Curriculum Gap Filling: A Benchmark for Course Catalog Completion

Anonymous ACL submission

Abstract

Course recommendation systems rank resources for a learner or a goal, but course-catalog completion requires a different decision: whether the existing catalog already covers the required competencies and what should be added when it does not. We introduce CURRICULUM GAP FILLING, a new task that connects three decisions in a shared competency space: auditing an internal catalog, acquiring external resources only when they add missing coverage, and designing new course content for gaps that remain unresolved. We build a 500-scenario Chinese benchmark and a 100-scenario English transfer set. Internal courses expose verified course–competency links, while external resources are visible only as text, forcing systems to infer competency coverage from titles, descriptions, syllabi, prerequisites, and related content. The benchmark includes hard negatives, text-inferability audits, human-audited prerequisite targets, GAPFLOW as an inspectable reference baseline, and oracle controls for attributing errors across the pipeline. Experiments show that relevance-based retrieval and LLM prompting often select goal-related resources that add little curriculum coverage. Candidate-pool and oracle analyses identify external text-to-competency coverage estimation as the main acquisition bottleneck, while end-to-end and human evaluations show that compact missing-course design remains a separate challenge.

1 Introduction

Course catalogs are often treated as retrieval spaces: given a learner or a goal, a system ranks existing courses or recommends a learning path (Chen et al., 2023; Cheng et al., 2026; Ma et al., 2025). This framing supports advising and personalization while leaving a curriculum-planning question unanswered: a catalog can contain relevant courses and still miss required competencies. A machine-learning engineering catalog may cover data preprocessing, model training, and software testing while omitting containerization or CI/CD. Course rankings alone leave catalog sufficiency unresolved.

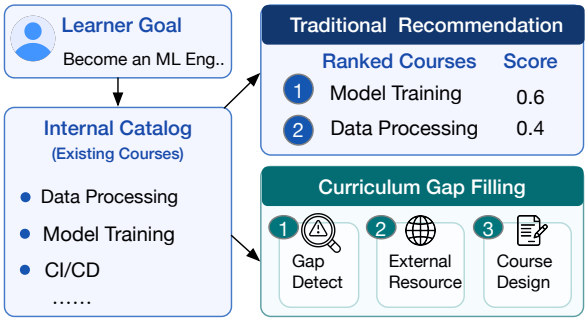

Figure 1: Existing recommendation settings rank available resources. CURRICULUM GAP FILLING evaluates catalog sufficiency, remaining gaps, and the decision to acquire or design.

Catalog completion requires coverage reasoning under an information asymmetry. Internal courses have verified links to a competency graph, reflecting institutional knowledge of one's own catalog. External resources arrive only as text: titles, descriptions, syllabi, prerequisites, or article summaries. A system must infer whether an external resource covers a missing competency, whether that coverage is new relative to the internal catalog, and whether remaining gaps require new course design.

We formalize this decision as CURRICULUM GAP FILLING: a system first audits internal catalog coverage, then acquires external text resources for uncovered competencies, and finally designs a missing course when acquisition leaves gaps unresolved (Figure 1) [1]. We instantiate the task as a controlled MOOC-based benchmark over MOOCCubeX across ten fields with a text-only external pool, and additionally construct an English Lec-

[1] We use "curriculum" in the educational sense of course catalogs, not in the machine-learning sense of ordering training examples.

tureBank set for an acquisition-focused transfer check.

Our experiments test whether the benchmark separates ordinary relevance from curriculum value. Results show that retrieval and prompting baselines often select resources that are related to the goal but add little gap coverage; oracle variants identify external text-to-competency coverage estimation as the main acquisition bottleneck; and end-to-end evaluation shows that compact missing-course design remains difficult even when external coverage improves. GAPFLOW serves as an inspectable reference baseline for locating these errors across the pipeline. Our contributions are:

- We define CURRICULUM GAP FILLING, a new task where systems audit an internal catalog, acquire external resources only when they add missing competency coverage, and design course content for unresolved gaps.
- We build two controlled benchmarks for this task: 500 Chinese MOOCCubeX scenarios and 100 English LectureBank scenarios. The data are constructed so that topically related resources can still be redundant, and gold acquisition labels require evidence from system-visible text.
- We evaluate systems by whether they change the curriculum coverage state, not only by whether they retrieve relevant resources. The benchmark includes graph-access controls, oracle variants, and GAPFLOW as an inspectable reference baseline for locating errors across the pipeline.

2 Related Work

Educational resources and prerequisite graphs. TutorialBank (Fabbri et al., 2018) collects tutorials and learning resources for NLP and related areas, with topic, prerequisite, and related-resource annotations. LectureBank (Li et al., 2019) organizes lecture materials from AI, ML, NLP, deep learning, and information retrieval courses with topic and prerequisite-chain annotations. MOOCCube and MOOCCubeX (Yu et al., 2020, 2021) provide Chinese MOOC repositories with course–concept links and fine-grained learning materials. These resources make course–concept and prerequisite structure available. CURRICULUM GAP FILLING uses such structure to ask a different question: given a partial catalog, which competencies remain uncovered, and how should they be filled?

Curricular analytics and alignment. Curricular analytics maps course text to skills, outcomes, or program requirements. Course-Skill Atlas (Javadian Sabet et al., 2024) infers skills from large-scale course syllabi and aligns them with occupational activities. CLO–PLO mapping uses NLP to support accreditation by aligning course outcomes with program outcomes (Zaki et al., 2023). Recent work also evaluates LLMs for extracting skills from curriculum documents (Xu et al., 2025). These lines of work estimate alignment between courses and external skill or outcome spaces. Our task starts from a shared competency space and evaluates the downstream decision: whether the catalog is sufficient, which external text resources add new coverage, and what course content is needed when acquisition is insufficient.

Knowledge states and skill extraction. Knowledge tracing estimates a learner's evolving mastery over knowledge components from interaction histories (Piech et al., 2015; Shen and Liu, 2024). CURRICULUM GAP FILLING also maintains a competency state, but at the catalog level: which competencies are already covered, newly acquired, or still missing. Skill-extraction work maps free text, such as job postings or occupational descriptions, to skill taxonomies such as O*NET or ESCO (Zhang et al., 2022; Senger et al., 2024). Our setting uses a similar text-to-label inference step, but evaluates whether the inferred coverage adds non-redundant value relative to an internal catalog.

Course recommendation and learning paths. A separate line of work selects or sequences existing learning items. Recent work studies LLM-based MOOC recommendation (Ma et al., 2025), T5-based course ranking and summarization (Bijl et al., 2024), recommendation explanations (Li et al., 2025), and prerequisite-aware learning-path planning (Chen et al., 2023; Cheng et al., 2026; Wang et al., 2026). Knowledge graphs and LLM grounding further improve recommendation and path planning (Abu-Rasheed et al., 2024, 2025; Hou et al., 2025). These systems optimize relevance, personalization, or path quality over resources. CURRICULUM GAP FILLING evaluates a catalog completion decision: the system must first detect missing coverage, then acquire resources only when they add non-redundant coverage, and finally design new content for unresolved gaps.

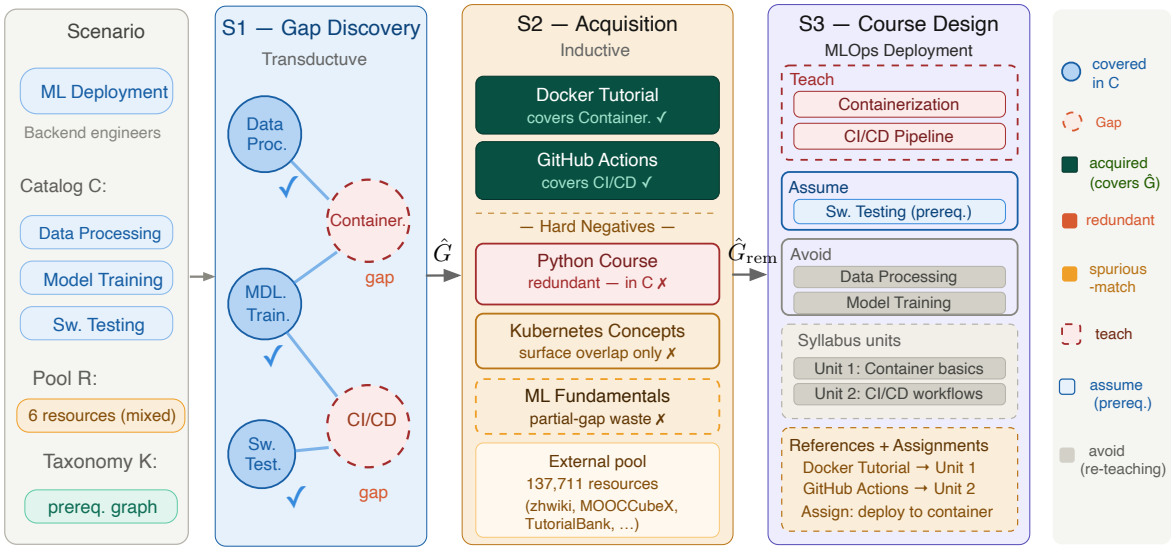

Figure 2: Task workflow for CURRICULUM GAP FILLING. Given a scenario, the system audits internal coverage, acquires external resources by added gap coverage, and designs a course for unresolved competencies. Colors mark covered competencies, gaps, acquisitions, hard negatives, and design constraints.

3 Task Formulation

We formulate CURRICULUM GAP FILLING as a catalog-completion task over a shared competency taxonomy. Each scenario is $x = (u, g, \mathcal{C}, \mathcal{R}, \mathcal{K})$, where u is a learner cohort, g is a learning goal, $\mathcal{C}$ is the internal catalog, $\mathcal{R}$ is a frozen external resource pool, and $\mathcal{K}$ is a competency taxonomy. Let $K_g \subseteq \mathcal{K}$ denote target competencies required by g. The system must maintain a coverage state over K_g: which competencies are already covered internally, which remain uncovered, which external resources add new coverage, and which gaps require new course design.

The task is defined by an information asymmetry. Each internal course $c \in \mathcal{C}$ is provided with verified course–competency links. Each external resource $r \in \mathcal{R}$ is provided only as text (title, description, syllabus or prerequisite text when available, or article lead text). External competency links are reserved for benchmark construction and evaluation. We call Subtask 1 transductive because the system reasons over the visible internal course–competency graph. We call Subtask 2 inductive because external resource–competency links are withheld and must be inferred from text. Figure 2 summarizes the three subtasks formalized below.

Subtask 1: internal coverage audit. Given $(u, g, \mathcal{C}, \mathcal{K})$, the system ranks internal courses and predicts a gap set $\hat{G} \subseteq K_g$. With verified links on internal courses, this subtask probes catalog sufficiency reasoning over a visible graph: a course may be topically related to g while leaving required competencies uncovered.

Subtask 2: external acquisition. Given $(\hat{G}, \mathcal{C}, \mathcal{R}_{\text{text}}, \mathcal{K})$, the system ranks external resources for acquisition. A useful resource must cover competencies in $\hat{G}$ while adding little redundant coverage over $\mathcal{C}$. This is an inductive text-to-competency inference problem: competency links for $\mathcal{R}$ are withheld at inference time, so the system must infer coverage from text alone.

Subtask 3: missing-course design. After acquisition, any competencies still unresolved form $\hat{G}_{\text{rem}}$. Given $\hat{G}_{\text{rem}}$, the internal catalog, and top acquired resources $\mathcal{R}^*_{\text{top}}$ as partial evidence, the system outputs a structured course proposal that teaches $\hat{G}_{\text{rem}}$, assumes appropriate prerequisites, and avoids re-teaching content already covered by $\mathcal{C}$.

Applicable scenarios. The task assumes verified internal course–competency links, reflecting institutional knowledge of courses already in a catalog. This assumption matches settings where curricula are maintained against learning outcomes or skill taxonomies, such as outcome-based accreditation criteria (International Engineering Alliance, 2021), CLO–PLO curriculum mapping (Zaki et al., 2023), MOOC knowledge graphs (Yu et al., 2021), and enterprise learning systems. Programs with only raw course text require an additional concept-extraction step before gap filling. This benchmark evaluates catalog completion after an internal competency mapping has already been established.

4 GAPFLOW

GAPFLOW is an inspectable reference baseline that follows the three subtasks of CURRICULUM GAP FILLING. It is included to make errors in the benchmark traceable, not to serve as a deployed curriculum-planning system. Given a scenario $x = (u, g, \mathcal{C}, \mathcal{R}, \mathcal{K})$, it produces three outputs: a predicted gap set, a ranked list of external resources, and a structured missing-course proposal.

GAPFLOW first audits the internal catalog. For each target competency $k \in K_g$, it aggregates evidence from all visible internal courses in $\mathcal{C}$, using the verified course–competency links and available course text. If the aggregated support for k is below a threshold, k is predicted as a gap. This step gives $\hat{G}$, the competencies that the catalog does not sufficiently cover.

Given $\hat{G}$, GAPFLOW assigns each external resource $r \in \mathcal{R}$ an acquisition score

$$s(r) = \mathrm{GapCov}(r; \hat{G}) - \beta\, \mathrm{Red}(r; \mathcal{C}) - \eta\, \mathrm{Unc}(r).$$

Here $s(r)$ is the score used for ranking external resources. $\mathrm{GapCov}(r; \hat{G})$ rewards predicted coverage of the current gap set, $\mathrm{Red}(r; \mathcal{C})$ penalizes overlap with competencies already covered by the internal catalog, and $\mathrm{Unc}(r)$ discounts uncertain text-based coverage matches. The weights β and η are tuned on the development split. The top-ranked resources are treated as acquisition candidates.

Finally, after acquisition, any still-unresolved competencies form $\hat{G}_{\mathrm{rem}}$. GAPFLOW builds a missing-course proposal for these remaining gaps by assigning competencies to three roles: content to teach, covered catalog content to avoid re-teaching, and prerequisite competencies to assume. The proposal is structured so that each design error can be traced to a missing gap, redundant content, or prerequisite choice. Implementation details and scoring formulas are in Appendix A.

We use two oracle variants to localize errors. GAPFLOW+Oracle Gaps replaces $\hat{G}$ with the gold gap set before external acquisition; the difference from GAPFLOW measures how much S1 gap-prediction error propagates downstream. GAPFLOW+Oracle Coverage also replaces the external text-based coverage estimates with gold resource–competency labels; the remaining difference measures the effect of external coverage estimation.

5 Benchmark Construction

We build CURRICULUM GAP FILLING as a controlled catalog-completion benchmark. The main dataset, CGF-MOOCCUBEX-CN, is built from MOOCCubeX course–competency graphs; a smaller CGF-LECTUREBANK-EN set applies the same protocol to English materials to test transfer (Appendix H). The construction turns a local course neighborhood into an incomplete catalog: we hide one to three competencies from a source course, keep internal courses that still cover the remaining target competencies, and evaluate whether systems can recover the held-out coverage by acquiring external resources or designing a new course.

Data and visibility. The main benchmark spans ten MOOCCubeX fields. Internal courses expose their verified course–competency links, reflecting the information an institution has about its own catalog. External resources come from out-of-catalog MOOCCubeX courses, auxiliary MOOCCubeX learning materials, TutorialBank and LectureBank passages, and Chinese Wikipedia articles. Systems see only external resource text, including titles, descriptions, syllabi, prerequisites, transcripts, problem text, and article leads. External resource–concept annotations are used only to construct labels and evaluate predictions.

Labels and hard negatives. Gold gaps are the held-out competencies that remain uncovered by the retained internal catalog. External resources are positive when their text substantively supports coverage of a gold gap. We filter resource–concept annotations with a text-inferability audit, keeping only majority-retained pairs. The candidate pool also contains hard negatives: resources that are topically related but redundant, resources that cover catalog content rather than gaps, and resources whose text overlaps with a gap term without teaching the competency. These cases test marginal curriculum value rather than topical similarity.

Construction quality. The text-inferability audit retains 100 of 276 audited OOC course–concept pairs (36.2%), with 76.4% three-way agreement and pairwise $\kappa = 0.565$–0.762. Prerequisite targets are separately audited from source-course evidence and capped at three per design-needed scenario. These audits keep the benchmark focused on coverage claims inferable from visible text.

Design supervision. Some gaps have positive external resources; others require missing-course design. For design-needed cases, prerequisite targets are audited from source-course prerequisite evidence, local prerequisite ancestors, and text-derived links. Only targets judged as valid assumed prerequisites enter $P_{\text{gold}}(x)$, capped at three per scenario to capture core prior knowledge. We mark a scenario as design-needed when the top-5 gold external acquisition set does not fully cover all held-out gaps, matching the $K = 5$ budget used by acquisition metrics. Appendix I.4 shows that the design-only share is stable across the swept K values but changes with pool composition.

Scale and scope. CGF-MOOCCUBEX-CN contains 500 scenarios across ten domains and a 137k-resource external pool. The benchmark evaluates controlled gap recovery from course–competency graphs. Real curriculum revision also involves degree requirements, instructor availability, budgets, and learner outcomes. Those factors are left for future work. The LectureBank transfer set checks whether the same construction protocol can be applied to English course materials. All labels are fixed before running any baseline.

6 Experiments

The experiments test whether CURRICULUM GAP FILLING behaves as a diagnostic benchmark. We first compare retrieval, prompting, course recommendation, and coverage-aware reference systems on the main benchmark. We then check candidate-pool effects, evaluate acquisition transfer on LectureBank, and test whether isolated subtask gains translate into end-to-end catalog completion. Additional component diagnostics are in Appendix I.

6.1 Evaluation Setup

Datasets. CGF-MOOCCUBEX-CN contains 500 scenarios across ten MOOCCubeX fields, each with an incomplete internal catalog of eight visible courses, a target competency set, one to three held-out gap competencies, and a frozen external retrieval pool. CGF-LECTUREBANK-EN contains 100 scenarios built from LectureBank by aggregating lecture-level topics and prerequisite links into instructor-grouped courses under the same protocol; it differs in language, source corpus, and domain coverage.

Metrics. Subtask 1: NDCG@5, CoverageRank over non-gap competencies (CR-v), gap precision and recall (Gap-P, Gap-R). Reporting precision and recall separately makes high-recall shortcuts visible. Subtask 2: Marginal Gap Coverage over the externally acquirable gap subset (MGC-acq@K), Redundancy Rate (RR@K, overlap between acquired coverage and the internal catalog), and Mean Utility (MU@K, the number of newly covered competencies normalized by the top-K acquisition budget). Subtask 3: CompCov (CC), Design Redundancy (DRed; lower is better), and Prerequisite Over-listing Rate (POLR; lower is better). End-to-end: Final Gap Resolution (FGR), False Completion Rate (FCR; lower is better), and Prerequisite-Compact Completion (PCC). Unless stated otherwise, all values average over the full $N{=}500$ benchmark. Formal definitions are in Appendix E.

Relevance-only baselines. BM25 (Robertson and Zaragoza, 2009), MiniLM, mContriever (Izacard et al., 2022), and BGE-M3 (Chen et al., 2024) rank resources by text similarity; Hybrid combines BM25 and mContriever via reciprocal rank fusion; ConceptMatch uses concept names and aliases without graph structure.

LLM baselines. We evaluate Qwen3-8B (Yang et al., 2025) in six variants reported in Table 1: *Direct* runs all three subtasks in a single pass; *Stepwise* decomposes them into separate calls; *OracleRR* adds pairwise reranking with gold positives mixed into the candidate pool to upper-bound in-pool selection; *+IntGraph* and *+Closure* progressively prepend the visible internal course–competency graph and its transitive coverage closure; *CourseRec** adapts the course-recommendation prompt design of Ma et al. (Ma et al., 2025) to Qwen3-8B: it asks the model to recommend useful resources and complete the curriculum, but does not require an explicit coverage-state update after each acquisition. We also run Claude Haiku (Stepwise) with the same decomposed S1–S3 prompt protocol to test whether a stronger LLM backbone changes the full-pipeline pattern. Prompts are in Appendix F.

Diagnostic reference. GAPFLOW uses the visible internal graph for Subtask 1 and text-only coverage inference for external resources. GAPFLOW+Oracle Gaps and GAPFLOW+Oracle Coverage supply gold gaps and gold external coverage labels respectively.

Statistical testing. We use paired bootstrap over scenarios (1000 iterations, seed 20260520) on

Table 1: Main results on CGF-MOOCCUBEX-CN (N=500). Lower is better for RR, DRed, and POLR. **Bold**: strongest non-oracle entry per column. GAPFLOW is a diagnostic reference with its own full-corpus phrase-anchored index; LLM variants are compared under the same retrieved K=50 pool. Pool-restricted diagnostics are in Appendix I.2. High Gap-R should be read with Gap-P; several retrieval baselines obtain high recall by over-predicting gaps.

	S1: internal				S2: acquisition			S3: design		
Method	NDCG@5	CR-v	Gap-P	Gap-R	MGC-acq@5	RR@5↓	MU@5	CC	DRed↓	POLR↓
Random	0.717	0.767	0.449	0.175	0.000	0.024	0.000	–	–	–
BM25	0.732	0.781	0.461	0.997	0.016	0.389	0.030	–	–	–
MiniLM	0.778	0.853	0.461	**1.000**	0.001	0.093	0.000	–	–	–
BGE-M3	0.767	0.767	0.503	0.939	0.033	0.304	0.019	–	–	–
mContriever	0.796	0.822	0.463	0.999	0.013	0.197	0.009	–	–	–
Hybrid	0.810	0.832	0.461	**1.000**	0.028	0.369	0.037	–	–	–
ConceptMatch	**0.982**	**0.910**	0.493	0.752	0.000	0.126	0.003	–	–	–
Qwen3-8B (Direct)	0.754	0.810	0.479	0.668	0.051	0.253	0.127	0.320	0.896	0.950
Qwen3-8B (Stepwise)	0.829	0.798	0.561	0.618	0.065	0.296	0.087	0.188	0.554	0.580
Qwen3-8B (OracleRR)	0.829	0.798	0.561	0.618	0.116	0.288	0.063	0.183	0.576	0.610
Qwen3-8B (+IntGraph)	0.865	0.830	0.608	0.500	0.065	0.284	0.087	0.181	0.530	0.545
Qwen3-8B (+Closure)	0.740	0.787	0.114	0.311	0.036	0.252	0.043	0.121	0.394	0.424
CourseRec*	0.807	0.838	0.461	0.997	0.058	0.379	0.081	**0.345**	0.933	0.960
Claude Haiku (Stepwise)	0.953	0.902	**0.848**	0.470	0.137	0.127	0.169	0.109	0.187	0.260
GAPFLOW	0.976	0.909	0.812	0.332	**0.155**	**0.018**	**0.234**	0.030	**0.015**	**0.168**
+Oracle Gaps	0.976	0.909	1.000	1.000	0.184	0.018	0.234	0.030	0.015	0.168
+Oracle Coverage	0.976	0.909	1.000	1.000	1.000	0.177	0.148	0.408	0.136	0.991

MGC-acq@5. Full bootstrap confidence intervals are in Appendix I.

6.2 Main Results

Table 1 yields three main observations, each tied to a capability that CURRICULUM GAP FILLING is designed to test.

Relevance ranking does not reduce to curriculum value. Dense retrievers and BM25 score well on S1 ranking but have negligible MGC-acq@5 and high RR@5, showing that semantic similarity often retrieves already-covered content. ConceptMatch attains the highest non-oracle NDCG@5 and CR-v while contributing essentially no acquisition value. Their high Gap-R is mostly a high-recall shortcut: they mark nearly every target competency as missing while Gap-P stays low. S1 ranking quality and S2 acquisition value are therefore separable axes on this benchmark.

The two oracle variants attribute S2 error to coverage estimation, not gap discovery. The jump from GAPFLOW to GAPFLOW+Oracle Gaps on MGC-acq@5 is modest, indicating that S1 gap-prediction error propagates only weakly downstream. The larger jump to GAPFLOW+Oracle Coverage identifies external text-to-competency coverage estimation as the main acquisition bottleneck. Appendix I.1 further shows that a supervised TF–IDF verifier improves pairwise coverage prediction but still yields weak acquisition under poor candidate exposure, so pairwise accuracy alone is insufficient.

LLM prompting with explicit graph context does not collapse the S2 gap. *+IntGraph* and *+Closure* expose the visible internal graph and its transitive closure, but neither approaches GAPFLOW's MGC-acq@5; *+Closure* therefore serves as a graph-access control rather than a successful repair. *OracleRR* is the closest Qwen3-8B variant but still trails, while CourseRec* improves design coverage without fixing acquisition. On Subtask 3, LLM variants over-list prerequisites and achieve low CompCov, showing that missing-course design is not measured by acquisition alone. A forced-design control in Appendix I.8 shows the same pattern under gold gaps and gold references: stronger LLMs improve coverage only partially, while redundant course content remains common.

6.3 Candidate Pool Recall

A natural concern is whether LLM acquisition is limited by the retrieved candidate pool. Table 2 measures this ceiling on the 477 CGF-MOOCCUBEX-CN scenarios with at least one gold-positive resource. The 3-way hybrid pool (BM25-CJK ∪ mContriever ∪ phrase) improves as K grows: Recall rises from 0.392 at $K = 50$ to

Table 2: Candidate-pool recall on CGF-MOOCCubeX-CN, restricted to the 477 scenarios with ≥ 1 gold-positive resource. **Pool oracle**: greedy in-pool MGC-acq@5 ceiling under the Table 1 gap normalization.

K	BM25-CJK	mContriever	Hybrid	Reach	Pool oracle
50	0.174	0.048	0.392	68.1%	0.261
100	0.218	0.078	0.473	74.6%	0.292
200	0.278	0.103	0.543	79.7%	0.314
500	0.365	0.161	0.603	83.9%	0.335

Table 3: Acquisition results on CGF-LectureBank-EN. MGC-acq@5 with 95% paired-bootstrap CI. We abbreviate GapFlow and Oracle Coverage as GF and OC.

Method	**MGC-acq@5**	**95% CI**
BM25	0.052	[0.015, 0.100]
BGE-M3	0.000	[0.000, 0.000]
mContriever	0.005	[0.000, 0.015]
Hybrid	0.180	[0.120, 0.252]
LLM Direct	0.400	[0.310, 0.487]
LLM Stepwise	0.020	[0.000, 0.045]
LLM Reranker	0.037	[0.007, 0.073]
GF	**0.637**	[0.548, 0.715]
GF+OC	1.000	[1.000, 1.000]

0.603 at $K = 500$, and scenario reach rises from 68.1% to 83.9%. Candidate generation is therefore a real bottleneck.

The pool oracle separates this retrieval ceiling from in-pool selection. Even at $K = 50$, the in-pool ceiling is 0.261 MGC-acq@5, well above the LLM scores in Table 1; at $K = 500$, the ceiling increases to 0.335. Thus, LLM acquisition is limited by both candidate generation and coverage judgment within the pool. Since all LLM variants in Table 1 use the same $K = 50$ pool, their comparisons remain controlled for retrieval access. GapFlow uses a separate phrase-anchored index over the full external corpus, so its score should be read as a diagnostic reference rather than as a same-pool reranker comparison.

6.4 Results on CGF-LectureBank-EN

Table 3 reports an acquisition-focused transfer check on the 100-scenario CGF-LectureBank-EN set. Relevance-only retrievers remain weak and the oracle ceiling remains high, matching the main acquisition pattern. At the same time, LLM Direct is much stronger on LectureBank than on CGF-MOOCCubeX-CN, indicating that acquisition difficulty depends on language, source corpus, and candidate density. These results support protocol portability for acquisition evaluation, rather than distribution-invariant performance across all subtasks. Full English results, including GapFlow ablations, are in Appendix H.

Table 4: End-to-end completion on CGF-MOOCCubeX-CN (N=500). **FGR**: gap resolution. **FCR**↓: false completion (claims complete with gaps unresolved). **PCC**: prerequisite-compact completion.

Method	FGR	FCR↓	PCC
Qwen3-8B (Direct)	0.306	0.307	0.182
CourseRec*	0.354	0.167	0.220
Claude Haiku (Stepwise)	0.177	0.777	0.113
GapFlow	0.155	0.773	0.102
GapFlow+Oracle Coverage	0.790	0.000	0.700

6.5 End-to-End Completion

Isolated subtask scores do not directly translate to end-to-end catalog completion. Table 4 reports FGR, FCR, and PCC for representative LLM pipelines, GapFlow, and GapFlow+Oracle Coverage. These metrics are diagnostic rather than human judgments of pedagogical quality: they measure label-level gap resolution, false completion, and prerequisite compactness.

End-to-end completion exposes errors that isolated subtasks hide. CourseRec* obtains the highest non-oracle FGR and lowers FCR, but remains far below the oracle on PCC. GapFlow shows the opposite failure, often judging the catalog complete and emitting empty designs. Oracle Coverage removes false completion and improves PCC, but FGR remains below 1.0. Thus, better external coverage estimation helps, but does not by itself solve usable missing-course design.

6.6 Human Evaluation of Course Designs

We further check whether the automatic S3 metrics reflect human judgments of generated course designs. We sample 50 design-required scenarios from CGF-MOOCCubeX-CN. To isolate Subtask 3 from upstream acquisition errors, each system receives the gold gap set and gold reference resources and is asked to produce a missing-course proposal. We restrict this evaluation to generative course-design baselines: Qwen3-8B (Direct), CourseRec*, and Claude Haiku (Stepwise). Outputs are anonymized and randomly ordered. Two annotators score each proposal on a 1–5 scale for gap coverage, non-redundancy, prerequisite appropriateness, and overall usability; we report mean scores across annotators.

Table 5: Human evaluation of generated missing-course designs on 50 design-required scenarios. Scores are on a 1–5 scale; higher is better.

Method	Cov.	Non-red.	Prereq.	Overall
Qwen3-8B	3.01	1.91	1.97	2.46
CourseRec*	2.82	2.28	1.72	2.19
Claude Haiku	3.40	2.60	2.25	2.75

Table 5 shows that the stronger LLM backbone improves human-rated coverage and usability, but all systems remain weak on prerequisite appropriateness. The results also match the main diagnostic message of the automatic S3 metrics: designs that repeat catalog content or list broad prerequisites are less usable, even when they cover some target gaps. We therefore treat S3 metrics as diagnostic signals for coverage, redundancy, and prerequisite compactness, not as substitutes for full expert review of course quality.

6.7 Failure Analysis

Figure 3 decomposes wasted top-5 acquisition slots using the hard-negative types from benchmark construction: partial-gap waste, topical redundancy, and spurious surface matches.

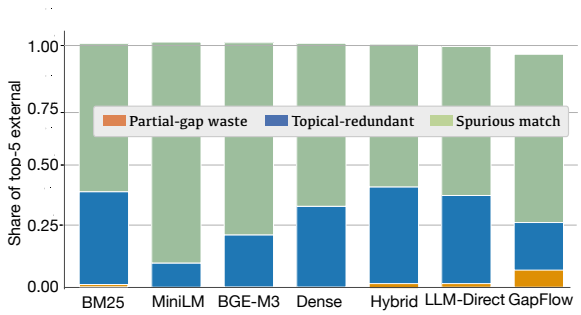

Figure 3: Top-5 external acquisition failure modes on CGF-MOOCCubeX-CN. Smaller bars indicate fewer wasted acquisition slots.

Retrieval and LLM-prompting baselines mainly fail through topical redundancy and spurious matches: they select resources related to the goal or gap terms without adding the missing competency coverage. The coverage-aware reference changes the error mix but still produces spurious matches, showing that external text-to-competency coverage estimation remains difficult. Thus, the benchmark separates acquisition failures into distinct error types rather than treating all mistakes as generic ranking errors. Per-domain and source-wise breakdowns in Appendix I.6 show that difficulty also depends on the availability and type of gap-covering resources in the external pool.

Limitations

The benchmark assumes verified internal course–competency links; programs with only raw course text require an additional concept-extraction step before catalog completion. Gold labels are bounded by the visible resource text and source annotation quality, even after text-inferability filtering. The external pool is source-imbalanced: in CGF-MOOCCubeX-CN, most S2 positives come from out-of-catalog MOOCCubeX courses, while TutorialBank and LectureBank contribute few positives. The English LectureBank set checks acquisition-protocol portability, but does not prove distribution-invariant performance across all subtasks. Extending the pool to cross-institution offerings or commercial textbooks may change the acquirable/design-only split. Because GapFlow uses a separate full-corpus index, it should be read as a diagnostic reference rather than a same-pool leaderboard competitor. Finally, S3 automatic metrics are diagnostic proxies: they test label-level coverage, redundancy, and prerequisite compactness, but do not replace full expert review of pedagogical quality. The human evaluation covers 50 design-required scenarios and three generative baselines, and does not measure long-term learner outcomes.

7 Conclusion

We introduced Curriculum Gap Filling, a task and MOOC-based benchmark for completing an incomplete course catalog across three linked decisions: internal gap discovery, non-redundant external acquisition, and missing-course design. The information asymmetry between verified internal links and text-only external resources separates catalog audit from inductive coverage inference. The English LectureBank set provides an acquisition-focused transfer check, with weak relevance-only retrieval and a high oracle ceiling under the same construction protocol. Experiments show that high relevance scores do not imply marginal curriculum coverage, that external text-to-competency coverage remains the main acquisition bottleneck, and that better acquisition does not by itself solve compact course design. GapFlow serves as an inspectable diagnostic baseline that decomposes these errors across gap discovery, coverage estimation, redundancy, and prerequisite handling.

References

Hasan Abu-Rasheed, Constance Jumbo, Rashed Al Amin, Christian Weber, Veit Wiese, Roman Obermaisser, and Madjid Fathi. 2025. LLM-assisted knowledge graph completion for curriculum and domain modelling in personalized higher education recommendations. In *IEEE Global Engineering Education Conference*.

Hasan Abu-Rasheed, Christian Weber, and Madjid Fathi. 2024. Knowledge graphs as context sources for llm-based explanations of learning recommendations. *arXiv preprint arXiv:2403.03008*.

Thijmen Bijl, Niels van Weeren, and Suzan Verberne. 2024. Efficient course recommendations with t5-based ranking and summarization. *arXiv preprint arXiv:2406.19018*.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335. Association for Computational Linguistics.

Xianyu Chen, Jian Shen, Wei Xia, Jiarui Jin, Yakun Song, Weinan Zhang, Weiwen Liu, Menghui Zhu, Ruiming Tang, Kai Dong, Dingyin Xia, and Yong Yu. 2023. Set-to-sequence ranking-based concept-aware learning path recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5027–5035.

Xinghe Cheng, Zihan Zhang, Jiapu Wang, Liangda Fang, Chaobo He, Quanlong Guan, Shirui Pan, and Weiqi Luo. 2026. GraphRAG-induced dual knowledge structure graphs for personalized learning path recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(17):14610–14620.

Alexander Fabbri, Irene Li, Prawat Trairatvorakul, Yijiao He, Weitai Ting, Robert Tung, Caitlin Westerfield, and Dragomir Radev. 2018. TutorialBank: A manually-collected corpus for prerequisite chains, survey extraction and resource recommendation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 611–620, Melbourne, Australia. Association for Computational Linguistics.

Bo Hou and 1 others. 2025. Kg-plppm: A knowledge graph-based personal learning path planning method used in online learning. *Electronics*, 14(2):255.

International Engineering Alliance. 2021. Graduate attributes and professional competencies. `https://www.internationalengineeringalliance.org/about/documents`. Version 4, September 2021. Accessed 2026-05-26.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Alireza Javadian Sabet, Sarah H. Bana, Renzhe Yu, and Morgan R. Frank. 2024. Course-skill atlas: A national longitudinal dataset of skills taught in u.s. higher education curricula. *Scientific Data*, 11(1086).

Irene Li, Alexander R. Fabbri, Robert R. Tung, and Dragomir R. Radev. 2019. What should i learn first: Introducing LectureBank for NLP education and prerequisite chain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6674–6681.

Jiawei Li, Qianru Lyu, Wei Qiu, and Andy W. H. Khong. 2025. Improving course recommendation systems with explainable AI: LLM-based frameworks and evaluations. In *Proceedings of the 18th International Conference on Educational Data Mining*, pages 205–214.

Boxuan Ma, Md Akib Zabed Khan, Tianyuan Yang, Agoritsa Polyzou, and Shin'ichi Konomi. 2025. How good are large language models for course recommendation in MOOCs? *Preprint*, arXiv:2504.08208.

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*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Elena Senger and 1 others. 2024. A survey on skill extraction and classification from job postings. In *Proceedings of NLP4HR*.

Shuanghong Shen and Qi Liu. 2024. A survey of knowledge tracing: Models, variants, and applications. *IEEE Transactions on Learning Technologies*.

Jiapu Wang and 1 others. 2026. Ib-grpo: Aligning llm-based learning path recommendation with educational objectives. *arXiv preprint arXiv:2601.14686*.

Zhen Xu, Xinjin Li, Yingqi Huan, Veronica Minaya, and Renzhe Yu. 2025. From course to skill: Evaluating large language model performance in curricular analytics. In *Proceedings of Artificial Intelligence in Education*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jifan Yu, Gan Luo, Tong Xiao, Qingyang Zhong, Yuquan Wang, Wenzheng Feng, Junyi Luo, Chenyu Wang, Lei Hou, Juanzi Li, Zhiyuan Liu, and Jie Tang. 2020. MOOCCube: A large-scale data repository for

NLP applications in MOOCs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3135–3142. Association for Computational Linguistics.

Jifan Yu, Yuquan Wang, Qingyang Zhong, Gan Luo, Yiming Mao, Kai Sun, Wenzheng Feng, Wei Xu, Shulin Cao, Kaisheng Zeng, Zijun Yao, Lei Hou, Yankai Lin, Peng Li, Jie Zhou, Bin Xu, Juanzi Li, Jie Tang, and Maosong Sun. 2021. MOOCCubeX: A large knowledge-centered repository for adaptive learning in MOOCs. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pages 4643–4652. ACM.

Nazar Zaki and 1 others. 2023. Automating the mapping of course learning outcomes to program learning outcomes using natural language processing for accurate educational program evaluation. *Education and Information Technologies*.

Mike Zhang, Kristian Nørgaard Jensen, Sif Dam Sonniks, and Barbara Plank. 2022. SkillSpan: Hard and soft skill extraction from english job postings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

A GapFlow Implementation Details

Course–competency verification. GapFlow applies different verification strategies reflecting the transductive/inductive split in Section 3. For an *internal course* $c \in \mathcal{C}$, the verifier uses known course–competency graph links in $\mathcal{K}$ directly, supplemented by embedding-based similarity for competencies without explicit links. For an *external resource* $r \in \mathcal{R}$, no pre-labeled mappings are available; the verifier infers coverage from text alone using normalized string matching, alias lookup, and embedding-based similarity. The verifier produces a normalized coverage score $\hat{v}(d,k) \in [0,1]$ and an uncertainty score $u(d,k) \in [0,1]$ derived from the margin between matched and unmatched competency evidence.

Coverage aggregation. For each target competency $k \in K_g$, GapFlow computes an internal sufficiency score:

$$\begin{aligned} S_{\mathcal{C}}(k) = w_{\max} \max_{c \in \mathcal{C}} \hat{v}(c,k) \\ + w_{\text{cum}} \min\Big(1, \textstyle\sum_{c \in \mathcal{C}} \hat{v}(c,k)\Big) \\ + w_{\text{match}} m(k) - w_{\text{unc}} \bar{u}(k), \quad (1) \end{aligned}$$

where $m(k)$ is a concept-matching feature and $\bar{u}(k) = |\mathcal{C}|^{-1} \sum_{c \in \mathcal{C}} u(c,k)$. A competency is predicted as a gap when $S_{\mathcal{C}}(k) < \tau_{\text{gap}}$, selected on a held-out development split.

External acquisition. Given the predicted gap set G, GapFlow scores each resource r:

$$s(r) = \text{GapCov}(r) - \beta\, \text{Red}(r) - \eta\, \text{Unc}(r),$$

where $\text{GapCov}(r)$ measures coverage of predicted unresolved gaps, $\text{Red}(r)$ penalizes overlap with competencies already covered by the internal catalog, $\text{Unc}(r)$ discounts uncertain coverage judgments, and $\beta, \eta \geq 0$ are set on a held-out development split.

Course design. After acquisition, GapFlow constructs a structured missing-course plan using the top-K resources from Subtask 2 as design references. It separates competencies into three roles: unresolved gaps to teach, existing catalog competencies to avoid re-teaching, and prerequisite competencies to assume or introduce. The implementation materializes this plan directly as a templated course-design record, preserving target competencies, assumed prerequisites, syllabus units, assignments, and relation to existing courses. No language model is used inside GapFlow.

B Benchmark Example

Table 6 shows a simplified scenario illustrating how controlled holdout creates synthetic gaps and how prerequisite targets are attached.

C Construction Details

C.1 Default Parameters

C.2 Course and concept filtering

We retain only courses with valid titles and at least m_c linked concepts. We retain only concepts linked to at least m_k courses within the same domain. Domains that do not meet the minimum course and concept counts are excluded. Duplicate resources are merged by canonical URL and normalized title.

C.3 Domain selection

The processed run contains ten fields: Computer Science and Technology, Mathematics, History, Foreign Languages and Literature, Arts, Physics, Clinical Medicine, Applied Economics, Psychology, and Biology. Domain selection was verified by a 3-seed dry-run requiring every domain to pass the four-axis joint gate (`acq_frac`≥ 0.45, `gap_cov_at_5`≥ 0.5, `scn_with_acq_gap_pct`≥ 0.65, `zero_cov_pct`≤ 0.3) on all three seeds.

C.4 External learning-resource mapping

OOC course candidates. Gold competency coverage on OOC MOOCCubeX courses is derived from verified concept-identifier links and then filtered through a text-inferability audit: three annotators independently judge each unique (course, concept) pair for whether visible resource text substantively supports the coverage claim; only majority-retained pairs enter the S2 gold set (Table 10).

Auxiliary course content. MOOCCubeX provides a much larger pool of non-course learning materials: course-content fragments, video-segment transcripts, and problem passages. These share the same concept-identifier annotation graph as OOC courses. A 60-pair dev audit finds course-content fragments at $\geq 95\%$ link precision, while video-segment annotations stand at 45% (9/20) and problem-passage annotations at 60% (12/20). We therefore gate the noisier sources by requiring both an annotation link and explicit text evidence (concept canonical name or alias appearing in the resource's title, description, outline, or transcript) before counting a (resource, gap) pair as a gold positive. Other sources keep an annotation-or-text-fallback rule.

TutorialBank and LectureBank. English-language tutorial and lecture passages from TutorialBank and LectureBank are mapped to MOOCCubeX competencies by Unicode-normalized canonical-name/alias substring match.

zhwiki articles. Chinese Wikipedia articles supplement the external pool for competencies not covered by any OOC course. We match MOOCCubeX competency names to zhwiki article titles and registered redirects after Unicode NFC normalization. Table 8 summarizes pool composition; Table 9 reports zhwiki-specific diagnostics.

Text-inferability audit of OOC gold labels. Three annotators independently judged all 276 unique (course, concept) pairs, assigning **1** (text supports coverage) or **0** (not inferable). Of 276 audited pairs, **100 are retained** (36.2%). Table 10 reports per-annotator statistics and pairwise Cohen's κ; Table 11 shows worked audit examples.

C.5 Human-Audited Prerequisite Targets

For each design-needed scenario x, auditors review candidate prerequisite triples (x, g, p) using the held-out source course's prerequisite text as primary evidence, assigning one of four labels: *valid assumed prerequisite*, *valid but should be taught*, *generic background*, or *not a prerequisite*. Only *valid assumed prerequisite* triples enter $P_{\text{gold}}(x)$, capped at three.

D Query Composition Ablation

English-only queries contribute essentially no retrieval signal, Chinese concept names alone reach the strongest recall, and mixing English with Chinese lowers recall. Low LLM acquisition scores therefore reflect the coverage-estimation challenge of the task rather than a Chinese/English adaptation artifact.

E Evaluation Details

E.1 Ranking

We use NDCG@K for both internal course ranking and external acquisition ranking.

E.2 Gap recovery

$$\text{Prec} = \frac{|P_{\text{gap}} \cap G_{\text{gap}}|}{\max(|P_{\text{gap}}|, 1)}.$$

Table 6: Simplified benchmark example from the Mathematics field.

Field	Example
Learner cohort	Undergraduate students with calculus and linear algebra background, pursuing a mathematics or statistics major.
Learning goal	Master the theory and application of probability and statistics.
Target competencies	Probability theory, random variables, statistical inference, hypothesis testing, regression analysis.
Initial catalog (before hold-out)	MOOCCubeX courses covering all five target competencies.
Held-out competencies	Hypothesis testing and regression analysis.
Catalog after holdout	Courses primarily covering hypothesis testing and regression analysis are removed; remaining courses still cover probability theory, random variables, and statistical inference.
Resulting gaps	Hypothesis testing and regression analysis are no longer covered internally.
External resources	OOC MOOCCubeX courses on hypothesis testing and regression analysis are high-value acquisitions; a descriptive-statistics course is treated as redundant; a broad probability survey is a spurious match for hypothesis testing. A Chinese Wikipedia article on hypothesis testing is included as an auxiliary candidate.
Missing-course design (if needed)	If no external resource covers a gap substantively, the specification targets the unresolved competencies, lists already-covered competencies to avoid re-teaching, and attaches human-audited prerequisites (e.g., single-variable calculus), capped at three.

$$\mathrm{Rec} = \frac{|P_{\mathrm{gap}} \cap G_{\mathrm{gap}}|}{\max(|G_{\mathrm{gap}}|, 1)}.$$

E.3 External acquisition

Let T_K be the top-K acquired resources, $K(r)$ the competencies covered by resource r, and $K(\mathcal{C}) = \bigcup_{c \in \mathcal{C}} K(c)$:

$$\text{MGC-acq@K} = \frac{\left|G_{\mathrm{gap}}^{\mathrm{acq}} \cap \bigcup_{r \in T_K} K(r)\right|}{\max(|G_{\mathrm{gap}}^{\mathrm{acq}}|, 1)}.$$

$$\text{MGC-all@K} = \frac{\left|G_{\mathrm{gap}} \cap \bigcup_{r \in T_K} K(r)\right|}{\max(|G_{\mathrm{gap}}|, 1)}.$$

$$\text{RR@K} = \frac{\left|K(\mathcal{C}) \cap \bigcup_{r \in T_K} K(r)\right|}{\max\left(\left|\bigcup_{r \in T_K} K(r)\right|, 1\right)}.$$

$$\text{MU@K} = \frac{\left|\bigcup_{r \in T_K} K(r) \setminus K(\mathcal{C})\right|}{K}.$$

E.4 Course design

$$\mathrm{CC} = \frac{|K_{\mathrm{gen}}^{\mathrm{content}} \cap G_{\mathrm{unres}}|}{\max(|G_{\mathrm{unres}}|, 1)}.$$

$$\mathrm{DRed} = \frac{|K_{\mathrm{gen}}^{\mathrm{content}} \cap A_{\mathrm{avoid}}|}{\max\left(|K_{\mathrm{gen}}^{\mathrm{content}}|, 1\right)}.$$

E.5 Prerequisite-sensitive metrics

$$\mathrm{POLR} = \frac{|P_{\mathrm{gen}} \setminus P_{\mathrm{gold}}(x)|}{\max(|P_{\mathrm{gen}}|, 1)}.$$

$$\mathrm{PRecall} = \frac{|P_{\mathrm{gen}} \cap P_{\mathrm{gold}}(x)|}{\max(|P_{\mathrm{gold}}(x)|, 1)}.$$

E.6 End-to-end metrics

$$\mathrm{FGR} = \frac{|G_{\mathrm{gap}} \cap K_{\mathrm{final}}|}{\max(|G_{\mathrm{gap}}|, 1)}.$$

$$\mathrm{FCR} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\mathrm{complete}_i = 1 \wedge G_{\mathrm{gap},i} \nsubseteq K_{\mathrm{final},i}].$$

$$\mathrm{PCC} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[G_{\mathrm{gap},i} \subseteq K_{\mathrm{final},i} \wedge \mathrm{POLR}_i = 0].$$

F Prompt Templates

Direct prompt. Given a cohort profile, learning goal, incomplete internal curriculum (with concept-graph links), concept taxonomy, and external resource corpus (text only, no concept labels), complete the curriculum. Return internal course recommendations, remaining gaps, external resources to acquire, and missing-course designs for unresolved gaps.

Stepwise prompts. S1: rank useful internal courses and identify remaining gaps. S2: rank external resources for acquisition over predicted gaps, avoiding redundant coverage. S3: design a missing course covering unresolved gaps with a structured proposal (title, target competencies, assumed prerequisites, syllabus units, assignments, relation to existing courses and reference resources).

+Taxonomy / +IntGraph / +Closure. Each variant runs all three subtasks in a single pass with additional context prepended: the competency taxonomy (+Taxonomy), the visible internal course–competency graph (+IntGraph), or the transitive coverage closure of the taxonomy (+Closure).

Table 7: Default construction and evaluation parameters.

Component	Parameter	Value
Scenario construction	Domain source	MOOCCubeX course fields selected by scenario availability
	External learning-resource sources	MOOCCubeX-OOC courses (3,781; concept-ids gold), MOOCCubeX video-segment transcripts (31,544) and problem passages (16,853) with text-evidence gating, MOOCCubeX other course-content fragments (64,858), TutorialBank lecture passages (20,139), LectureBank lecture passages (408), zhwiki articles (128; title/redirect match)
	Number of domains	10
	Scenarios per domain	50
	Internal courses per scenario	8
	Gap size distribution	45% size-1 / 40% size-2 / 15% size-3
	Design-needed criterion	top-K (K=5) external resources do not fully resolve all held-out gaps
	System-visible features ($\mathcal{C}$)	text + verified concept-graph links
	System-visible features ($\mathcal{R}$, OOC courses)	title, description, syllabus outline, prerequisites (where available)
	System-visible features ($\mathcal{R}$, zhwiki)	title, lead section
Prerequisite construction	Candidate sources	source-file local edges, local ancestors, direct text-derived links
	Audit evidence (primary)	held-out source course prerequisite text
	Audit cap	3 hard prerequisites per design-needed scenario
External acquisition	Candidate budget	100 (BM25) + 100 (mContriever) + 100 (concept graph), merged
	OOC leakage filter	exclude $\mathcal{C}$, $h(x)$, near-duplicates (title overlap & concept Jaccard)
Evaluation	Main cutoff	$K = 5$

G Human Annotation Protocols

G.1 Prerequisite candidate audit

For each design-needed scenario x, the auditor reviews candidate prerequisite triples (x, g, p) and assigns one of four labels: *valid assumed prerequisite*, *valid but should be taught*, *generic background*, or *not a prerequisite*. Only *valid assumed prerequisite* triples enter $P_{\text{gold}}(x)$, capped at three.

G.2 Generated course evaluation

Generated missing-course proposals are evaluated on a 1–5 scale using the rubric in Table 15.

H English Dataset: Full Results

The English dataset contains 100 scenarios constructed from LectureBank (Li et al., 2019) following the holdout procedure of Section 5. LectureBank's lecture-level topic–prerequisite graphs are aggregated into instructor-grouped courses (minimum 5 lectures and 3 distinct topics per course); scenario construction, gap holdout, joint-gate filtering, and gold-label synthesis proceed unchanged. The dataset draws on LectureBank's CS and NLP topic graphs and uses no Chinese resources.

I Additional Results

This appendix collects diagnostics referenced in Section 6.

I.1 Verifier Diagnostic

To factorise the S2 bottleneck, we separate verifier quality from positive exposure. A verifier predicts a binary coverage decision from resource text and a gap competency. We measure verifier generalisation on a leakage-free test pair set with disjoint train/dev/test *resources* and *concepts*; this controls for the near-ceiling pairwise F1 obtained on the original by-scenario split, which is traceable to resource and concept overlap with the train set. Positive exposure is measured independently as candidate Recall@K from each verifier's prefilter.

The supervised TF–IDF verifier improves substantially over the BM25-as-verifier baseline on the clean split, yet both pipelines pulling from BM25 top-100 collapse to negligible MGC-acq@5: any pipeline is exposure-bounded regardless of pairwise F1. GapFlow's native phrase-anchored verifier supplies full candidate recall. Pairwise coverage F1 alone therefore does not predict pipeline utility on this benchmark.

Table 8: External pool composition. **Visible fields**: provided to systems at inference. **Gold-only fields**: used for constructing gold labels only. **Pos. contrib.**: share of total S2 gold positive (resource, gap) rows on the Chinese 500-scenario set after text-evidence gating. **S2-eval. scen.**: number of scenarios with at least one gold positive from the source.

Source	#Res.	Type	Visible fields	Gold-only fields	Pos. contrib.	S2-eval. scen.
MOOCCubeX-OOC	3,781	Course	title, description, outline, prereqs[†]	concept_ids	54.3%	378
MOOCCubeX-other	64,858	Fragment	title, text	concept_ids	28.8%	270
MOOCCubeX-video[‡]	31,544	Video transcript	title, transcript	concept_ids (gated)	12.1%	128
MOOCCubeX-problem[‡]	16,853	Problem passage	title, text	concept_ids (gated)	3.3%	60
zhwiki (aux.)	128	Article	title, lead section	title/redirect match	1.3%	33
LectureBank	408	Lecture	title, text	substring match	0.2%	4
TutorialBank	20,139	Lecture	title, text	substring match	<0.1%	1
Total / union	137,711	–	–	–	100%	477

[†]Prerequisites available for 31.8% of OOC courses. [‡]Annotation links require co-occurring text evidence to enter the gold set.

Table 9: zhwiki concept-matching diagnostics per domain. Of the 128 zhwiki articles in the external corpus, 84 link to a scenario concept.

Field	#Articles	#Concepts	Title%
CS&T	20	88	22.7%
Math	21	84	25.0%
History	17	70	24.3%
Foreign Lang.	14	81	17.3%
Arts	12	87	13.8%
Physics	0	68	0.0%
Med.	0	84	0.0%
Econ.	0	75	0.0%
Psy.	0	60	0.0%
Bio.	0	104	0.0%
Overall	84	780	10.8%

Table 10: Inter-annotator agreement for the OOC gold text-inferability audit (276 pairs).

	Pos.	Sole dissent	κ vs. others
Annotator 1	37.0%	8	0.762 / 0.658
Annotator 2	44.9%	24	0.762 / 0.565
Annotator 3	25.0%	33	0.658 / 0.565
3-way full agreement: 211/276 (76.4%)			
Majority-vote retain: 100/276 (36.2%)			

I.2 Pool-Size Sweep

Table 18 sweeps the hybrid candidate-pool size $K \in \{10, 30, 50, 100, 200, 500\}$ on the 477 scenarios with ≥ 1 gold-positive resource. Pool-Restricted GAPFLOW replaces GAPFLOW's token-matched candidate set with the hybrid top-K; everything else (marginal coverage, redundancy, uncertainty) is unchanged. We report non-LLM pool diagnostics only.

Three patterns: hybrid reachability rises with K but plateaus well below saturation, leaving an irreducible retrieval shortfall; the pool-oracle ceiling rises sharply with K while Pool-Restricted GAPFLOW barely moves, attributing the widening gap to coverage verification rather than candidate generation.

I.3 Source-wise Acquisition

Annotation precision varies across the external pool. Table 19 reports MGC-acq@5 restricted to gold positives from each source, on the subset of scenarios with at least one gold positive from that source.

The benchmark separates source-precision regimes. On the highest-precision OOC course source, the LLM reranker exploits long descriptions effectively; on noisier or sparser sources, the text-evidence verifier aligns with the gating rule used to construct gold labels there. The oracle row shows a residual retrieval gap on OOC courses that no in-pool verifier can recover.

I.4 Construction Sensitivity

We sweep two construction choices behind the design-only share reported in Section 5: the top-K cutoff used by the gold greedy ranking that defines acquirable coverage, and the external pool composition. Table 20 reports pool-sensitivity at K=5; Table 21 reports K-sensitivity at the paper pool. Predictions are held fixed across cells: saved top-100 candidates are post-hoc filtered to each pool and re-scored.

The design-only share rises substantially when the pool drops the fine-grained MOOCCubeX video/problem fragments, and is invariant to K in the swept range. GAPFLOW's FCR stays high across all pool variants. Qwen3-8B Direct's mea-

Table 11: Example audit decisions on (course, concept) gold pairs. **Retained**: visible text substantively supports the coverage claim. **Discarded**: concept is mentioned only in passing or absent from teaching content.

Course title	Annotated concept	Audit decision (rationale)
(RISC-V) (*Operating Systems, RISC-V*)	(*interrupt mechanism*)	Retained: an operating-systems course explicitly covers interrupt handling, matching the annotated concept.
C++-91 (*C++ Programming*)	(*pointer types*)	Retained: pointer types are a core syllabus topic in any C++ course; the title and standard course outline support the coverage claim.
C++-91 (*C++ Programming*)	(*screen*)	Discarded: the annotated concept is a generic I/O term that appears at most as a passing reference, not as substantive teaching content.
C++-91 (*C++ Programming*)	(*option/menu item*)	Discarded: the concept is generic UI vocabulary unrelated to the C++ programming syllabus; annotation is spurious.
(RISC-V) (*Operating Systems, RISC-V*)	(*mouse click*)	Discarded: the annotation links a desktop-interaction concept that the visible course text does not teach in any substantive sense.

Table 12: Capped human-audited prerequisites by source field. **Rich**: $\geq$2 validated prerequisites.

Source field	Design	Mean	Median	P95	Max	Rich	Rich %
History	20	1.25	1.0	2.0	2	5	25.0
Foreign Languages and Literature	21	3.00	3.0	3.0	3	21	100.0
Mathematics	35	1.74	2.0	3.0	3	18	51.4
Arts	20	0.95	1.0	2.0	2	3	15.0
Computer Science and Technology	14	2.21	3.0	3.0	3	9	64.3
Overall	110	1.81	2.0	3.0	3	56	50.9

Table 13: Examples of audited hard prerequisite targets.

Missing-course topic	Audited prerequisites
Chinese Ancient Literature	Ancient Chinese
Academic English Communication	English vocabulary; grammar; expression
Advanced Modern Life Aesthetics	Modern Life Aesthetics
Probability and Mathematical Statistics	Single-variable calculus
Linux Kernel Analysis and Applications	Linux basics; C language; OS basics

Table 14: BM25 Recall@K under five query compositions (retriever fixed).

Query composition	R@10	R@30	R@50	R@100
EN: goal + cohort	0.002	0.002	0.002	0.116
Opaque IDs	0.000	0.142	0.142	0.142
ZH: gap names	**0.429**	**0.578**	**0.643**	**0.726**
ZH: gap names + aliases	0.429	0.578	0.643	0.726
EN + ZH: gap names + aliases	0.302	0.537	0.599	0.673

sured FGR shifts only marginally because most of its completion is routed through generated design content; the high FGR is consistent with low PCC across all configurations. Even with GAPFLOW+Oracle Coverage, PCC stays at or below 0.7 in every cell, reinforcing prerequisite-compact design as a separate axis.

I.5 Diagnostic Component Controls

Table 22 reports component-control diagnostics that isolate graph access, text fallback, set-cover ranking, and marginal-utility scoring inside GAPFLOW.

I.6 Per-Domain Breakdown

I.7 S2 Acquisition Gap Decomposition

I.8 Forced-Design Control

J Error Examples

False synthetic gap. A model predicts a competency as missing even though it remains covered by the incomplete internal catalog, indicating underestimation of internal coverage.

Missed gap. A model fails to predict a held-out competency as missing because it overestimates coverage from a weakly related internal course.

Redundant acquisition. A model recommends an external resource related to the learning goal while mostly overlapping with competencies already covered internally.

Prerequisite mismatch. A model lists a prerequisite without scenario validation and outside the ancestor set of the generated target competencies.

Table 15: Human evaluation rubric for generated missing-course designs.

Dimension	Description
Competency completeness (CC-H)	The proposal addresses all unresolved gap competencies.
Prerequisite appropriateness (PA)	Listed prerequisites are reasonable, appropriately scoped, and avoid merely generic background.
Pedagogical feasibility (PF)	The course could realistically be taught given the cohort background and implied course length.
Catalog integration (CI)	The proposal forms a coherent sequence with existing courses, assumes the right prior knowledge, and avoids duplicating existing content.

Table 16: English dataset (100 scenarios). 95% paired-bootstrap CI in brackets (1000 iterations, seed 20260520). †: significantly different from GapFlow at 95% on MGC-acq@5.

Method	**MGC-acq@5**	**MGC-all@5**	**End-to-end**
Random	0.000 [0.000, 0.000]†	0.000 [0.000, 0.000]	0.310
BM25	0.052 [0.015, 0.100]†	0.037 [0.012, 0.068]	0.309
MiniLM	0.050 [0.013, 0.097]†	0.032 [0.008, 0.057]	0.313
BGE-M3	0.000 [0.000, 0.000]†	0.000 [0.000, 0.000]	0.305
mContriever	0.005 [0.000, 0.015]†	0.003 [0.000, 0.010]	0.305
Hybrid	0.180 [0.120, 0.252]†	0.133 [0.090, 0.182]	0.357
LLM Direct	0.400 [0.310, 0.487]†	0.275 [0.217, 0.335]	0.384
LLM Stepwise	0.020 [0.000, 0.045]†	0.015 [0.000, 0.035]	0.166
LLM Reranker	0.037 [0.007, 0.073]†	0.027 [0.005, 0.052]	0.163
GapFlow	**0.637** [0.548, 0.715]	**0.437** [0.380, 0.495]	**0.652**
w/o coverage verifier	0.000 [0.000, 0.000]†	0.000 [0.000, 0.000]	0.304
w/o redundancy penalty	0.665 [0.578, 0.745]	0.450 [0.395, 0.508]	0.601
w/o prerequisite graph	0.637 [0.548, 0.715]	0.437 [0.380, 0.495]	0.652
w/o structured design	0.637 [0.548, 0.715]	0.437 [0.380, 0.495]	0.638
GapFlow+Oracle Coverage	1.000 [1.000, 1.000]†	0.692 [0.645, 0.738]	0.910

Table 17: Coverage-verification diagnostic on CGF-MOOCCubeX-CN, restricted to scenarios with ≥ 1 gold positive (477 of 500). *Pairwise*: verifier coverage F1 (**Cov-F1**) on the leakage-free pair set. *Pipeline*: MGC-acq@5, RR@5, and FCR obtained by substituting each scorer into GF; BM25 and Sup. TF-IDF pull candidates from BM25 top-100 (**Cand-Rec**: candidate recall@100). Pipeline MGC and RR use the same gap normalization as Table 1. The Sup. TF-IDF pipeline uses the original-split (leaky) classifier; clean-split Cov-F1 measures generalisation separately.

	Pairwise	*Pipeline*			
Verifier	Cov-F1	Cand-Rec	MGC	RR↓	FCR↓
GF-text	–	1.000	0.155	0.018	0.773
BM25	0.305	0.128	0.009	0.080	0.988
Sup. TF-IDF	0.739	0.128	0.005	0.076	0.992
GF+OC	1.000	1.000	1.000	0.176	0.000

Underspecified course design. A generated course has a plausible title and syllabus while leaving unresolved gaps uncovered or omitting prerequisite structure.

K Artifact, Annotation, and Reproducibility Details

K.1 Artifact Licenses and Access Conditions

We use both existing artifacts and artifacts created in this work. Existing artifacts include MOOCCubeX, TutorialBank, LectureBank, Chinese Wikipedia, retrieval models, and LLM backbones. We cite the original creators of all third-party artifacts in the main paper and references.

MOOCCubeX is used as the main source of course metadata, course–concept links, concept graphs, video/problem/resource metadata, and external MOOC resources. We use the educational-resource and concept-link portions needed for benchmark construction. We do not use or release learner-level behavior records, raw user profiles, or any personally identifying learner information. Users of the released benchmark should obtain MOOCCubeX from the original maintainers and comply with its access conditions and license.

TutorialBank and LectureBank are used as public educational-resource metadata and topic/prerequisite resources. We use resource titles,

Table 18: Candidate-pool sensitivity on CGF-MOOCCubeX-CN, 477 scenarios with ≥ 1 gold positive. MGC values use the Table 1 gap normalization. **Oracle**: greedy in-pool MGC-acq@5 ceiling. **Pool-R. GapFlow**: Pool-Restricted GapFlow.

	Pool Recall@K				**MGC-acq@5**	
K	BM25-CJK	mContriever	Hybrid	**Reach%**	**Oracle**	**Pool-R. GapFlow**
10	0.088	0.027	0.194	42.8%	0.157	0.140
30	0.143	0.042	0.335	60.6%	0.232	0.158
50	0.174	0.048	0.392	68.1%	0.261	0.160
100	0.218	0.078	0.473	74.6%	0.292	0.164
200	0.278	0.103	0.543	79.7%	0.313	0.166
500	0.365	0.161	0.603	83.9%	0.335	0.167

Table 19: Source-wise MGC-acq@5 on CGF-MOOCCubeX-CN under the Table 1 gap normalization. *aux.* groups zhwiki, TutorialBank, and LectureBank.

Method	OOC	other	video	problem	aux.
BM25	0.015	0.020	0.025	0.007	0.022
Hybrid	0.030	0.037	0.054	0.068	0.055
Qwen3-8B (Direct)	0.037	0.083	0.054	0.082	0.094
Qwen3-8B (OracleRR)	0.127	0.146	0.145	0.150	0.176
GF	0.100	**0.196**	**0.255**	**0.184**	**0.198**
+OC	0.268	1.000	1.000	1.000	1.000

Table 20: Pool-sensitivity at K=5 (N=500), using the same gap normalization as Table 1.

Pool	%D-only	**Method**	FGR	FCR↓	PCC
OOC	52.4	GF	0.066	0.948	0.044
		Direct	0.292	0.350	0.171
		GF+OC	0.658	0.207	0.555
+Wiki	51.8	GF	0.075	0.930	0.051
		Direct	0.292	0.350	0.171
		GF+OC	0.666	0.197	0.562
no V/P	43.4	GF	0.128	0.822	0.085
		Direct	0.303	0.317	0.180
		GF+OC	0.743	0.060	0.658
Paper	40.0	GF	0.155	0.773	0.102
		Direct	0.306	0.307	0.182
		GF+OC	0.790	0.000	0.700

Table 21: K-sensitivity at the paper pool (N=500), using the same gap normalization as Table 1.

K	%D-only	**Method**	FGR	FCR↓	PCC
3	40.0	GF	0.135	0.806	0.091
		Direct	0.304	0.313	0.180
		GF+OC	0.790	0.000	0.700
5	40.0	GF	0.155	0.773	0.102
		Direct	0.306	0.307	0.182
		GF+OC	0.790	0.000	0.700
10	40.0	GF	0.191	0.703	0.127
		Direct	0.306	0.307	0.182
		GF+OC	0.790	0.000	0.700
20	40.0	GF	0.214	0.646	0.149
		Direct	0.306	0.307	0.182
		GF+OC	0.790	0.000	0.700

Table 22: Diagnostic component controls on CGF-MOOCCubeX-CN.

Variant	Gap-P	Gap-R	MGC	RR↓
Graph labels only	0.86	0.47	0.02	0.27
Graph labels + set cover	0.86	0.47	0.02	0.27
Text rank + graph gaps	0.86	0.47	0.08	0.24
Graph + text fallback	0.81	0.33	0.09	0.08
Graph planner w/o MU	0.49	0.75	0.00	0.13
GF	0.81	0.33	0.16	0.02
GF+OG	1.00	1.00	0.18	0.02
GF+OC	1.00	1.00	1.00	0.18

URLs, topic annotations, and short text fields needed for controlled benchmark construction and transfer evaluation. We preserve attribution to the original resources through their identifiers and URLs. When releasing our benchmark, we release derived scenario identifiers, labels, splits, and evaluation files, and refer users to the original repositories for the underlying source resources.

Chinese Wikipedia is used as an auxiliary public text source. We use article titles, redirects, and lead text for text-only external acquisition candidates. Wikipedia-derived text is subject to Wikimedia licensing terms, including attribution and share-alike requirements. Our released files will include page identifiers and derived benchmark labels; if article text is redistributed, it will be distributed with attribution and under compatible license terms.

For models and software, we use BM25 retrieval, mContriever, BGE-M3, Qwen3-8B, and Claude Haiku 4.5. Open-weight models are used according to their model licenses and model-card terms. Claude Haiku 4.5 is accessed only through the provider API; we do not redistribute model weights or proprietary model outputs as standalone re-

Table 23: Per-domain breakdown of MGC-acq@5 and FGR across the ten MOOCCubeX fields, using the same gap normalization as Table 1. Oracle Coverage is shown for MGC-acq@5; FGR oracle diagnostics are reported in aggregate in Table 4.

MGC-acq@5	CS	Math	Hist.	FL	Arts	Phys.	Med.	Econ.	Psy.	Bio.	Mean
BM25	0.000	0.018	0.013	0.000	0.042	0.053	0.017	0.000	0.017	0.000	0.016
MiniLM	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.009	0.000	0.001
BGE-M3	0.052	0.035	0.000	0.000	0.081	0.029	0.060	0.000	0.060	0.013	0.033
mContriever	0.009	0.013	0.004	0.000	0.055	0.000	0.009	0.000	0.038	0.000	0.013
Hybrid	0.061	0.013	0.008	0.009	0.042	0.008	0.034	0.000	0.106	0.000	0.028
GapFlow	**0.213**	**0.187**	**0.217**	**0.204**	**0.102**	**0.155**	**0.140**	**0.059**	**0.196**	**0.069**	**0.155**
GapFlow+Oracle Coverage	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
FGR	CS	Math	Hist.	FL	Arts	Phys.	Med.	Econ.	Psy.	Bio.	Mean
GapFlow	0.185	0.144	0.230	0.221	0.129	0.135	0.211	0.067	0.180	0.053	0.155
Qwen3-8B (Direct)	**0.306**	0.240	0.218	**0.272**	**0.282**	**0.370**	**0.366**	**0.343**	**0.339**	**0.325**	**0.306**
Qwen3-8B (Stepwise)	0.181	0.196	0.187	0.241	0.263	0.166	0.221	0.169	0.254	0.200	0.208
Qwen3-8B (OracleRR)	0.225	**0.246**	0.205	0.226	0.257	0.196	0.225	0.232	0.237	0.211	0.226
Qwen3-8B (+IntGraph)	0.244	0.244	**0.234**	0.242	0.204	0.132	0.146	0.188	0.196	0.152	0.198
Qwen3-8B (+Closure)	0.158	0.116	0.156	0.103	0.080	0.138	0.101	0.155	0.140	0.131	0.128

Table 24: Decomposition of the S2 acquisition gap into retrieval ceiling and in-pool selection. **MGC-acq@5 (all)**: 477 scenarios with ≥ 1 gold positive. **MGC-acq@5 (pool-reachable)**: the 325 of those scenarios whose Hybrid (3-way) top-50 contains ≥ 1 gold positive. A greedy pool-oracle picking 5 from the top-50 reaches MGC-acq@5=0.261 on all 477 scenarios under the Table 1 gap normalization, defining the in-pool ceiling.

Method	**MGC-acq@5 (all 477)**	**MGC-acq@5 (pool-reachable 325)**
BM25	0.016	0.024
Hybrid	0.028	0.038
GraphTextCoverage	0.091	0.128
Qwen3-8B (Direct)	0.051	0.074
Qwen3-8B (Stepwise)	0.065	0.089
Qwen3-8B (OracleRR)	0.116	0.124
Qwen3-8B (+IntGraph)	0.065	0.089
Qwen3-8B (+Closure)	0.036	0.048
GapFlow	**0.155**	**0.224**
GapFlow+Oracle Coverage	1.000	1.000
Pool-oracle (greedy)	0.261	0.383

sources. Our own benchmark construction scripts, evaluation code, and derived benchmark metadata will be released under the MIT License for code and CC BY 4.0 for derived benchmark metadata (scenario files, evaluation labels, and split definitions). Subsets that incorporate Chinese Wikipedia text are distributed under CC BY-SA 4.0 to maintain compatibility with Wikimedia licensing terms. We do not claim ownership over third-party resources.

K.2 Intended Use and Data Risk

The benchmark is intended for research on curriculum analytics, educational NLP, course-catalog auditing, resource acquisition, and diagnostic evaluation of coverage reasoning. It is not intended to make high-stakes decisions about individual learners, instructors, universities, or course providers. System outputs should be treated as decision support for curriculum analysis rather than as expert-certified curriculum recommendations.

The benchmark is constructed from course/resource metadata, concept annotations, public educational-resource text, and human-audited benchmark labels. We do not use raw learner behavior logs, learner profiles, grades, forum identities, or other learner-level personally identifying information. Some public educational resources may contain public course titles, instructor names, institutional names, URLs, or article titles inherited from the original source repositories. We preserve such fields only when needed for resource attribution or benchmark reproducibility.

We did not intentionally collect offensive or sensitive content. Because public web and educational

Table 25: Forced-design control on Subtask 3. Every method receives the gold gap set and gold reference resources and must return a non-empty design over 200 design-required scenarios. Claude Haiku 4.5+Closure is evaluated only in this regime.

Method	CC	DRed↓
Qwen3-8B (Direct)	0.757	0.933
Qwen3-8B (Stepwise)	0.807	0.838
Qwen3-8B (OracleRR)	0.807	0.838
Qwen3-8B (+IntGraph)	0.807	0.838
Qwen3-8B (+Taxonomy)	0.863	0.886
Qwen3-8B (+Closure)	0.887	0.910
Claude Haiku 4.5 (+Closure)	**1.000**	0.979
GF	**1.000**	**0.136**
GF+OC	**1.000**	**0.136**

resources can contain unexpected content, we limit the benchmark to course/resource text, article leads, and educational metadata, and we exclude learner-generated personal records. Future users should apply additional filtering before deploying systems trained or evaluated on this benchmark in student-facing settings.

K.3 Human Annotation Protocol

Human annotation is used in two places: the text-inferability audit for external resource–competency coverage labels and the prerequisite audit for course-design targets.

For the text-inferability audit, annotators were shown a resource text field (e.g., title, description, syllabus, transcript snippet, or article lead) and a candidate competency. They were instructed to judge whether the visible text substantively supports the claim that the resource teaches or covers the competency. Annotators assigned a binary label: *covered* if the resource text provides substantive evidence of teaching the competency, and *not covered* if the competency is only mentioned in passing, appears only as background/prerequisite text, appears through surface lexical overlap, or is not inferable from the visible text. Majority agreement determines whether a pair is retained in the gold acquisition set.

For the prerequisite audit, an expert auditor reviewed scenario–gap–candidate prerequisite triples using the held-out source course's prerequisite text and local prerequisite graph evidence. The auditor assigned one of four labels: valid assumed prerequisite, valid but should be taught, generic background, or not a prerequisite. Only triples labeled as valid assumed prerequisites enter the gold prerequisite set, capped at three per scenario.

Annotators were recruited from graduate students in the authors' research laboratory. They were selected based on Chinese and English reading proficiency, familiarity with MOOC course descriptions, and background in computer science, natural language processing, or educational informatics. Annotation was conducted as part of research participation in the laboratory; no separate monetary compensation was provided. Annotators were informed that their judgments would be used to construct an anonymous research benchmark. No annotator names or personally identifying information are released.

K.4 Ethics Review

The annotation tasks involve judging educational-resource text and candidate competency/prerequisite labels. The study does not collect sensitive personal information from annotators, does not involve student participants, and does not use learner-level private data. Under our institution's policy, this work did not require formal IRB review. Annotators were informed of the research purpose of the task and no annotator-level personal data is included in the released benchmark.

K.5 Computational Budget and Implementation Details

All experiments are run with fixed random seeds where applicable. Sparse retrieval uses BM25; dense retrieval uses MiniLM, mContriever, and BGE-M3 embeddings. LLM experiments use Qwen3-8B variants for the main open-weight baselines, and Claude Haiku 4.5 for the isolated forced-design control. Hyperparameters for GAPFLOW, including gap threshold, redundancy weight, and uncertainty weight, are selected on a held-out development split and kept fixed for test evaluation.

Experiments were run on the KAUST IBEX cluster using NVIDIA V100 (32 GB) and A100 (80 GB) GPUs, one GPU per job. The approximate compute budget is 40 GPU-hours for dense retrieval embedding (mContriever and BGE-M3 over the 137k-resource external pool) and Qwen3-8B inference across all baseline variants, plus approximately 10 CPU-hours for BM25 indexing, benchmark construction, and evaluation. Qwen3-8B inference uses greedy decoding (temperature = 0, `enable_thinking = False`), `max_new_tokens` = 2048, and `float16` precision;

Claude Haiku 4.5 uses temperature = 0 and `max_tokens` = 2048 via the provider API. Retrieval and verifier models use PyTorch 2.6.0 (CUDA 12.4), Hugging Face Transformers 4.52.4, sentence-transformers 5.1.1, FAISS-cpu $\geq$ 1.8.0, scikit-learn 1.7.2, and rank-bm25 $\geq$ 0.2.2. We will release scripts for scenario construction, baseline execution, metric computation, bootstrap testing, and sensitivity analyses.

K.6 Use of AI Assistants

The authors used AI assistants during the writing and revision process for language editing, reviewer-style feedback, checklist preparation, and non-authoritative coding/debugging discussion. AI assistants were not used to make final research decisions, assign gold labels, fabricate results, or replace human audit. All task definitions, dataset construction choices, experimental protocols, numerical results, claims, and final manuscript text were reviewed, verified, and approved by the authors. No AI assistant is listed as an author.