

SAGE: A Synthetic Graph Framework for Personalized Student Knowledge-State Modeling

Manzi Fabrice Niyigaba
Dartmouth College
fabrice.27@dartmouth.edu

Thomas Thesen
Dartmouth College
thomas.thesen@dartmouth.edu

Xuyuan Liu
Dartmouth College
xuyuan.liu.gr@dartmouth.edu

Yujun Yan
Dartmouth College
yujun.yan@dartmouth.edu

Abstract

Assessing a student’s knowledge state accurately and efficiently remains a fundamental yet unsolved challenge in intelligent tutoring systems. Existing approaches treat assessment as a static prediction task over pre-defined test sets and provide limited support for adaptive, interaction-efficient evaluation during learning. Although recent advances in LLM-based simulation have enabled the modeling of educational content and learner behavior, standardized testbeds for systematically studying interactive assessment remain scarce. In this work, we formalize personalized interactive knowledge assessment under explicit interaction budgets and introduce SAGE, a novel framework for Student Assessment via Graph Exploration. SAGE constructs knowledge graphs from real educational materials across multiple domains, defines overlapping local knowledge regions for context-aware question generation, and simulates learner interaction trajectories. Evaluating various methods on SAGE datasets, we find that techniques leveraging graph structure and interaction history consistently outperform baselines, while performance remains highly sensitive to individual learner characteristics—an important open challenge. We release SAGE to accelerate the development of interaction-efficient methods for precise and personalized assessment in education. Our data and code are available at [Link](#).

CCS Concepts

• Information systems → Data mining; • Applied computing → Computer-assisted instruction.

Keywords

AI for Education, Personalization, Intelligent Learning Systems

ACM Reference Format:

Manzi Fabrice Niyigaba, Xuyuan Liu, Thomas Thesen, and Yujun Yan. 2026. SAGE: A Synthetic Graph Framework for Personalized Student Knowledge-State Modeling. In *Proceedings of ACM SIGKDD Conference on Knowledge*

Discovery and Data Mining (KDD '26). ACM, New York, NY, USA, 9 pages.
<https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Assessment shapes how students learn and how instructors teach. For students preparing for high-stakes exams, self-assessment builds confidence in the material and clarifies exactly where to focus next. For instructors, it uncovers real understanding of students, exposes hidden misconceptions, and shapes what to teach next. In practice, effective assessment is inherently interactive and adaptive [9]: instructors continually revise questions based on a student’s past performance, targeting uncertain or weakly understood concepts [27]. However, this adaptivity is constrained by the cost of interaction: each additional question demands student effort and instructional time. Therefore, a truly effective system must infer a learner’s full knowledge state from only a handful of interactions by asking the most informative questions possible. Despite these requirements, decades of research in knowledge tracing have largely treated assessment as a static prediction problem, building increasingly large next-response predictors [22, 38] on fixed logs of past student responses [31]. These methods treat learner responses primarily as observations for prediction, rather than as dynamic feedback for selecting subsequent questions. Consequently, they provide limited support for real-time, personalized assessment under tight interaction budgets.

Developing interaction-efficient and personalized assessment models has long been constrained by a fundamental data bottleneck. Real-world assessment logs are often limited in subject coverage, learner diversity, and, most critically, the adaptive interactions needed to study personalized assessment. Collecting such data at scale is also difficult: without strong adaptive models, it is hard to generate high-quality adaptive interactions, yet such models themselves require adaptive interaction data for development and evaluation. This creates a persistent chicken-and-egg problem. Recent advances in LLM-based student simulation [34] offer a promising way to generate realistic synthetic interaction data at scale across diverse learner profiles. Building on this, we break the deadlock by introducing Student Assessment via Graph Exploration (SAGE), a novel framework for interactive student knowledge assessment under explicit interaction budgets. SAGE provides (i) diverse benchmark datasets, (ii) a controlled simulation environment for learner–assessor interactions, and (iii) strong baseline methods that establish reproducible performance and enable systematic comparison of interaction-efficient strategies.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '26, Jeju Island, South Korea

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
<https://doi.org/XXXXXXX.XXXXXXX>

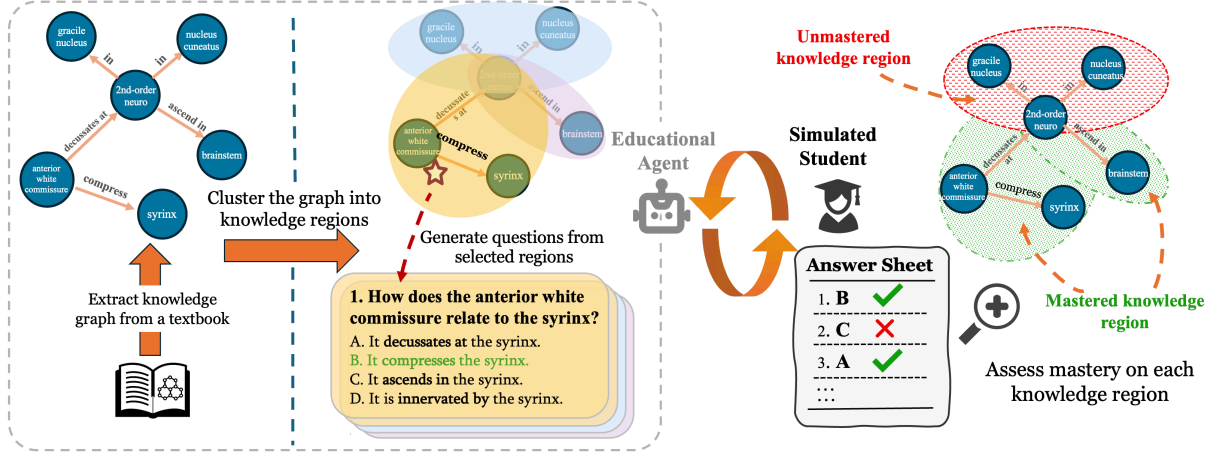


Figure 1: Overview of SAGE framework. It extracts a knowledge graph from a textbook, clusters it into coherent knowledge regions, and generates questions by conditioning an LLM on each region. Simulated students with diverse profiles answer these questions, and their responses are used to assess mastery across knowledge regions.

Under a limited interaction budget, what does a student know? To study this rigorously, we need a structured representation of the source material, questions grounded in that structure, and learners whose underlying knowledge can be verified. SAGE provides all three: concept graphs extracted from real educational materials, questions generated over local graph regions, and simulated students whose responses follow predefined knowledge profiles. The resulting question–answer sequences form the SAGE dataset. This setup supports methods that use interaction histories to select subsequent questions and update beliefs about a learner’s latent knowledge state from observed responses. To our knowledge, SAGE is the first framework to formalize student knowledge assessment as a graph-structured sequential decision problem under explicit interaction budgets, reflecting the practical constraints of adaptive assessment in educational settings. Taken together, our main contributions are:

- **New problem.** We formalize interactive student knowledge assessment as a graph-structured sequential decision problem under explicit query budgets.
- **New framework.** We introduce SAGE, a domain-agnostic framework built from real STEM textbooks, providing diverse concept graphs, overlapping knowledge regions, and simulated student interaction trajectories for standardized evaluation.
- **Practical baselines.** We provide baseline methods and show that leveraging graph structure and interaction history yields strong gains, while exposing an open challenge: sensitivity to learner profiles and the underlying graph topology.

2 Related Work

LLMs for Educational Content and Student Simulation. LLMs can extract knowledge graphs from raw text at scale, with methods spanning multi-turn dialogue, code generation, and agent-based workflows [33, 37]. More recent extensions tackle billion-node schemas, domain-specific extraction, and sparse-data settings [1, 3, 19]. In parallel, LLMs have shown surprising fidelity

in simulating human behavior, replicating social routines [21], diverse personalities [16], and real experimental outcomes [5]. This positions them as scalable human simulators [35, 39]. These capabilities have inspired adoption in educational settings, including personalized assessment, exercise retrieval, tutoring, student profile simulators, and multi-agent teaching [15, 17]. SAGE builds on these advances.

Knowledge Tracing and Student Modeling. Knowledge tracing methods aim to estimate a learner’s mastery from historical question responses [6]. Early neural methods used RNNs and memory networks to model temporal learning patterns [22], while later work incorporated explicit knowledge graphs through GNN-based propagation [20]. More recent LLM-based approaches enrich estimates of student knowledge with model-generated reasoning traces [25], leveraging LLMs’ ability to approximate student-like reasoning. However, these methods are largely trained and evaluated on fixed historical logs to predict the next response in a pre-recorded sequence. Since the logs are static, the question order is fixed in advance, and the model cannot decide what to ask next. SAGE departs from this paradigm by treating assessment as an interactive process, where each observed response informs the next question, allowing the system to recover the learner’s full knowledge state with far fewer queries.

3 Problem Setup

We study the problem of estimating a student’s knowledge state over a given text corpus in a sample-efficient manner. Unlike traditional approaches that rely on static assessments, our goal is to leverage both the corpus’s implicit knowledge structure and the student’s interaction trajectory to probe their level of concept mastery. Specifically, we first perform entity and relation extraction to construct a knowledge graph $G = (\mathcal{V}, \mathcal{E})$ from the corpus. Atomic knowledge units are represented as triplets in $\mathcal{T} = \{(h, r, t)\}$, where $h, t \in \mathcal{V}$ denote head and tail knowledge entities, and $r \in \mathcal{E}$ denotes the semantic relation between them. This graph provides an

explicit representation of concepts and their dependencies, transforming unstructured text into a structured schema. Then, to model local conceptual structure, for each anchor concept s , we define a localized knowledge region $c(s) \subseteq \mathcal{V}$. This region consists of a cluster of nodes related to s and is used to generate test questions associated with that concept. We also explicitly model student interactions by jointly modeling the posed question and the student’s answer. Specifically, student interactions are represented as a trajectory $\mathcal{I} = \{\tau_1, \dots, \tau_T\}$, where each interaction is defined as $\tau_t = (q_t, a_t^*, a_t)$. Here, q_t denotes the question, a_t^* denotes the ground truth answer, and a_t denotes the student’s answer. Our objective is to accurately infer the student’s latent mastery over the text corpus based on the interaction trajectory $\mathcal{I}$ using a minimal number of questions.

4 Framework Construction

To instantiate the problem setup defined in Section 3, we introduce the pipeline shown in Figure 1, which couples an educational agent with simulated learners. The agent’s workflow can be decomposed into 3 sub-tasks: (1) extracting a knowledge graph from a textbook, (2) partitioning it into localized knowledge regions, and (3) generating region-conditioned questions, which simulated learners then answer according to predefined mastery profiles. The resulting question–answer trajectories form the SAGE datasets. We detail each below.

4.1 Knowledge Graph Construction

We start by transforming an unstructured textbook into a structured representation of its concepts and their relationships, which is then used to reason about what a student knows. To achieve this, we adopt the LLM-based knowledge graph construction framework [19] to extract a knowledge graph from textbook corpora, where nodes represent knowledge entities and edges capture the semantic relationships between them. The resulting graphs are naturally aligned with the pedagogical structure of the source material, offering a hierarchical representation of concepts and their interrelations. Concretely, we partition the textbook into chunks and merge the resulting per-chunk knowledge graphs into a single global knowledge graph G_i summarizing the full textbook. To treat each factual triplet (h, r, t) as a single atomic knowledge unit for assessment, we convert G_i into its line graph $L(G_i)$ [11], where each node corresponds to one full triplet and thus preserves its head–relation–tail structure. We then embed each triplet using SentenceBERT [23], producing semantic node features $\mathbf{x}_u$ that capture the semantics of each knowledge unit and drive the clustering step described next. As shown in Table 1, our generated knowledge graphs cover different subjects and vary in scale and density. We report further details on the source textbooks used in the Appendix A.

Table 1: Statistics of the constructed knowledge graphs. The graphs span diverse subjects and vary in scale and density.

	ECON	CS	Pharm	Physio
No. of nodes	5,708	7,281	31,736	34,241
No. of edges	35,571	39,665	117,135	420,008
Avg. degree	12.46	10.89	7.38	24.53

4.2 Knowledge Graph Clustering

Each coherent knowledge concept presented in a textbook is supported by a small set of tightly related triplets, so we partition $L(G_i)$ into localized regions, each capturing one such concept to ground the assessment questions. In this view, a strict single-membership partition, however, is poorly suited here: foundational concepts naturally recur across multiple topics in educational content, and confining a triplet to one region would sever cross-topic links that are pedagogically essential. To overcome this, we utilize an overlapping clustering approach that allows shared triplets across multiple regions.

Specifically, we adopt Graph Convolutional Clustering (GCC) [8], which iteratively updates node representations using the graph neighborhood information and refines cluster assignments based on representational similarity. However, since the GCC objective does not explicitly encourage structural coherence, we augment it with an additional Newman–Girvan modularity regularizer [10] to bias clustering toward dense, well-connected regions of the concept graph, ensuring that each cluster corresponds to a coherent conceptual area. We refer to this formulation as **GCC_{hard}**. We report how we select the modularity weight λ in Appendix C.1, where we observe that our modification of explicitly augmenting GCC with a modularity regularizer consistently improves clustering quality across all datasets.

Since **GCC_{hard}** naturally enforces single-membership assignments, we additionally consider three overlapping strategies to better capture the inherent overlap in domain knowledge: **GCC_{soft}** duplicates the top $p\%$ of nodes with the highest assignment uncertainty, measured as the ratio of the distance to the nearest cluster centroid over the distance to the second nearest, allowing ambiguous concepts to participate in multiple clusters. Additionally, we introduce ego-based overlapping strategies; **Ego** is based on ego-splitting [7], which creates multiple personas for boundary nodes by splitting each boundary node into distinct instances, one per compatible cluster it belongs to. We define a boundary node u as a node with neighbors in more than one cluster and whose embedding lies close to the decision boundary between two cluster regions in the embedding space. Formally, u is a boundary node if $|\{c^*(v) : v \in \mathcal{N}(u)\}| \geq 2$ and $\|\mathbf{x}_u - \boldsymbol{\mu}_{c_1}\|_2 - \|\mathbf{x}_u - \boldsymbol{\mu}_{c_2}\|_2 \leq \delta$, where $\mathcal{N}(u)$ denotes the neighbors of u , $c_1 = c(u)$ is the primary cluster of u , c_2 is the nearest alternative cluster, and $\delta \geq 0$ controls how close to the decision boundary a node must be to qualify for duplication. Among all boundary nodes, we retain the top $p\%$ with the smallest boundary scores and apply ego-splitting to each, assigning it to both c_1 and all semantically compatible secondary clusters. Finally, **Ego_{mod}** further restricts ego-splitting by requiring that splitting a boundary node improves local modularity. All boundary nodes are evaluated simultaneously against the original cluster structure, and a split is accepted only if the normalized modularity gain exceeds a threshold ε : $\frac{Q_{\text{split}} - Q_{\text{base}}}{\deg(u)} \geq \varepsilon$. This constraint ensures that overlap is introduced only if it satisfies both semantic proximity and strengthens local structural cohesion.

Metrics. We compare **GCC_{soft}**, **Ego**, and **Ego_{mod}** soft clustering strategies along four complementary metrics that span structural quality and downstream utility for question generation. First, we consider **Overlapping modularity** Q_{ov} [4] that measures internal

cohesion adapted to multi-membership assignments. Second, we adopt **Conductance** that measures boundary separation, where lower values indicate well-separated clusters. Third, we use **Coverage** that measures how much of the graph’s connectivity is captured by the clustering. Since the ultimate goal of clustering is to support question generation, we additionally introduce **Duplication Utility** (DU), which measures how often duplicated concepts are referenced in questions from their secondary clusters. Formal definitions are provided in Appendix C.2. To summarize performance across all four dimensions in a single number, we report a **Combined** score given by the geometric mean of the metrics, defined as $\text{Comb} = (Q_{\text{ov}} \cdot (1 - \text{Cond}) \cdot \text{Cov} \cdot \text{DU})^{1/4}$, where Conductance is inverted to $(1 - \text{Cond})$ so that all four quantities follow the convention that higher is better. The geometric mean penalizes uneven scores, rewarding only strategies that are simultaneously structurally coherent and pedagogically useful.

Table 2: Clustering quality and duplication utility at duplication rate = 0.10. Ego methods outperform GCCsoft on duplication utility and combined score across all three datasets.

	Method	$Q_{\text{ov}} \uparrow$	Cond. $\downarrow$	Cov. $\uparrow$	DU $\uparrow$	Comb $\uparrow$
ECON	GCCsoft	0.70	0.18	0.82	0.07	0.43
	Ego	0.66	0.25	0.83	0.16	0.51
	Egomod	0.65	0.26	0.84	0.15	0.50
CS	GCCsoft	0.76	0.15	0.87	0.08	0.46
	Ego	0.68	0.21	0.89	0.20	0.56
	Egomod	0.69	0.22	0.89	0.15	0.52
Pharm	GCCsoft	0.81	0.13	0.87	0.08	0.47
	Ego	0.68	0.20	0.91	0.16	0.53
	Egomod	0.68	0.20	0.91	0.16	0.53

As shown in Table 2, GCC_{soft} achieves the highest scores on standard clustering metrics such as Q_{ov} and conductance. However, it underperforms on Duplication Utility. This shows that its duplicated nodes are semantically redundant in their secondary clusters and are largely ignored during question generation. In contrast, Ego and Ego_{mod} duplicate structurally active boundary nodes that bridge knowledge regions, producing duplications that are effectively utilized, as reflected in their higher DU scores. This suggests that standard clustering metrics alone might not be sufficient for evaluating soft clustering in interactive assessment. Based on this result, we adopt Ego as the default clustering strategy for SAGE, as it achieves comparable quality to Ego_{mod} at substantially lower computational cost.

4.3 Question Generation

Given the clustered concept regions identified in Section 4.2, the next task for the educational agent is to generate assessment questions that evaluate the learner’s mastery of these regions. Recent work has shown that LLMs can produce high-quality multiple-choice questions (MCQs) that are often indistinguishable from human-written ones, even when judged by domain experts [24]. Building on these findings, we leverage LLMs to generate MCQs conditioned on the concept clusters. We design the prompt following proven strategies that achieve human-comparable question quality, including instructor role-playing [29], few-shot prompting with textbook MCQs from the same source [28], and the CTQA [32] strategy. We instruct the model to match the style, difficulty, and domain ontology of expert-written assessments. As an additional consistency check, each generated question is reviewed by a checker

LLM from a different model family. The checker verifies that the concepts reported by the educational agent are actually reflected in the question. .

Evaluation. The standard way to evaluate LLM-generated content is to ask another LLM to score it. However, recent work has shown this signal to be unreliable. The single token a judge outputs as its score can systematically diverge from the underlying distribution that produced it [30]. Therefore, rather than reading the judge’s output token directly, we prompt it to rate each question on a 0–5 scale along five dimensions, Naturalness, Clarity, Answerability, Difficulty, and Distractor Quality, and extract the log-probabilities it assigns to score tokens 0 through 5, recovering a calibrated signal from the full distribution. To compare against expert human-written questions, we measure the Wasserstein distance between SAGE’s score distributions and those of held-out human-written questions excluded from few-shot prompting, repeating the evaluation across judge models from different families and architectures for robustness.

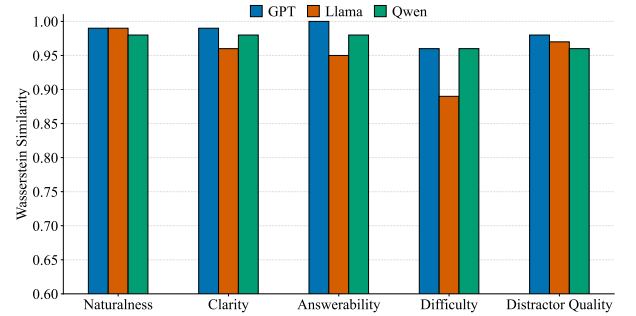


Figure 2: Wasserstein similarity between SAGE-generated and human-written questions across five quality dimensions. Across all dimensions and evaluator LLMs, SAGE-generated questions closely match human-written ones.

As shown in Figure 2, our method produces questions whose score distributions closely match those of human-written questions across all judges and evaluation dimensions. We also noted that explicitly instructing the model to emulate the language style, domain ontology, distractor quality, and difficulty level of the few-shot examples significantly narrows the distribution gap between LLM-generated and human-written questions.

5 Generative Students

Prior studies show that LLMs can simulate student responses that closely align with human learners across diverse profiles and difficulty levels [26, 34]. We adopt the pipeline of [18], which has demonstrated strong alignment with real student behavior, and define two mastery states per knowledge region: a student is *mastered* when they fully understand the relevant concept cluster, and *unmastered* otherwise. For each question, the model role-plays a student at the assigned mastery level and produces a multiple-choice selection together with a reasoning trace grounded in the question’s concept terms.

5.1 Response Scoring

Given a question q_t generated from a subset of triplet nodes $S_t \subseteq c_t$ in cluster c_t , the student model produces a response $(a_t, \mathcal{X}_t)$, where a_t is the selected answer and $\mathcal{X}_t$ is the accompanying reasoning trace. We convert this into node-level signals over the queried cluster c_t . For directly tested nodes $v \in S_t$, the answer signal is $a_t(v) = +1$ if $a_t = a_t^*$ and -1 otherwise. An LLM judge analyzes $\mathcal{X}_t$ and classifies each node $v \in c_t$ as mastered, confused, or unknown, yielding a reasoning signal $r_t(v) \in \{+1, -1, 0\}$ defined over all nodes in the cluster. For each node $v \in c_t$, the combined evidence, clipped to $[-1, 1]$, is $e_t(v) = w_a a_t(v) + w_r r_t(v)$ if $v \in S_t$ and $e_t(v) = w_r r_t(v)$ if $v \in c_t \setminus S_t$, where w_a and w_r are the answer and reasoning signal weights respectively. Nodes in S_t receive both the answer signal and the reasoning signal, while nodes in $c_t \setminus S_t$ receive only the reasoning signal.

5.2 Mastery Belief Representation

The per-node evidence $e_t(v)$ produced in Section 5.1 is accumulated into a Beta-Bernoulli belief at each node, yielding a posterior estimate of mastery that updates with every interaction. Each concept node $v \in \mathcal{V}$ maintains a Beta-Bernoulli belief initialized at $(\alpha_v, \beta_v) = (1, 1)$, representing maximum uncertainty. We track two derived quantities per node: $k_v = \frac{\alpha_v - \beta_v}{\alpha_v + \beta_v} \in [-1, 1]$ and $u_v = \frac{\sqrt{\alpha_v \beta_v} / ((\alpha_v + \beta_v)^2 (\alpha_v + \beta_v + 1))}{\sqrt{1/12}} \in [0, 1]$, where k_v is the signed knowledge estimate and u_v is the normalized uncertainty. After each interaction, queried nodes are updated directly: $\alpha_v \leftarrow \alpha_v + \max(0, e_t(v))$ and $\beta_v \leftarrow \beta_v + \max(0, -e_t(v))$. Positive evidence increments α_v , negative evidence increments β_v , and zero evidence leaves both unchanged. Mastery is predicted by thresholding the signed knowledge: $\hat{y}_v = 1[k_v > 0]$.

6 Baseline Methods

SAGE supports two key types of methodological development: (1) **belief propagation**, deciding how to update mastery beliefs from observed responses and extend them across the concept graph to regions not yet queried, and (2) **question selection policy**, deciding which knowledge region to query next. We next present several baselines for each category.

6.1 Mastery-Level Belief Propagation

Without structural guidance, belief updates are confined to directly queried concepts, leaving unqueried concepts with no evidence. We address this by propagating the signed knowledge vector $\mathbf{k}_t = (k_v)_{v \in \mathcal{V}}$ across the graph after each interaction. The propagated signal is: $\sigma_t = \Pi(\mathbf{k}_t; G)$ where Π is a diffusion operator. Unqueried nodes then receive a soft belief update proportional to their graph proximity to the queried set S_t . Letting $\omega_v = 1/(d(v, S_t) + 1)$ denote the inverse-distance weight and λ the propagation strength: $\alpha_v \leftarrow \max(1, \alpha_v + \lambda \max(0, \sigma_t(v)) \omega_v)$ and $\beta_v \leftarrow \max(1, \beta_v + \lambda \max(0, -\sigma_t(v)) \omega_v)$. Positive propagated signal adds evidence for mastery, negative signal adds evidence against mastery, and nodes far from the queried set receive weaker updates. We consider three propagation operators widely used in the graph

learning literature, each with a distinct inductive bias over how beliefs spread across the graph. **PPR/APPNP** [2, 14] applies anchored local diffusion. **Heat Kernel** [36] applies smooth isotropic diffusion. **Chebyshev** [12] applies finite-hop polynomial diffusion. Furthermore, we consider a **Distance-Weighted** baseline that applies no diffusion operator, setting $\sigma_t = \mathbf{k}_t$ and updating unqueried nodes solely through the inverse-distance weight $\omega_v = 1/(d(v, S_t) + 1)$, without spreading beliefs across the graph.

6.2 Question Selection Policy

Since each question probes only a localized set of concepts, the selection policy governs which parts of the knowledge graph are evaluated, and thus how efficiently a learner's mastery state can be inferred. Under a limited query budget, an effective selector should target the clusters whose mastery is most uncertain, as querying well-understood ones yields little information. At each step, we score every candidate cluster by the total uncertainty it covers and select the highest. All selectors share a base score summing the uncertainty of the cluster and its neighborhood, excluding nodes already covered by previous queries: $\text{Base}(S_t) = \sum_{v \in S_t} u_v + \lambda_{\text{cov}} \sum_{v \in R_h(S_t) \setminus C_t} u_v$, where u_v is the uncertainty at node v , $R_h(S_t)$ the h -hop neighborhood of the cluster, and C_t the set of already-covered nodes. **Submodular** uses this base score directly, picking the cluster with the most total uncertainty.

Uncertainty alone may be insufficient, so we ask whether structural information, such as which clusters lie at mastery boundaries, improves selection. We consider three variants, each adding a structural preference to the base score. **Boundary** targets clusters with mixed mastery, adding $\mu \sum_{(u,v) \in E, u,v \in R_h(S_t)} \mathbf{1}\{k_u k_v < 0\}$, which counts neighborhood edges with one endpoint mastered and the other not. **CutMass** favors well-connected clusters, whose answers propagate widely, adding $\mu_{\text{cut}} |\{(u,v) \in E \mid u \in R_h(S_t), v \notin R_h(S_t)\}|$, the edges leaving the neighborhood. **Frontier** encourages outward exploration, prioritizing clusters at the edge of the covered set by adding $\mu_f \sum_{v \in R_h(S_t) \setminus C_t} u_v \mathbf{1}\{v \in \partial C_t\}$, the uncertainty at the covered-set boundary. Finally, **Random** selects clusters uniformly without replacement as a non-adaptive baseline.

7 Method Evaluation

Question selection and belief propagation interact directly: the selector determines which concepts receive direct evidence, and the propagator determines how that evidence spreads to unqueried nodes. To isolate the effect of question selection policy and belief propagation, we adopt a two-stage evaluation: we first fix question selection to random and compare propagation operators, then fix the propagator and compare question selectors.

In summary our experiments aim to answer the following research questions:

- (Q1) How does exploiting graph structure improve mastery recovery under limited interaction budgets?
- (Q2) How does the effectiveness of different diffusion operators vary across learner profiles and graph structures?
- (Q3) Which graph-structural signal most effectively guides question selection?

For fair comparison, all methods are evaluated on the same datasets, with performance averaged over three seeds.

Table 3: Question selection policies (Boundary, Cutmass, Frontier, Submodular) under three propagation operators. Macro-F1 (F_a), averaged over three seeds. Best F_a per row highlighted. Adaptive question selection consistently improves performance over random selection (Table 4).

	Budget	PPR/APPNP				Heat Kernel				Chebyshev			
		Bnd	Cut	Frn	Sub	Bnd	Cut	Frn	Sub	Bnd	Cut	Frn	Sub
20%M Pharm	50	0.564	0.474	0.561	0.567	0.488	0.451	0.489	0.490	0.604	0.482	0.616	0.607
	100	0.631	0.508	0.638	0.631	0.531	0.460	0.537	0.537	0.665	0.493	0.683	0.664
	200	0.706	0.567	0.715	0.706	0.591	0.472	0.635	0.614	0.725	0.532	0.738	0.722
	300	0.735	0.603	0.749	0.736	0.653	0.480	0.706	0.680	0.758	0.567	0.767	0.758
50%M CS	50	0.592	0.602	0.617	0.598	0.543	0.568	0.577	0.556	0.624	0.603	0.656	0.627
	100	0.711	0.581	0.723	0.705	0.654	0.557	0.662	0.656	0.710	0.613	0.730	0.701
	200	0.803	0.648	0.824	0.801	0.684	0.611	0.688	0.675	0.752	0.606	0.779	0.753
	300	0.832	0.726	0.843	0.828	0.682	0.651	0.681	0.676	0.796	0.610	0.808	0.795
90%M ECON	50	0.733	0.849	0.686	0.706	0.760	0.875	0.740	0.729	0.648	0.747	0.624	0.629
	100	0.801	0.845	0.749	0.778	0.884	0.875	0.876	0.855	0.640	0.737	0.637	0.627
	200	0.899	0.909	0.867	0.889	0.957	0.940	0.957	0.956	0.667	0.695	0.642	0.643
	300	0.932	0.931	0.909	0.928	0.960	0.957	0.962	0.961	0.671	0.672	0.652	0.660

(Q1) Exploiting graph structure yields substantial gains. As shown in Table 4, on the propagation axis alone, all three diffusion operators (PPR/APPNP, Heat Kernel, Chebyshev) consistently outperform the Distance Weighted (DW) baseline across every profile and budget. We attribute these gains to the propagation distributing evidence into regions of the graph that direct querying alone cannot reach, which is essential under tight budgets. As shown in Table 3, pairing diffusion with adaptive selectors further lifts Macro- F_a above 0.75 across all three datasets with as few as 300 queries. This demonstrates the importance of jointly leveraging graph structure across both propagation and question selection, and confirms the central design choice of SAGE.

Table 4: Mastery belief propagation under random question selection, reporting Macro-F1 (F_a) averaged over three seeds. Propagation-based methods consistently outperform the Distance-Weighted (DW) baseline. Best F_a per row highlighted.

	Budget	PPR	Heat	Cheb	DW
20%M Pharm	50	0.521	0.483	0.643	0.449
	100	0.509	0.465	0.621	0.453
	300	0.545	0.478	0.567	0.466
	500	0.555	0.480	0.547	0.472
50%M CS	50	0.538	0.466	0.575	0.343
	100	0.511	0.455	0.591	0.358
	300	0.579	0.453	0.636	0.416
	500	0.605	0.444	0.649	0.441
90%M ECON	50	0.633	0.822	0.590	0.128
	100	0.607	0.813	0.551	0.155
	300	0.592	0.742	0.545	0.234
	500	0.592	0.742	0.549	0.288

(Q2) Diffusion operator effectiveness depends on both the mastery regime and the graph topology. In Table 4, under low-mastery regimes (Pharmacology 20%M, Computer Science 50%M) the locality-preserving operators (PPR/APPNP, Chebyshev) dominate, whereas under high-mastery profiles (Economics 90%M) the globally isotropic Heat Kernel wins. We attribute this to each operator’s inductive bias: locality-preserving propagation confines the positive signal to its true neighborhood, avoiding false positives in the dominant unmastered class when mastered concepts are rare, while broad isotropic aggregation reinforces correct labels when

positive evidence is dense. Topology matters too: larger, more heterogeneous graphs (Pharmacology) underperform smaller, cohesive ones (Economics), and the gap persists even with adaptive selection. This motivates adaptive operators that jointly account for student profile and local topology.

(Q3) Effective question selection depends on targeting uncertain concepts outside already-queried regions. Table 3 shows adaptive selection significantly outperforms random, with Frontier best overall, indicating that pairing uncertainty with outward expansion into unqueried concepts is most beneficial. Cutmass, by contrast, underperforms in most settings: it favors structurally central clusters but ignores the learner’s evolving belief state, reinforcing that effective selection must combine belief state with topology, as neither alone suffices. Under high-mastery profiles the gap between selectors shrinks, as positive evidence is dense enough for any selector to recover the state. Still, performance stays sensitive to the learner profile: even the best methods recover mastery far better on small graphs where mastery is the majority and struggle on large heterogeneous graphs when mastery is rare, an open challenge.

8 Conclusion

We study interactive student knowledge assessment and introduce SAGE, a framework that formulates assessment as adaptive querying over graph-structured knowledge spaces under explicit interaction budgets. SAGE deploys an educational agent to construct textbook-derived knowledge graphs, cluster localized knowledge regions, and generate questions conditioned on these regions, then simulates learner interactions to support standardized evaluation of assessment accuracy and query efficiency. Evaluating across knowledge graphs from multiple domains, diverse learner profiles, and a range of querying strategies, we show that methods leveraging graph structure and propagating signals across related concepts outperform baselines. Our results also expose clear limitations of existing methods, notably degraded recovery of the minority mastered class under skewed distributions, highlighting an open challenge for robust, interaction-efficient assessment. SAGE thus provides a general framework for developing and comparing adaptive algorithms that recover student knowledge under limited interactions.

References

- [1] Jiaxin Bai, Wei Fan, Qi Hu, Qing Zong, Chunyang Li, Hong Ting Tsang, Hongyu Luo, Yauwai Yim, Haoyu Huang, Xiaofang Zhou, Feng Qin, Tianshi Zheng, Xi Peng, Xin Yao, Huiwen Yang, Leijie Wu, Yi Ji, Gong Zhang, Renhai Chen, and Yangqiu Song. 2025. AutoSchemaKG: Autonomous Knowledge Graph Construction through Dynamic Schema Induction from Web-Scale Corpora. *CoRR abs/2505.23628* (2025). arXiv:2505.23628 doi:10.48550/ARXIV.2505.23628
- [2] Sergey Brin and Lawrence Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Comput. Networks* 30, 1-7 (1998), 107–117. doi:10.1016/S0169-7552(98)00110-X
- [3] Haoting Chen, Sergio José Rodríguez Méndez, and Pouya Ghiasnezhad Omran. 2025. Open Local Knowledge Graph Construction from Academic Papers Using Generative Large Language Models. In *Companion Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025 - 2 May 2025*, Guodong Long, Michale Blumstein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov (Eds.). ACM, 2551–2559. doi:10.1145/3701716.3717820
- [4] Mingming Chen, Konstantin Kuzmin, and Boleslaw K. Szymanski. 2014. Extension of Modularity Density for overlapping community structure. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*. 856–863. doi:10.1109/ASONAM.2014.6921686
- [5] Yaoyu Chen, Yuheng Hu, and Yingda Lu. 2025. Predicting Field Experiments with Large Language Models. *CoRR abs/2504.01167* (2025). arXiv:2504.01167 doi:10.48550/ARXIV.2504.01167
- [6] Albert T. Corbett and John R. Anderson. 2005. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction* 4 (2005), 253–278. <https://api.semanticscholar.org/CorpusID:19228797>
- [7] Alessandro Epasto, Silvio Lattanzi, and Renato Paes Leme. 2017. Ego-Splitting Framework: from Non-Overlapping to Overlapping Clusters. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Halifax, NS, Canada) (KDD '17)*. Association for Computing Machinery, New York, NY, USA, 145–154. doi:10.1145/3097983.3098054
- [8] Chakib Fettaf, Lazhar Labiod, and Mohamed Nadif. 2022. Efficient Graph Convolution for Joint Node Representation Learning and Clustering. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (Virtual Event, AZ, USA) (WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 289–297. doi:10.1145/3488560.3498533
- [9] Smadar Galili. 2014. Cooperative Learning by Teacher Trainers as a Tool for Improving Teacher Training in Heterogeneous Elementary School Track Classes. *American Journal of Educational Research* 2, 8 (2014), 568–576. doi:10.12691/education-2-8-2
- [10] M. Girvan and M. E. J. Newman. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99, 12 (2002), 7821–7826. arXiv:<https://www.pnas.org/doi/pdf/10.1073/pnas.122653799> doi:10.1073/pnas.122653799
- [11] Frank Harary and R. Z. Norman. 1960. Some properties of line digraphs. *Rendiconti del Circolo Matematico di Palermo* 9 (1960), 161–168. <https://api.semanticscholar.org/CorpusID:122473974>
- [12] Mingguo He, Zhewei Wei, and Ji-Rong Wen. 2022. Convolutional neural networks on graphs with chebyshev approximation, revisited. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (New Orleans, LA, USA) (NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 527, 13 pages.
- [13] Mark Jerrum and Alistair Sinclair. 1988. Conductance and the rapid mixing property for Markov chains: the approximation of permanent resolved. In *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing (Chicago, Illinois, USA) (STOC '88)*. Association for Computing Machinery, New York, NY, USA, 235–244. doi:10.1145/62212.62234
- [14] Johannes Klicpera, Aleksandar Bojchevski, and Stephan Günnemann. 2019. Predict then Propagate: Graph Neural Networks meet Personalized PageRank. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. <https://openreview.net/forum?id=H1gL-2A9Ym>
- [15] Ehsan Latif, Gengchen Mai, Matthew Nyaaba, Xuansheng Wu, Ninghao Liu, Guoyu Lu, Sheng Li, Tianming Liu, and Xiaoming Zhai. 2023. Artificial General Intelligence (AGI) for Education. *CoRR abs/2304.12479* (2023). arXiv:2304.12479 doi:10.48550/ARXIV.2304.12479
- [16] Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. 2025. BIG5-CHAT: Shaping LLM Personalities Through Training on Human-Grounded Data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, 20434–20471. <https://aclanthology.org/2025.acl-long.999/>
- [17] Yuxin Liu, Zeqing Song, Jiong Lou, Chentao Wu, and Jie Li. 2025. AgentTutor: Empowering Personalized Learning with Multi-Turn Interactive Teaching in Intelligent Education Systems. arXiv:2601.04219 [cs.CY] <https://arxiv.org/abs/2601.04219>
- [18] Xinyi Lu and Xu Wang. 2024. Generative Students: Using LLM-Simulated Student Profiles to Support Question Item Evaluation. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale (Atlanta, GA, USA) (L@S '24)*. Association for Computing Machinery, New York, NY, USA, 16–27. doi:10.1145/3657604.3662031
- [19] Belinda Mo, Kyssen Yu, Joshua Kazdan, Proud Mpala, Lisa Yu, Chris Cundy, Charilaos I. Kanatsoulis, and Sanmi Koyejo. 2025. KGGen: Extracting Knowledge Graphs from Plain Text with Language Models. *CoRR abs/2502.09956* (2025). arXiv:2502.09956 doi:10.48550/ARXIV.2502.09956
- [20] Hiromi Nakagawa, Yusuke Iwasawa, and Yutaka Matsuo. 2019. Graph-based Knowledge Tracing: Modeling Student Proficiency Using Graph Neural Network. In *IEEE/WIC/ACM International Conference on Web Intelligence (Thessaloniki, Greece) (WI '19)*. Association for Computing Machinery, New York, NY, USA, 156–163. doi:10.1145/3350546.3352513
- [21] Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023 - 1 November 2023*, Sean Follmer, Jeff Han, Jürgen Steimle, and Nathalie Henry Riche (Eds.). ACM, 2:1–2:22. doi:10.1145/3586183.3606763
- [22] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep Knowledge Tracing. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/bac9162b47c56fc8a4d2a519803d51b3-Paper.pdf
- [23] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3980–3990. doi:10.18653/V1/D19-1410
- [24] Nicy Scaria, Suma Dharani Chenna, and Deepak Subramani. 2024. How Good are Modern LLMs in Generating Relevant and High-Quality Questions at Different Bloom's Skill Levels for Indian High School Social Science Curriculum?. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Ekaterina Kochmar, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anais Tack, Victoria Yaneva, and Zheng Yuan (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 1–10. <https://aclanthology.org/2024.bea-1.1/>
- [25] Alexander Scarlatos, Ryan S. Baker, and Andrew Lan. 2025. Exploring Knowledge Tracing in Tutor-Student Dialogues using LLMs. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25)*. Association for Computing Machinery, New York, NY, USA, 249–259. doi:10.1145/3706468.3706501
- [26] Alexander Scarlatos, Nigel Fernandez, Christopher Ormerod, Susan Lottridge, and Andrew Lan. 2025. SMART: Simulated Students Aligned with Item Response Theory for Question Difficulty Prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 25071–25094. doi:10.18653/v1/2025.emnlp-main.1274
- [27] Kim Schildkamp, Fabienne M. van der Kleij, Maaike C. Heitink, Wilma B. Kippers, and Bernard P. Veldkamp. 2020. Formative assessment: A systematic review of critical teacher prerequisites for classroom practice. *International Journal of Educational Research* 103 (2020), 101602. doi:10.1016/j.ijer.2020.101602
- [28] Patrick Sutanto, Joan Santoso, Esther Irawati Setiawan, and Aji Prasetya Wibawa. 2025. LLM Distillation for Efficient Few-Shot Multiple Choice Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 8502–8530. doi:10.18653/v1/2025.findings-emnlp.452
- [29] Anyi Wang, Dong Shu, Yifan Wang, Yunpu Ma, and Mengnan Du. 2025. Improving LLM Reasoning through Interpretable Role-Playing Steering. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 731–751. doi:10.18653/v1/2025.findings-emnlp.39
- [30] Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7407–7416. doi:10.18653/v1/2024.findings-acl.441
- [31] Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Yordan Zaykov, José Miguel Hernández-Lobato, Richard E. Turner, Richard G. Baraniuk, Craig Barton, Simon Peyton Jones, Simon Woodhead, and Cheng Zhang. 2021. Instructions and Guide for Diagnostic Questions: The NeurIPS 2020 Education Challenge. arXiv:2007.12061 [cs.CY] <https://arxiv.org/abs/2007.12061>

- [32] Zichao Wang, Jakob Valdez, Debshila Basu Mallick, and Richard G. Baraniuk. 2022. Towards Human-Like Educational Question Generation with Large Language Models. In *Artificial Intelligence in Education - 23rd International Conference, AIED 2022, Durham, UK, July 27-31, 2022, Proceedings, Part 1 (Lecture Notes in Computer Science, Vol. 13355)*, Maria Mercedes T. Rodrigo, Noboru Matsuda, Alexandra I. Cristea, and Vania Dimitrova (Eds.). Springer, 153–166. doi:10.1007/978-3-031-11644-5_13
- [33] Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2024. ChatIE: Zero-Shot Information Extraction via Chatting with ChatGPT. arXiv:2302.10205 [cs.CL]. <https://arxiv.org/abs/2302.10205>
- [34] Tao Wu, Jingyuan Chen, Wang Lin, Mengze Li, Yumeng Zhu, Ang Li, Kun Kuang, and Fei Wu. 2025. Embracing Imperfection: Simulating Students with Diverse Cognitive Levels Using LLM-based Agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 9887–9908. doi:10.18653/v1/2025.acl-long.488
- [35] Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can Large Language Model Agents Simulate Human Trust Behavior?. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). http://papers.nips.cc/paper_files/paper/2024/hash/1cb57fcf7f3f6d37eebae5becc9e6d-Abstract-Conference.html
- [36] Gui Yang, Jiale Zou, Yan Chen, Liang Du, and Peng Zhou. 2024. Heat Kernel Diffusion for Enhanced Late Fusion Multi-View Clustering. *IEEE Signal Processing Letters* 31 (2024), 2310–2314. doi:10.1109/LSP.2024.3449229
- [37] Bowen Zhang and Harold Soh. 2024. Extract, Define, Canonicalize: An LLM-based Framework for Knowledge Graph Construction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 9820–9836. doi:10.18653/v1/2024.emnlp-main.548
- [38] Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. 2017. Dynamic Key-Value Memory Networks for Knowledge Tracing. In *Proceedings of the 26th International Conference on World Wide Web (Perth, Australia) (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 765–774. doi:10.1145/3038912.3052580
- [39] Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, Jianing Yin, Shuai Wang, Quanyu Dai, Zhenhua Dong, Hongning Wang, and Minlie Huang. 2025. SocialEval: Evaluating Social Intelligence of Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2025, Vienna, Austria, July 27 - August 1, 2025, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, 30958–31012. <https://aclanthology.org/2025.acl-long.1496/>

A Textbooks Used for Graph Construction

Table 5 lists the publicly available textbooks used as source materials for knowledge graph construction. The resulting graphs (Table 1) vary in scale and density, from the 5,708-node Economics graph to the 34,241-node Physiology graph. Although we use textbooks, the framework applies equally to other resources such as lecture notes and slides.

Table 5: Textbooks used for knowledge graph construction.

Domain	Abbr.	Textbook
Economics	ECON	<i>Principles of Economics</i> , Mankiw
CS	CS	<i>Data Communications and Networking</i> , Forouzan
Pharmacology	Pharm	<i>Basic & Clinical Pharmacology</i> , Katzung et al.
Physiology	Physio	<i>Guyton and Hall Medical Physiology</i> , Hall

B Comparison with Knowledge Tracing Methods

SAGE differs fundamentally from existing knowledge tracing families. DKT [22] (LSTM over concept-response pairs), DKVMN [38] (key-value memory), GKT [20] (GNN propagation over expert-tagged skills), and recent LLM-based methods [25] all perform next-response prediction on offline logs and assume a flat or expert-curated concept space with each concept assessed repeatedly, typically requiring 10K+ annotated interaction turns to reach a stable ceiling. These assumptions are incompatible with our setting, where questions are generated over thousands of automatically extracted triplet nodes, each queried once under a strict budget. In contrast, SAGE obtains solid mastery estimates from as few as 50–300 adaptive queries, reflecting the real-world constraint the benchmark models.

C Clustering

C.1 Hyperparameter Selection for GCC

We select the propagation power and modularity weight λ for GCC_{hard} in two stages. Stage 1 is a coarse joint search over power $\in \{5, 10, 15, 20, 25\}$ and $\lambda \in \{0, 0.001, 0.01, 0.1, 0.5, 1.0\}$ (3 runs, 50 iters); power is evenly spaced across local/global regimes, λ log-spaced with 0 as a no-regularization baseline. Stage 2 refines around the top-3 combinations within ± 2 power and a $\times 4$ λ range (5 runs, 100 iters). All are scored by $Q + \text{silhouette} - \text{conductance}$. Modularity regularization contributes meaningfully to quality; final values are in Table 6.

Table 6: Selected GCC hyperparameters per domain.

Dataset	power	λ	Q	Silh.	Cond.	Score
Physiology	27	0.075	0.773	0.290	0.227	0.836
CS	26	0.075	0.783	0.312	0.137	0.958
Economics	25	0.050	0.737	0.244	0.163	0.818
Pharm	27	0.075	0.837	0.255	0.118	0.974

C.2 Soft Clustering Evaluation Metrics

We use four metrics. **Overlapping modularity** [4], $Q_{ov} = \sum_c \sum_{i,j \in c} (A_{ij} - k_i k_j / 2m) O_i^{-1} O_j^{-1}$ (O_i : clusters containing node i), and **conductance** [13], $\phi(c) = E_c^{\text{out}} / (2E_c^{\text{in}} + E_c^{\text{out}})$, measure community structure and separation (higher Q_{ov} , lower ϕ are better). **Coverage**, $\text{Cov} = |\{(i, j) \in E : \mathcal{M}(i) \cap \mathcal{M}(j) \neq \emptyset\}| / |E|$, is the fraction of edges whose endpoints share a cluster. **Duplication utility**, $\text{DU}(u) = \frac{1}{|\mathcal{M}(u)|} \sum_{c \in \mathcal{M}(u)} |\{q \in \mathcal{Q} : u \in q\}| / K$, measures how consistently a duplicated node u is referenced across the K questions of its clusters, justifying its placement.

D Belief Propagation and Question Selection

D.1 Diffusion Operator Formulations

PPR/APPNP [2, 14]: anchored local diffusion via $r(I - (1 - r)\hat{A})^{-1}$, $\hat{A} = D^{-1}A$, with restart r controlling anchoring near queried nodes. **Heat Kernel** [36]: isotropic diffusion via e^{-tL} , $L = D - A$, with time t setting the spread radius. **Chebyshev** [12]: finite-hop polynomial diffusion $\sum_{j=0}^K c_j T_j(\tilde{L})$, limited to a K -hop neighborhood.

Distance-Weighted (DW): no diffusion; sets $\sigma_t = \mathbf{k}_t$ and updates unqueried nodes via $w_v = 1/(d(v, \mathcal{S}_t) + 1)$. Table 4 shows propagation operators consistently beat DW, with performance best on the smaller Economics graph, then CS, then Pharmacology.

D.2 Sensitivity to LLM Choice

To rule out self-preference bias from using one model in all roles, we compare a single-model pipeline (gpt-oss-120b generates, answers, judges) against a cross-model one (gpt-oss generates, Qwen2.5-72B answers, Llama-3.3-70B judges) on CS at 50%M with PPR and Frontier. Macro-F1 differs by under 0.03 at every budget (Table 7), indicating SAGE’s findings reflect methodological design rather than single-model artifacts.

Table 7: Sensitivity to LLM choice (CS, 50%M, PPR + Frontier).

Budget	Single (gpt-oss)	Cross (oss/Qwen/Llama)	Δ
50	0.617	0.632	+0.015
100	0.723	0.710	−0.013
200	0.824	0.797	−0.027
300	0.843	0.824	−0.019

D.3 Reproducibility

Table 8 lists all hyperparameters, obtained via grid search over the four mastery profiles, selecting the configuration maximizing mean Macro-F1 for robustness across profiles.

Table 8: Hyperparameters for all methods.

Component	Parameter	Value
PPR/APPNP	r, λ	0.15, 0.1
Heat Kernel	t, λ	1.0, 0.1
Chebyshev	K, λ	3, 0.1
All selectors	λ_{cov}, h	0.3, 1
Boundary	μ	0.4
CutMass	μ_{cut}	0.25
Frontier	μ_f	0.5
Response scoring	w_a, w_r	0.75, 0.25