Filling the Long Tail: Structure-Aware Curriculum-Gap Completion for Medical Education with LLMs

Wenjie Lin
School of Applied and Creative Computing
Purdue University
lin1790@purdue.edu

Abstract

Medical education resources are dense for common diseases but often sparse for under-covered conditions, atypical presentations, and fine-grained concept distinctions. This creates curriculum gaps that are difficult to repair manually, especially in long-tail domains where structured teaching materials are limited. We introduce Curriculum-Gap Completion (CGC), a new task for Large Language Model (LLM)-based medical education in which a model reconstructs missing educational units from a partially specified curriculum graph. Given topic nodes, pedagogical relations, and structured teaching slots, the model predicts omitted concepts, restores missing instructional links, and completes automatically verifiable teaching content. We instantiate this setting in a long-tail medical case study (hyperhidrosis) and evaluate five LLMs under three methods: direct prompting, retrieval-augmented prompting, and our proposed Structure-Aware Curriculum-Gap Completion (SACGC) framework. Across models, SACGC achieves the strongest overall performance, with the largest gains on structurally demanding masking settings. Ablation results show that explicit graph structure is the most important component, while schema constraints provide additional benefit. These findings suggest that LLMs are better suited for reconstructing an under-specified educational structure than for unrestricted medical tutoring, and they motivate CGC as a new natural language processing (NLP) problem for healthcare education.

1 Introduction

Medical education is necessarily selective: curricula devote substantial attention to common diseases, canonical presentations, and core management patterns, while more limited space is available for under-covered conditions, atypical presentations, and fine-grained distinctions that sit in the clinical long tail (Helderman et al., 2024). This imbalance has practical consequences. In rare-disease and underrepresented-condition education, recent studies have documented persistent gaps in learner awareness, sparse curricular exposure, and calls for more systematic early teaching rather than isolated exposure later in training (Huynh et al., 2025).

At the same time, LLMs are rapidly entering medical education. Recent reviews describe growing use of LLMs for tutoring, content generation, assessment support, and virtual patients, and they suggest that these systems can improve access to educational materials and interactive learning support (Zhang et al., 2025). However, the same reviews also highlight recurring limitations: unstable factual grounding, weak control over reasoning quality, and uncertainty about how these systems should be integrated into structured learning rather than open-ended question answering alone (Lucas et al., 2024a).

A second limitation is methodological. Much of the recent medical LLM literature evaluates performance through exam-style question answering or final-answer accuracy (Lin and Wei-Kocsis, 2025). Although these settings are useful, they do not directly capture whether a model can support curriculum design, recover missing prerequisite structure, or expand educational coverage in domains where teaching material is sparse. Recent analyses in clinical reasoning and benchmark design have explicitly argued that high benchmark scores can mask reasoning failures and that final-answer evaluation alone is often insufficient for understanding model behavior in medically meaningful settings (McCoy et al., 2025). In this paper, we argue that an important and underexplored use of LLMs in healthcare education is CGC. Instead of treating the model as an unrestricted tutor, we ask whether it can help recover missing educational structure from a partially specified curriculum: omitted concepts, missing prerequisite links, and missing contrastive teaching units. This framing is motivated by a sim-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 941–950

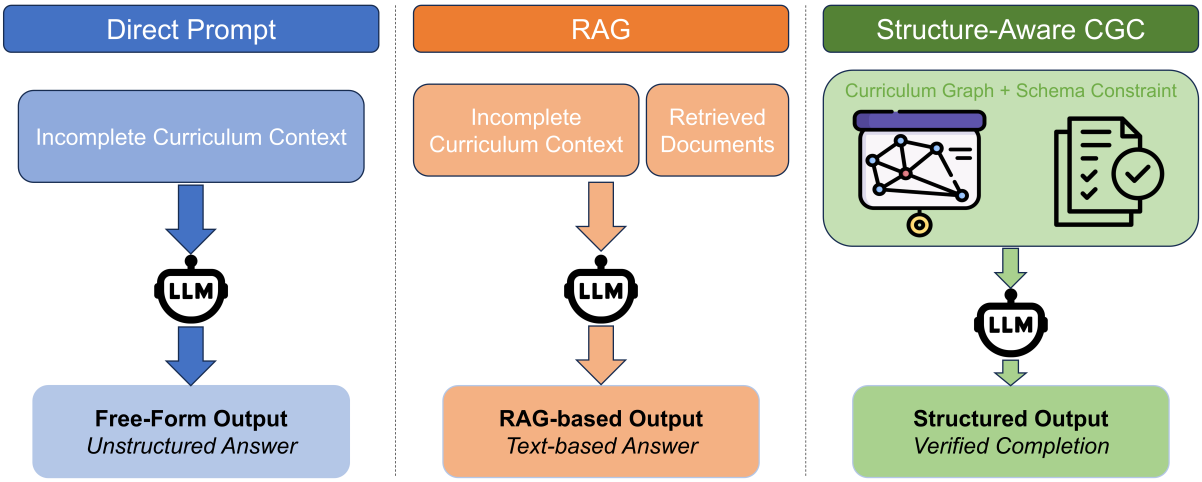

Figure 1: Comparison of three approaches for CGC. **Direct Prompt** relies on incomplete contexts and produces unstructured outputs. **RAG** augments context with retrieved evidence but remains text-based. In contrast, **Structure-Aware CGC** conditions on explicit curriculum graph structure and schema constraints, enabling structured and verifiable completion of missing concepts, relations, and teaching units.

ple educational observation: long-tail domains are often not entirely absent from curricula, but they are frequently under-connected, under-sequenced, or weakly scaffolded. A system that can repair those missing links may be more useful than one that only answers isolated questions.

To operationalize this idea, we introduce CGC for medical education, a new task in which an LLM receives a partial curriculum structure and must reconstruct missing educational units in automatically verifiable forms. Our framing draws on prior work in educational knowledge graphs and concept-graph learning, where prerequisite structure and concept relations are treated as central to learning progression, but adapts these ideas to long-tail medical education and LLM-based reasoning (Liang et al., 2015).

We instantiate this setting in a long-tail medical case study and study whether LLMs can recover missing concepts and instructional relations more effectively when reasoning over partial curriculum structure than when answering in a direct prompt-based manner. This task is appealing for medical NLP because it is educationally meaningful, automatically evaluable, and distinct from conventional medical QA. Rather than asking whether a model knows the final answer to a question, we ask whether it can help *reconstruct what should be taught, in what order, and through which conceptual distinctions*.

Our work makes three contributions. First, we formulate CGC as a new NLP task for healthcare education. Second, we propose SACGC, a structured framework for recovering missing curriculum units, prerequisite relations, and teaching attributes from partial curriculum graphs. Third, we provide an automatically verifiable evaluation setting in a long-tail medical domain and show that structure-aware completion offers a more suitable use of LLMs for under-covered educational content than unrestricted medical tutoring.

2 Related Work

2.1 LLMs in the Medical Domain and Medical Education

LLMs have been studied broadly across the medical domain, including clinical question answering (Lucas et al., 2024b), diagnostic support (Tretow et al., 2025; Lin and Wei-Kocsis, 2025), reasoning assistance (Goh et al., 2024), summarization (Balde et al., 2025), and patient-facing communication (Diego et al., 2026). Much of this literature has focused on benchmark performance on medical exams, licensing-style questions, and structured clinical tasks, often showing strong surface-level performance while also raising concerns about hallucination, calibration, and weak evidence grounding (Singhal et al., 2025; Wang et al., 2025). More recent evaluations have argued that strong benchmark scores do not necessarily reflect robust clinical reasoning or safe real-world use, especially when tasks are dominated by common conditions and final-answer metrics (McCoy et al., 2025).

Within medical education, LLMs have been in-

creasingly explored for tutoring, question generation, feedback, content drafting, learner support, and virtual patient simulation. Recent reviews suggest that these models can improve access to learning resources and support interactive educational workflows, but they also highlight recurring limitations such as factual instability, difficulty controlling pedagogical structure, and the lack of clear integration into curriculum design (Lucas et al., 2024a; Vrdoljak et al., 2025). Existing educational uses are therefore centered largely on assisting isolated learning activities, rather than repairing missing curriculum structure in under-covered domains.

Our work differs from this line of research in two ways. First, we do not study unrestricted tutoring or free-form educational generation. Second, we focus on long-tail medical knowledge and experiment on a case study for hyperhidrosis, which is relatively under-recognized in clinical practice and often lacks systematic coverage in medical education despite its clinical relevance (Lin and Fang, 2022). Here, the main challenge is often not answering a single question but identifying what concepts and relations are missing from a coherent learning sequence. This makes our problem formulation closer to curriculum modeling than to conventional medical QA.

2.2 Curriculum Structure, Prerequisite Modeling, and Educational Knowledge Graphs

Curriculum mapping has long been used in health professions education to identify omissions, redundancies, and sequencing problems in instructional programs (Komenda et al., 2015). In parallel, educational data mining and NLP have studied prerequisite learning as a structured prediction problem, where concepts are connected by dependency relations that shape learning order and instructional progression (Liang et al., 2015; Watson et al., 2020). This work has motivated the use of concept graphs and educational knowledge structures to support learning analytics, content sequencing, and instructional planning.

Despite this progress, prior prerequisite and curriculum-graph work has largely been developed for general educational settings or broad scientific domains, rather than for healthcare education and the medical long tail. In medicine, the problem is especially important because under-covered conditions and atypical distinctions are often not entirely absent from curricula, but weakly connected to prerequisite concepts or downstream teaching units. This makes long-tail medical education a natural setting for structure-aware completion rather than only content generation.

Our work builds on the intuition from curriculum mapping and prerequisite modeling that educational quality depends not only on what is taught, but also on how concepts are connected. We extend this perspective to LLM-based medical education by formulating missing educational content as a recoverable graph-structured object and by designing automatically verifiable completion tasks over partially observed curriculum structures.

3 Problem Formulation

Medical curricula are constrained by time and content overload, which makes complete and well-sequenced coverage difficult in practice. This problem is especially visible in the long tail of medicine, where under-covered conditions, atypical presentations, and fine-grained distinctions are more likely to appear as omissions or weak prerequisite links than as completely absent topics. Curriculum mapping has long been used to expose such structural gaps in health professions education, while educational NLP has studied prerequisite relations as a formal dependency structure between concepts.

We formalize CGC as a structured prediction task over a partially observed curriculum graph. Let a curriculum be represented as a directed graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}),$$

where $\mathcal{V}$ is a set of curriculum units, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is a set of directed prerequisite or instructional-support relations, and $\mathcal{A}$ denotes node attributes such as learning objectives, evidence passages, concept type, or assessment items.

Each node $v \in \mathcal{V}$ corresponds to a fine-grained educational unit, for example, a disease concept, a subtype distinction, a red-flag pattern, or a management principle. Each edge $(u, v) \in \mathcal{E}$ indicates that mastering u is pedagogically useful or necessary before v.

In the CGC task, the model is given an incomplete curriculum graph

$$\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}}, \tilde{\mathcal{A}})$$

obtained by masking a subset of nodes, edges, or node attributes from the full graph $\mathcal{G}$. The goal is to recover the missing curriculum content

$$\Delta = \{\mathcal{V} \setminus \tilde{\mathcal{V}},\ \mathcal{E} \setminus \tilde{\mathcal{E}},\ \mathcal{A} \setminus \tilde{\mathcal{A}}\},$$

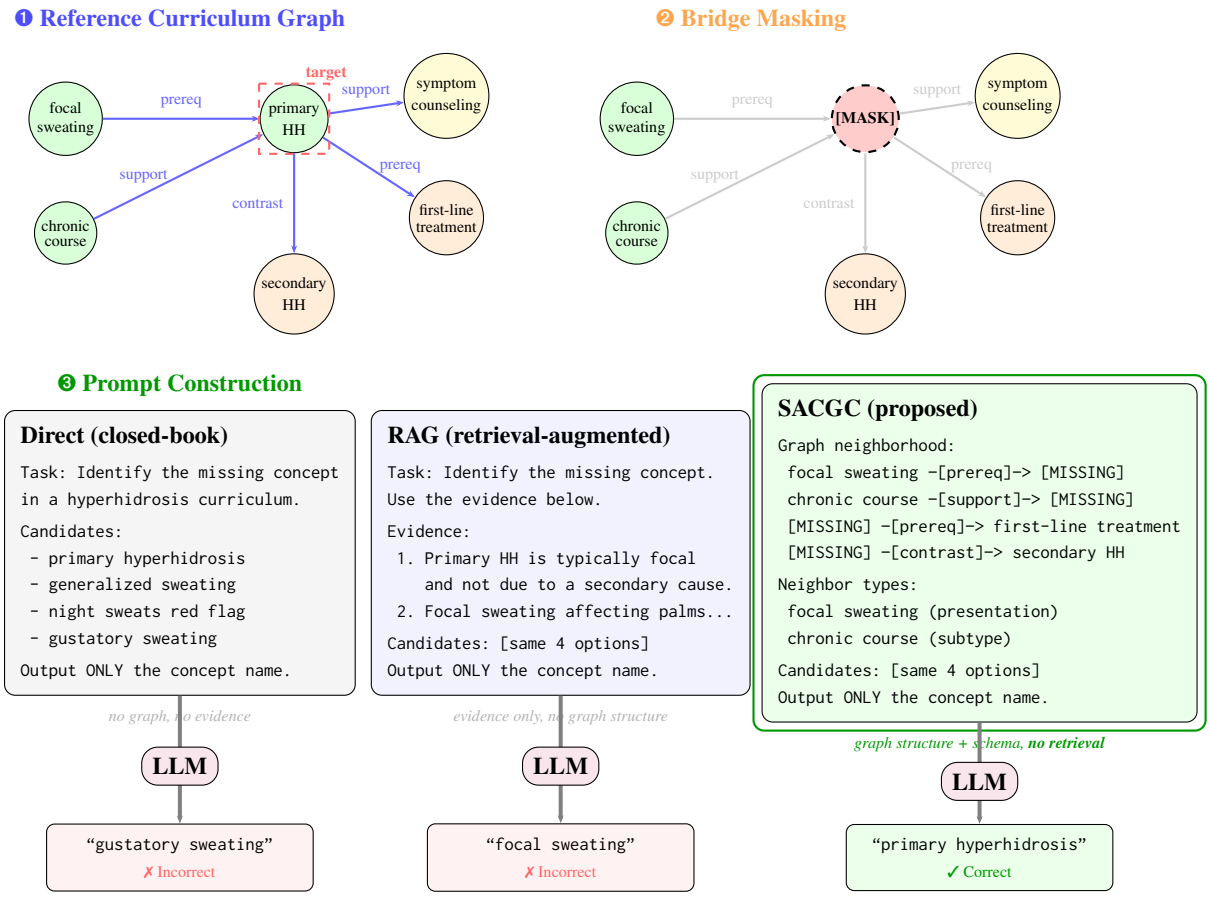

Figure 2: A Demonstration of the workflow of CGC task on a node recovery task (bridge masking). **Top**: The node *primary hyperhidrosis* is masked from the reference graph, breaking the learning path from symptom patterns to treatment. **Middle**: Three prompting conditions receive the same task. Direct and RAG lack structural context; SACGC sees typed edges and neighbor metadata. **Bottom**: Only SACGC correctly infers the missing concept by reasoning over the graph topology.

subject to structural and semantic constraints.

We consider three concrete subtasks.

First, **missing concept recovery** predicts omitted curriculum units from local graph context and source evidence. Formally, given a masked neighborhood $\mathcal{N}(v)$, the system predicts the missing node $\hat{v}$.

Second, **prerequisite relation completion** predicts whether a directed edge should exist between two nodes. This is a binary or multiclass edge prediction task.

Third, **teaching-unit completion** fills structured node attributes such as learning objective, contrastive distinction, common misunderstanding, and assessment item. Unlike unrestricted generation, these outputs are schema-constrained and automatically checkable against held-out gold fields.

The central hypothesis is that long-tail educational gaps are more naturally modeled as *missing structure* than as isolated question answering. Rather than asking whether a model can answer a medical question, CGC asks whether it can reconstruct what should be taught, how it should be connected, and what minimal additions restore a coherent learning path.

4 Overview of SACGC Model

As shown in Figure 1, our proposed SACGC framework solves curriculum repair through *structure-aware prompting* and *schema-constrained prediction*. Given a partially observed curriculum graph $\tilde{\mathcal{G}}$, SACGC predicts the missing unit by conditioning on its local pedagogical context rather than treating the instance as an isolated text problem. The framework is designed to preserve two properties of the task: explicit curriculum structure during inference and deterministic verification at output time.

4.1 Local Curriculum Context Construction

For each masked target, SACGC first extracts a local subgraph centered on the missing node, edge, or attribute. This context includes the visible neighboring units, their concept types, and any surviving pedagogical relations. The resulting neighborhood serves as a compact curriculum state that identifies what is already known and what remains to be reconstructed.

This step is important because the missing target is often ambiguous without local instructional structure. For example, the same medical concept may play different pedagogical roles depending on whether it appears between symptom description and subtype distinction, or between subtype distinction and management. SACGC grounds its prediction in the *position* of the missing content within the curriculum graph, not only in its textual description.

4.2 Structure-Aware Prompting

SACGC converts each local curriculum context into a task-specific prompt. The prompt explicitly exposes the information needed for structured recovery: visible neighboring nodes, node types, available edge labels, and the required prediction format.

For **missing concept recovery**, the prompt describes the masked neighborhood and asks the model to infer the omitted curriculum unit. For **prerequisite relation completion**, the prompt presents two nodes together with their surrounding context and requires one label from a closed set such as `prerequisite`, `support`, `contrastive`. For **teaching-unit completion**, the prompt specifies the target node and requests a fixed schema, such as learning objective, key distinction, and misconception to avoid.

Unlike direct prompting, SACGC does not ask the model to answer a medical question in isolation. Instead, it asks the model to reconstruct a missing curriculum object within an explicit pedagogical structure.

4.3 Schema-Constrained Prediction

A central design choice in SACGC is that outputs are constrained to automatically verifiable forms. Node predictions are normalized to canonical concept identifiers, edge predictions are restricted to a closed relation inventory, and teaching-unit generation is limited to a predefined slot schema. This reduces ambiguity at both inference and evaluation time.

These constraints are not only technical conveniences. They also align the model with the goal of curriculum repair. In this setting, a fluent answer is not sufficient unless it can be mapped back to a concrete educational unit, relation, or structured teaching field. Schema constraints encourage the model to produce outputs that are curriculum-compatible rather than merely plausible.

4.4 Inference Objective

Given a masked target z and its serialized local curriculum context, SACGC predicts

$$\hat{z} = g_\theta\big(\phi(\tilde{\mathcal{G}}, z)\big),$$

where $\phi(\tilde{\mathcal{G}}, z)$ denotes the structure-aware prompt constructed from the partial graph and g_θ is the underlying LLM. Depending on the task, $\hat{z}$ may be a missing node, an edge label, or a structured attribute set. The output is then normalized into its canonical representation before evaluation.

In this way, SACGC treats curriculum repair as structured completion over partial instructional graphs. The framework is intended to recover not only missing content, but missing pedagogical organization, which is especially important in long-tail medical education, where weak sequencing and absent conceptual bridges often matter as much as the content itself.

5 Experiments

5.1 Experimental Setup

We evaluate the CGC task on a case study domain of hyperhidrosis, a long-tail medical topic with fine-grained distinctions and limited coverage in standard benchmarks. The workflow is illustrated in Figure 2. We construct a reference curriculum graph with 50 nodes (concepts) and 59 edges (relations) following a two-phase annotation process. In the first phase, node candidates were identified by reviewing ten publicly available medical education resources on hyperhidrosis, including clinical review articles and guideline-style summaries, which form our evidence bank of 73 text chunks. Each node was required to correspond to a distinct, nameable instructional unit. In the second phase, edges were assigned according to a fixed rubric based on the three relation types defined in Section 4. Node attributes were written to be faithful to source documents and cross-checked against the evidence

Model	Direct				RAG				**SACGC (Ours)**			
	Overall	Random	Long-tail	Bridge	Overall	Random	Long-tail	Bridge	Overall	Random	Long-tail	Bridge
Llama-3.1-8B-Instruct	0.304±0.050	0.318±0.147	**0.367±0.229**	0.311±0.184	0.316±0.084	0.193±0.062	0.228±0.108	0.400±0.130	**0.496±0.035**	**0.567±0.125**	0.323±0.136	**0.585±0.117**
Qwen2.5-7B-Instruct	0.289±0.095	0.219±0.072	0.136±0.042	0.340±0.169	0.506±0.059	0.409±0.064	0.310±0.028	0.588±0.189	**0.602±0.146**	**0.624±0.057**	**0.455±0.053**	**0.614±0.206**
Llama-3.1-70B-Instruct	0.425±0.064	0.247±0.100	0.283±0.149	0.524±0.215	0.523±0.072	0.378±0.050	0.294±0.156	0.642±0.017	**0.642±0.129**	**0.470±0.216**	**0.601±0.174**	**0.737±0.069**
Qwen2.5-72B-Instruct	0.499±0.059	0.414±0.065	0.335±0.169	0.550±0.174	0.471±0.046	0.454±0.124	0.234±0.082	0.503±0.028	**0.595±0.138**	**0.568±0.130**	**0.370±0.050**	**0.667±0.122**
GPT-4o	0.434±0.099	0.316±0.050	0.341±0.070	0.455±0.172	0.437±0.067	0.394±0.086	0.301±0.038	0.527±0.087	**0.698±0.089**	**0.606±0.092**	**0.512±0.093**	**0.731±0.081**

Table 1: Main comparison across five LLMs. For each LLM, we compare **Direct**, **RAG**, and **SACGC** using the aggregate **CGC-Score** (mean ± std). Under each method, we report the overall score and the scores under the three masking settings: **Random**, **Long-tail**, and **Bridge**.

Model	Direct				RAG				**SACGC (Ours)**			
	Node Acc	Edge F1	Slot F1	Path EM	Node Acc	Edge F1	Slot F1	Path EM	Node Acc	Edge F1	Slot F1	Path EM
Llama-3.1-8B-Instruct	0.093±0.081	0.486±0.166	0.136±0.021	0.444±0.385	0.111±0.096	0.615±0.309	**0.165±0.030**	0.778±0.385	**0.556±0.096**	**0.651±0.056**	0.143±0.014	**0.889±0.192**
Qwen2.5-7B-Instruct	0.537±0.142	0.793±0.297	0.179±0.006	0.833±0.289	0.463±0.038	0.873±0.219	**0.187±0.028**	**0.889±0.192**	**0.778±0.192**	**0.883±0.208**	0.164±0.021	0.611±0.347
Llama-3.1-70B-Instruct	0.130±0.116	0.870±0.224	0.176±0.012	0.833±0.289	0.259±0.160	0.664±0.143	**0.188±0.020**	1.000±0.000	**0.833±0.167**	**0.877±0.214**	0.174±0.012	**1.000±0.000**
Qwen2.5-72B-Instruct	0.519±0.028	0.485±0.150	0.167±0.002	0.000±0.000	0.444±0.160	**0.594±0.314**	**0.174±0.019**	**0.167±0.289**	**0.537±0.112**	0.496±0.106	0.169±0.017	0.000±0.000
GPT-4o	0.130±0.116	0.870±0.224	0.176±0.012	0.833±0.289	0.259±0.160	0.664±0.143	**0.188±0.020**	1.000±0.000	**0.833±0.167**	**0.877±0.214**	0.174±0.012	**1.000±0.000**

Table 2: Overall metric breakdown across five LLMs and three methods (mean ± std).

bank. Each node is associated with structured attributes (e.g., learning objective, key distinction, misconception) and linked evidence snippets from a curated evidence bank for the RAG method.

To simulate curriculum incompleteness, we generate masked graphs $\tilde{\mathcal{G}}$ by removing nodes, edges, or attributes. We consider three masking regimes: **Random**, **Long-tail**, and **Bridge**. Random masking removes nodes or edges uniformly at random and serves as a general completion setting. Long-tail masking preferentially removes low-frequency or distinction-heavy concepts, simulating realistic under-coverage in medical curricula. Bridge masking removes intermediate prerequisite units and tests restoration of coherent learning paths.

5.2 Tasks

We evaluate three core tasks: (1) **Concept recovery**: predict missing curriculum nodes from local graph context; (2) **Relation completion**: predict edge labels from a closed set (`prerequisite`, `support`, `contrastive`, `none`); (3) **Slot completion**: fill structured attributes (e.g., learning objective, key distinction). We additionally evaluate **path restoration** on a subset of bridge-masked instances.

5.3 Models

We compare three settings: **Direct**: prompt the model with incomplete curriculum context; **RAG**: augment context with retrieved evidence; **SACGC (ours)**: provide graph-structured context and schema-constrained prompts. All outputs are normalized to canonical forms before scoring.

5.4 Evaluation Metrics

All tasks are evaluated with deterministic metrics. We report: Node Acc for concept (node) recovery; Edge F1 for relation (edge) completion; Slot F1 for attribute completion; Path EM (Exact Match) for path restoration. We also report a combined score: CGC-Score $= 0.35\,\mathrm{NodeAcc} + 0.30\,\mathrm{EdgeF1} + 0.25\,\mathrm{SlotF1} + 0.10\,\mathrm{PathEM}$. Weights are set proportional to task centrality and metric reliability; all component metrics are also reported individually.

5.5 Ablation

We study two ablations of SACGC: (1) **w/o Graph**: remove structured neighborhood information; (2) **w/o Schema**: replace constrained outputs with free-form generation.

5.6 Hypotheses

We test three hypotheses: (H1) SACGC improves structured recovery over Direct and RAG, especially under long-tail and bridge masking; (H2) retrieval alone is insufficient for curriculum reconstruction; (H3) both graph context and schema constraints are necessary for strong performance.

6 Results

6.1 Main Comparison Across LLMs and Methods

Table 1 reports the primary comparison across five LLMs under **Direct**, **RAG**, and **SACGC**. The main result is clear: **SACGC** achieves the best overall CGC-Score for all five models. This pattern holds across both smaller and larger models, indicating that curriculum-aware completion is beneficial beyond any single model family or scale.

Model	SACGC				SACGC w/o Graph				SACGC w/o Schema			
	Overall	Random	Long-tail	Bridge	Overall	Random	Long-tail	Bridge	Overall	Random	Long-tail	Bridge
Llama-3.1-8B-Instruct	**0.496±0.035**	**0.567±0.125**	**0.323±0.136**	**0.585±0.117**	0.329±0.101	0.193±0.062	0.251±0.085	0.400±0.130	0.289±0.024	0.346±0.267	0.240±0.139	0.359±0.142
Qwen2.5-7B-Instruct	**0.602±0.146**	**0.624±0.057**	0.455±0.053	**0.614±0.206**	0.506±0.060	0.409±0.064	0.310±0.028	0.588±0.189	0.498±0.022	0.458±0.154	**0.468±0.249**	0.511±0.067
Llama-3.1-70B-Instruct	**0.642±0.129**	0.470±0.216	**0.601±0.174**	**0.737±0.069**	0.522±0.072	0.377±0.051	0.293±0.156	0.641±0.016	0.591±0.056	**0.528±0.125**	0.560±0.156	0.650±0.173
Qwen2.5-72B-Instruct	0.595±0.138	**0.568±0.130**	0.370±0.050	**0.667±0.122**	0.483±0.031	0.514±0.152	0.227±0.072	0.503±0.028	**0.605±0.189**	0.565±0.147	**0.535±0.192**	0.575±0.210
GPT-4o	**0.698±0.089**	**0.606±0.092**	**0.512±0.093**	**0.731±0.081**	0.454±0.081	0.396±0.083	0.326±0.044	0.538±0.078	0.622±0.077	0.563±0.090	0.475±0.090	0.611±0.025

Table 3: Ablation study for the proposed SACGC model. We compare the full **SACGC** model against **SACGC w/o Graph** and **SACGC w/o Schema**. Each cell reports the aggregate **CGC-Score** (mean ± std) for the overall setting and for each masking type.

The gains are especially pronounced for stronger models. On GPT-4o, SACGC improves the overall score from 0.434 under Direct and 0.437 under RAG to **0.698**. Similar gains appear for Llama-3.1-70B-Instruct, where SACGC reaches **0.642**, compared with 0.425 for Direct and 0.523 for RAG. Qwen2.5-72B-Instruct and Qwen2.5-7B-Instruct show the same pattern, with SACGC outperforming both baselines overall. Even on Llama-3.1-8B-Instruct, where the absolute scores are lower, SACGC remains the strongest setting.

A second important pattern appears in the masking-specific results. SACGC performs particularly well under Bridge masking, which is the most structurally demanding setting because it requires recovering missing intermediate prerequisite units rather than only local content. For all five models, the best Bridge score is achieved by SACGC. This is important because Bridge masking most directly reflects the motivating educational problem in this paper: long-tail curricula are often incomplete, not because concepts are entirely absent, but because conceptual bridges and prerequisite links are missing.

SACGC also performs strongly under Random masking, again achieving the best score for every model. On Long-tail masking, SACGC is best in four out of five models and is only slightly below Direct on Llama-3.1-8B-Instruct. Overall, these results suggest that curriculum-aware prompting is especially valuable when the missing content has structural importance, and that its benefits generalize across different model families.

Finally, the comparison with RAG is informative. Although RAG is often stronger than Direct, it does not match SACGC consistently. This suggests that evidence access alone is not sufficient for this task. What matters is not only providing the model with more text, but guiding it with explicit curriculum structure so that it can infer what educational unit is missing and how that unit should connect to the rest of the learning path.

6.2 Metric Breakdown

Table 2 breaks the overall score into its four components: node accuracy, edge macro-F1, slot token-F1, and path exact match. This analysis clarifies where the gains from SACGC come from.

The most consistent improvements appear in Node ACC, Edge F1, and Path EM. For GPT-4o, SACGC reaches **0.833** node accuracy, **0.877** edge macro-F1, and **1.000** path exact match. Llama-3.1-70B-Instruct shows the same pattern, with SACGC producing the strongest values on all three of these metrics. Qwen2.5-7B-Instruct also benefits substantially from SACGC, especially on node accuracy and edge macro-F1. These trends indicate that the main strength of the proposed framework lies in reconstructing curriculum structure: identifying missing concepts, recovering pedagogical relations, and restoring coherent instructional paths.

In contrast, slot F1 remains relatively low across all methods and all models. Even when SACGC gives the best overall score, its advantage on slot completion is small, and in several cases, RAG is slightly stronger on this metric. This shows that structured teaching-unit generation remains the hardest component of the task. Recovering a missing node or edge is easier than generating a compact but educationally appropriate textual slot, such as a learning objective or misconception field. We therefore view slot completion as an open challenge rather than a fully solved part of the framework.

The metric breakdown also helps explain why SACGC improves the overall score even when the slot gains are modest. Since CGC-Score combines node recovery, edge recovery, slot quality, and path restoration, improvements in structural metrics can outweigh smaller differences in slot token-F1. This matches the intended use of the framework: SACGC is designed primarily to repair missing curriculum structure, not to maximize unrestricted text generation quality.

A final observation is that stronger models ben-

efit more clearly from structure-aware completion. GPT-4o and Llama-3.1-70B-Instruct show the most complete improvements across metrics, while smaller models exhibit gains that are more selective. This suggests that curriculum-aware prompting does not replace model capability, but instead gives stronger models a better inductive bias for recovering missing educational structure.

6.3 Ablation Study

Table 3 evaluates two ablations of the proposed method: **SACGC w/o Graph** and **SACGC w/o Schema**. These experiments isolate the contribution of the two main design choices in our framework: explicit curriculum structure and schema-constrained generation.

Removing graph structure causes a substantial performance drop in most cases. The effect is particularly large for GPT-4o, where the overall score falls from 0.698 to 0.454, and for Llama-3.1-70B-Instruct, where it drops from 0.642 to 0.522. Llama-3.1-8B-Instruct shows a similarly clear decline, from 0.496 to 0.329. These results show that the graph is not merely a convenient representation; it provides useful structural information that the model relies on to recover missing curriculum content. The same pattern is visible under Bridge masking, where the full SACGC model consistently outperforms the graph-ablated version. This is strong evidence that explicit curriculum structure is central to the success of the method.

Removing schema constraints also hurts performance, but the effect is somewhat more variable across models. For Llama-3.1-8B-Instruct, the score drops from 0.496 to 0.289, showing that structured output formats are particularly helpful for weaker models. GPT-4o and Llama-3.1-70B-Instruct also decline without schema constraints, although the degradation is smaller than in the no-graph setting. For Qwen2.5-72B-Instruct, however, the no-schema variant is slightly stronger overall than the fully constrained model. This suggests that schema constraints are generally useful for stabilizing structured generation, but their effect depends more on model family than graph structure does.

Taken together, the ablations indicate that **graph structure is the more important component** of the proposed framework. Schema constraints are also beneficial in most cases, but they play a secondary role relative to the explicit representation of instructional units and pedagogical links. This finding is important because it strengthens the main claim of the paper: the core gain from SACGC does not come only from output formatting, but from treating missing educational content as missing curriculum structure.

Overall, the ablation results confirm that both components contribute to performance, but they do so differently. Graph structure provides the strongest and most stable gains, especially on structurally demanding masking settings, while schema constraints improve the reliability of structured outputs and are most helpful for models with weaker instruction following.

7 Conclusion

We introduced CGC as a new NLP task for medical education. Instead of treating LLMs as unrestricted tutors, our framework models under-covered educational content as missing curriculum structure: omitted concepts, broken pedagogical links, and incomplete teaching units. Across five LLMs, the proposed SACGC framework consistently outperformed direct prompting and retrieval-augmented prompting in overall score, with especially strong gains on structurally demanding settings. Ablation results further showed that explicit curriculum graph structure is the main source of improvement, while schema constraints provide additional support for reliable structured generation.

These results suggest that LLMs may be most useful in medical education not as free-form tutors, but as tools for repairing incomplete instructional structure in long-tail domains. While this study focuses on a single domain, the CGC framework is domain-agnostic: applying it to a new long-tail medical topic requires only a domain-specific curriculum graph and a linked evidence bank. Future work will explore automated graph construction and multi-domain evaluation to reduce annotation overhead and broaden coverage. More broadly, our work identifies CGC as a practical and automatically evaluable direction at the intersection of NLP, medical education, and knowledge organization.

8 Limitations

Our study has several limitations. First, we evaluate the framework on a single case-study domain, which limits claims about cross-domain generalization, though the framework formulation and pipeline are not specific to hyperhidrosis. Second, the reference curriculum graph is manually designed, so the choice of instructional units and

prerequisite relations reflects modeling decisions that other educators may define differently. Third, our evaluation focuses on automatically verifiable outputs. This improves reproducibility, but it does not capture broader pedagogical qualities such as explanatory richness, learner engagement, or downstream learning gains.

In addition, completion performance depends on the quality and coverage of the evidence bank. In long-tail domains, sparse or uneven source material may constrain both retrieval and generation. Finally, our framework is intended as a structured support tool for curriculum analysis and repair, not as an autonomous medical educator or a substitute for expert-designed instruction.

References

Gunjan Balde, Soumyadeep Roy, Mainack Mondal, and Niloy Ganguly. 2025. Evaluation of llms in medical text summarization: The role of vocabulary adaptation in high oov settings. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22989–23004.

Fajardo V Diego, Oleksii Proniakin, Victoria-Elisabeth Gruber, and Razvan Marinescu. 2026. Medpi: Evaluating ai systems in medical patient-facing interactions. *medRxiv*, pages 2025–12.

Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, and 1 others. 2024. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA network open*, 7(10):e2440969.

Ron Helderman, Carolyn M Macica, Adam Weinstein, Richard Feinn, and Maya Doyle. 2024. Addressing challenges in diagnosis and management of rare disease through interprofessional education. *Rare*, 2:100044.

Sharon Huynh, Eric L. Wan, Angelette Pham, Robin Yoon, Scott Dorris, Nada Yazigi, and Jessica M. Jones. 2025. Rare disease education in medical schools: patient-centered and innovative strategies. *Orphanet Journal of Rare Diseases*, 20(1):596.

Martin Komenda, Martin Víta, Christos Vaitsis, Daniel Schwarz, Andrea Pokorná, Nabil Zary, and Ladislav Dušek. 2015. Curriculum mapping with academic analytics in medical and healthcare education. *PLOS ONE*, 10(12):e0143748.

Chen Liang, Zhaohui Wu, Wenyi Huang, and C. Lee Giles. 2015. Measuring prerequisite relations among concepts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1668–1674, Lisbon, Portugal. Association for Computational Linguistics.

Wenjie Lin and Yajun Fang. 2022. Primary hyperhidrosis: A systematic review of current status and potential interventions. In *2022 6th International Conference on Universal Village (UV)*, pages 1–8. IEEE.

Wenjie Lin and Jin Wei-Kocsis. 2025. Llm4sweat: A trustworthy large language model for hyperhidrosis support. *arXiv preprint arXiv:2508.15192*.

Harrison C. Lucas, Jeffrey S. Upperman, and Jamie R. Robinson. 2024a. A systematic review of large language models and their implications in medical education. *Medical Education*, 58(11):1276–1285.

Mary M Lucas, Justin Yang, Jon K Pomeroy, and Christopher C Yang. 2024b. Reasoning with large language models for medical question answering. *Journal of the American Medical Informatics Association*, 31(9):1964–1975.

Liam G. McCoy, Natasha Sagar, Shanjida Bacchi, Jeffrey M. N. Fong, Nicholas C. K. Tan, and Abraham Rodman. 2025. Assessment of large language models in clinical reasoning: A novel benchmarking study. *NEJM AI*, 2(10).

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip A. Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Isabel Tretow, Moritz Schwebel, Stefan Feuerriegel, Theresa Treffers, and Isabell M Welpe. 2025. The effect of llm assistance on diagnostic accuracy: A meta-analysis. *medRxiv*, pages 2025–12.

Josip Vrdoljak, Zvonimir Boban, Marino Vilović, Marko Kumrić, and Joško Božić. 2025. A review of large language models in medical education, clinical decision support, and healthcare administration. *Healthcare*, 13(6):603.

Chong Wang, Mengyao Li, Junjun He, Zhongruo Wang, Erfan Darzi, Zan Chen, Jin Ye, Tianbin Li, Yanzhou Su, Jing Ke, Kaili Qu, Shuxin Li, Yi Yu, Pietro Liò, Tianyun Wang, Yu Guang Wang, and Yiqing Shen. 2025. A survey for large language models in biomedicine. *Artificial Intelligence in Medicine*, 170:103268.

Eilean G. S. Watson, Carole Steketee, Kylie Mansfield, Maxine Moore, Bronwen Dalziel, Arvin Damodaran, Ben Walker, Robbert J. Duvivier, and Wendy Hu. 2020. Curriculum mapping for health professions education: a typology. *Focus on Health Professional Education: A Multi-Professional Journal*, 21(1):19–37.

Qi Zhang, Zijing Huang, Yuqiang Huang, Geng Wang, Riping Zhang, Jianling Yang, Yinglin Cheng, Binyao Chen, Hongxi Wang, Kunliang Qiu, and 1 others.

2025. Generative ai in medical education: feasibility and educational value of llm-generated clinical cases with mcqs. *BMC Medical Education*, 25(1):1502.