowledge graph, 42 receive at least one LLM-assigned training label; together with an “unknown” abstention class for off-topic questions, this gives an effective classifier output space of 43 labels. The remaining 12 topics have no training questions and are not predicted. Per-class label counts among the 42 substantive topics are highly skewed (median = 5, max = 89, min = 1), reflecting the long-tailed nature of student interest; the abstention class itself accumulates 464 examples. To counteract this imbalance during training we downsample the abstention class to 80 examples and augment each topic with fewer than 10 examples by three short paraphrases derived from its KG description, yielding a balanced pool of 737 training items.

3.3. Survey Validation and H2 Test

For H2, we compare topic-level assistant question volume (total questions per topic across all students) against self-reported difficulty from an independent mid-semester Qualtrics survey ($n = 280$ respondents). Both are aggregate class-wide signals, making raw question volume the appropriate assistant-side counterpart to the aggregate survey measure. We rank both by topic and compute Spearman’s rank correlation ρ over the 28 topics present in both datasets.

Matching survey topics to graph topics. The mid-semester survey item asks respondents to rank

which topics they found most difficult from a fixed list authored by the course instructor in collaboration with the teaching team, prior to and independently of the knowledge-graph extraction. We construct the matched set of 28 topics by manually aligning each survey topic to its corresponding node in the 54-topic graph. Matching is straightforward where the survey topic name and graph topic name agree verbatim (e.g., *Analogical Reasoning*, *Planning*); where wording differs but the underlying concept is the same (e.g., the survey’s *Generate & Test* maps to the graph’s *Generate-and-Test*), we record the alignment in a matching table that is shipped with the analysis code. Survey topics with no clear graph counterpart (8 topics, mostly late-semester project or administrative items) and graph topics with no corresponding survey entry (18 topics, mostly low-traffic specialized concepts) are excluded from the comparison. The final matched set of 28 topics is fixed before the correlation is computed.

4. Evaluation

4.1. Dataset

We analyze **1,340 question events** from **164 students** in a graduate-level AI course offered online by a US R1 university. A *question event* is a single Caliper MessageEvent record from a conversational AI teaching assistant in which the student actor authored a non-empty question; system-generated prompts, internal acknowledgements, and non-question utterances are excluded. The 1,340 raw question events are deduplicated and filtered to 1,046 unique student-authored questions, which form the pool used for downstream LLM annotation and classifier training. Caliper events are ingested through the A4L data architecture [17, 18] and processed by the configurable analytics pipeline of [19], which provides the per-cohort event extraction and deduplication described below. The mid-semester survey has **280 respondents** who self-reported which topics they found most difficult.

Data handling. All analyses were conducted on pseudonymized logs under an institutionally approved research protocol. Direct identifiers (name, email, student ID) were removed at ingestion and replaced with opaque per-student hashes before any processing. No raw question text is released with this paper; examples quoted in the error analysis are paraphrased to remove content that could identify individual students or cohorts.

4.2. H1: Classification Accuracy

The classifier was evaluated on an initial human-annotated gold set of 70 questions, labeled by a course teaching assistant using the course textbook as ground truth. The set is small relative to the 43 target classes; expanding it is a priority for ongoing work.

Table 1

Classifier evaluation on the 70-question gold set (43-class label space). FastFit is our pipeline; GPT-4 direct is the same closed-label prompt used to generate the training set, applied at inference time as a baseline.

Method	Accuracy	Macro F1
GPT-4 direct (closed-label prompt)	0.786	0.654
FastFit (ours)	0.800	0.669
FastFit weighted F1: 0.856.		

Error analysis. The 14 misclassified gold items cluster into three interpretable patterns: (i) *Surface-keyword traps* (7/14), where a strongly cued word pulls the prediction toward an unrelated topic (e.g., a question about Schank’s primitive actions classified as *Classification* rather than *Commonsense Reasoning*); (ii) *adjacent reasoning topics* (5/14), mostly gold-label *Analogical Reasoning* questions classified into linked problem-solving topics (*Problem Reduction*, *Knowledge Representation*) whose textbook vocabulary (“similarity”, “mapping”, “structural”) overlaps; and (iii) *vague questions and abstention misfires* (2/14), where the question is too fragmentary to disambiguate or the abstention class fires

inappropriately. Pattern (i) reflects a classifier limitation that scaling the training pool would address; pattern (ii) reflects topic boundaries that a multi-label extension could resolve. The gap between macro F1 (0.669) and weighted F1 (0.856) is driven by the long tail of singleton-support classes in the gold set, the expected behavior of a few-shot classifier on a heavily imbalanced class set.

Per-topic performance. The unknown abstention class dominates support (22/70 gold items) and is well classified ($F1 = 0.93$), confirming that the model reliably filters off-topic chat (course logistics, trolling, out-of-curriculum questions). Among substantive topics, the higher-support classes (Analogical Reasoning at $F1 = 0.55$ and Commonsense Reasoning at $F1 = 0.67$) are where most misclassifications concentrate, since these topics share vocabulary with adjacent reasoning topics in the curriculum and account for most of the macro-F1 gap.

H1 is supported: 80% accuracy appears sufficient for preliminary aggregate topic-level analysis in this setting, and the error structure is concentrated in predictable, interpretable failure modes.

4.3. H2: Survey Alignment

Table 2

Survey validation: assistant question volume vs. self-reported difficulty.

Metric	Value
Spearman ρ	0.491
p -value	0.008
Matched topics	28
Significant ($\alpha = 0.05$)	Yes

Topic-level assistant question volume tracks self-reported difficulty significantly across 28 matched topics. Analogical Reasoning generates the most assistant questions (89 events) and is the fifth most-reported survey difficulty topic (19.3% of respondents). Generalization (33 events) and Constraint Propagation (28 events) also rank highly in both measures.

Per-topic rank agreement. Grouping the 28 matched topics by absolute rank difference $|\Delta|$ between assistant volume rank and survey difficulty rank: 14 of 28 fall in strong agreement ($|\Delta| \leq 3$), 9 in moderate disagreement ($4 \leq |\Delta| \leq 10$), and 5 are outliers ($|\Delta| > 10$). The outliers split into two interpretable subpatterns: late-semester meta-level topics that students rate hard but rarely query, and concept-introduction topics that generate terminology questions out of proportion to their conceptual difficulty.

H2 is supported: assistant question volume is a statistically significant surrogate for aggregate topic difficulty.

5. Discussion and Limitations

Practical implication. The pipeline converts passively logged assistant interactions into topic-level difficulty signals without additional student burden. Instructors are shown which topics attract the most questions and can verify the ranking against independent survey evidence.

Relevance to teachers and the AIED setting. The student-side input is naturally occurring chat data, but the evaluation is grounded in instructor-authored curriculum artifacts (the prerequisite knowledge graph and the gold set): the model is trained on student behavior but assessed against curriculum representations the instructor signed off on, making the resulting signal actionable in an AIED workflow without requiring an item bank or a Bayesian skill model. The primary use we envision is a cohort-wide instructor dashboard; the same signal can be passed to the conversational assistant as *suggestions* for prerequisite review when a student asks about a downstream topic with an unresolved prerequisite, or surfaced through the LMS as student-specific recommendations. A related line of work in our group surfaces isolated learners through outcome-independent mediation of feedback between

teachers and students using AI [20], framing the teacher–AI loop as the unit of analysis rather than the student–AI dyad alone.

Why topic classification rather than per-question solving. Per-question correctness assessment in free-form student questions is a multi-hop reasoning problem that current LLMs handle unreliably. Aggregating classified questions over a student’s history sidesteps this: the diagnostic signal is the *distribution of help-seeking across topics* rather than the correctness of any individual question, which is empirically more robust and matches what an instructor can realistically act on.

Limitations. Three caveats bound the present claims. (i) The 20% classification error rate introduces label noise into the topic-volume counts, concentrated in topics with overlapping vocabulary (e.g., Analogical Reasoning vs. Knowledge Representation); this likely attenuates the measured correlation. (ii) Evaluation is on one graduate-level AI course at a US R1 university; replication across courses and disciplines is needed. (iii) The H2 result is aggregate at the topic level and does not by itself identify which individual students are struggling on which topic; we return to this point in the Future directions paragraph below.

Future directions. The topic-level volume signal reported here points naturally toward three complementary extensions on the same data. First, the *per-student gap signal*: students whose question distribution over topics diverges anomalously from the cohort can be flagged for targeted support, in contrast to the aggregate ranking this paper reports. Second, *prerequisite-edge gap propagation*: the directed edges in the knowledge graph become diagnostically useful once per-student gaps are available, because an observed gap on a downstream topic can be traced back along its prerequisite chain to candidate upstream root-cause topics that the student has not consolidated. This is the analytical step the graph edges are designed for; we are preparing a companion analysis on the same cohort. Third, a closed-loop *intervention study* on the resulting instructor-facing dashboard: surfacing per-student gap chains alongside classifier confidence would allow a controlled test of whether curriculum-grounded remediation prompts measurably improve outcomes relative to a generic-help control.

6. Conclusion

We presented a work-in-progress pipeline that maps conversational AI student questions onto a GPT-4-extracted prerequisite knowledge graph via a few-shot classifier, and tested whether such logs surface topic-level knowledge gaps matching an independent difficulty measure. Both hypotheses are supported: 80.0% classifier accuracy on a 43-class few-shot task (H1), and significant correlation between assistant question volume and self-reported difficulty (H2: $\rho = 0.491$, $p = 0.008$). At the *course level*, the topics that draw the most assistant questions are, with statistical reliability, the same topics students report as most difficult—evidence that conversational AI interaction logs, mapped onto curriculum structure, surface course-level knowledge gaps without requiring an item bank or per-question correctness assessment.

Acknowledgments

We thank members of the A4L team in Georgia Tech’s Design Intelligence Lab (DILab; dilab.gatech.edu) for their contributions to this research. This research is supported by a US National Science Foundation grant (Grant No. 2247790) to the National AI Institute for Adult Learning and Online Education (AI-ALOE; aialoe.org).

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude in order to: Draft content, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication’s content.

References

- [1] A. K. Goel and L. Polepeddi, Jill Watson: A virtual teaching assistant for online education, in: *Learning Engineering for Online Education*, Routledge, 2018, pp. 120–143.
- [2] P. Maiti, A. K. Goel, Can an AI partner empower learners to ask critical questions?, in: *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*, 2025, pp. 314–324.
- [3] Q. Fu, Q. Wu, D. Williams, Mining the gold: Student-AI chat logs as rich sources for automated knowledge gap detection, arXiv preprint arXiv:2512.22404, 2025.
- [4] L. Cabral, R. Pinto, G. Gonçalves, AI-powered learning analytics dashboards: A systematic review of applications, techniques, and research gaps, *Discover Education*, 2025.
- [5] A. T. Corbett, J. R. Anderson, Knowledge tracing: Modeling the acquisition of procedural knowledge, *User Modeling and User-Adapted Interaction* 4(4) (1994) 253–278.
- [6] C. Piech, et al., Deep knowledge tracing, in: *Advances in Neural Information Processing Systems*, 2015.
- [7] A. Yehudai, E. Bandel, et al., FastFit: Fast and effective few-shot text classification with a multitude of classes, in: *Proceedings of NAACL 2024 (System Demonstrations)*, 2024.
- [8] P. Chen, Y. Lu, V. W. Zheng, Y. Pian, Prerequisite-driven deep knowledge tracing, in: *IEEE International Conference on Data Mining (ICDM)*, 2018, pp. 39–48.
- [9] L. Pan, C. Li, J. Li, J. Tang, Prerequisite relation learning for concepts in MOOCs, in: *Proceedings of ACL*, 2017.
- [10] C. Liang, J. Ye, S. Wang, B. Pursel, C. L. Giles, Investigating active learning for concept prerequisite learning, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [11] I. Roll, V. Aleven, B. M. McLaren, K. R. Koedinger, Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system, *Learning and Instruction* 21(2) (2011) 267–280.
- [12] J. Park, Y. Medhat, H. P. Wai, P. Thajchayapong, A. K. Goel, Evaluating learner representations for differentiation prior to instructional outcomes, arXiv preprint arXiv:2604.05848, 2026.
- [13] J. Park, Y. Medhat, H. P. Wai, P. Thajchayapong, A. K. Goel, Memory-based vs. context-only conditioning produces distinct behavioral patterns in stateful personalization, arXiv preprint arXiv:2605.27389, 2026.
- [14] S. Basu, J. Brown, C. Lum, J. Park, A. K. Goel, Bidirectional feedback-based personalization of learning using multi-tier AI: A real-world assessment of its efficacy in classrooms, in: *Proceedings of the AAAI Symposium Series*, volume 5, 2025, pp. 50–51.
- [15] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: *Proceedings of EMNLP*, 2019.
- [16] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, O. Pereg, Efficient few-shot learning without prompts, arXiv preprint arXiv:2209.11055, 2022.
- [17] A. K. Goel, P. Thajchayapong, V. Nandan, H. Sikka, S. Rugaber, A4L: An architecture for AI-augmented learning, in: *AI Applications in Online Higher Education Administration*, Routledge, 2025, pp. 151–164.
- [18] P. Thajchayapong, S. Carbonaro, T. Couper, B. Helmick, S. Rugaber, A. K. Goel, Evolution of A4L: A data architecture for AI-augmented learning, arXiv preprint arXiv:2511.11877, 2025.
- [19] J. A. Santana, P. Thajchayapong, S. Rugaber, A. K. Goel, A highly configurable data analytics pipeline for learning, in: *Proceedings of EDULEARN25*, 2025, p. 8737.
- [20] J. Park, Y. Medhat, H. P. Wai, P. Thajchayapong, A. K. Goel, Surfacing isolated learners with outcome-independent mediation of feedback between teachers and students using AI, arXiv preprint arXiv:2605.29240, 2026.