

GroundEd: Curriculum-Grounded Small Language Models for Offline AIED

Anonymous Author(s)

Anonymous Institution

Abstract. Large language models (LLMs) offer increasingly capable educational assistance, yet their computational and connectivity requirements limit deployment in low-resource settings. Recent AIED Unplugged research has established deployment practices for optimized local models on constrained devices. Curriculum-grounded retrieval offers a complementary way to make those models contextually aware of the textbooks and curricular materials used in a school. However, it remains unclear how much model capacity is required when a small language model is grounded in such material and deployed entirely offline.

This paper presents *GroundEd*, a fully offline pipeline that converts educator-provided textbooks and curricular documents into a local retrieval index used by quantized small language models. Across multiple model scales, we compare ungrounded generation, naive retrieval-augmented generation (RAG), and curriculum-aware retrieval under a fixed low-cost hardware budget. In addition to general educational questions, the evaluation includes curriculum-dependent queries whose answers rely on specific, including deliberately obscure, information in the supplied materials. This design distinguishes successful use of retrieved curriculum evidence from answers based mainly on pre-existing parametric knowledge. We assess answer quality, faithfulness, pedagogical usefulness, out-of-scope behavior, latency, memory, storage, and energy use.

We therefore evaluate whether curriculum grounding can reduce the model capacity required for useful generative AIED while preserving fully offline operation on inexpensive, resource-constrained hardware.

Keywords: Artificial Intelligence in Education · Small Language Models · Retrieval-Augmented Generation · Offline AI · Curriculum Grounding · AIED Unplugged · Resource-Constrained Computing

1 Introduction

Generative language models support tutoring, explanation, feedback, and question answering, but most deployments assume cloud inference, reliable connectivity, and capable devices. These assumptions exclude many settings in which educational support is most needed.

AIED Unplugged argues that educational AI should conform to available infrastructure [4]. Work on pedagogical elasticity and offline-first applications extends this principle to local LLMs whose capabilities and model selection adapt to constrained devices [10, 11]; smartphone evidence shows that feasibility alone does not guarantee responsive use [9].

RAG offers a complementary strategy: retrieve local instructional evidence rather than requiring all knowledge parametrically [5]. Educational systems ground models in course materials and textbooks [6, 3], with curriculum-aware variants adding structured retrieval, pedagogical prompting, and answerability checks [2, 1, 8].

Together, these directions motivate the question: *How much language model do we actually need if the model only needs to teach a bounded curriculum?* GroundEd couples a local index of educator-provided materials with small quantized models and tests material-specific questions unlikely to be answered reliably from generic model knowledge alone. We evaluate model scale, grounding quality, pedagogical utility, and resource cost on Pi-class Linux hardware, treating hardware as the evaluation platform rather than the contribution.

Figure 1 summarizes GroundEd’s offline curriculum-preparation pipeline.

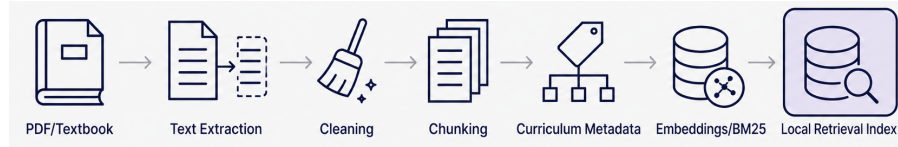


Fig. 1. GroundEd offline curriculum preparation pipeline. Curriculum materials are extracted, cleaned, segmented, annotated with curriculum metadata, and indexed locally using semantic and lexical retrieval representations.

2 Background and Related Work

2.1 AIED Unplugged and Resource-Constrained Educational AI

Isotani et al. define AIED Unplugged around scalable educational AI that conforms to underserved settings [4]. Uema et al. extend this agenda through *pedagogical elasticity*, adapting local-LLM capabilities to available compute [10], and show the importance of model–device matching across heterogeneous offline-first deployments [11].

Santos et al. quantify latency, memory, and storage barriers on constrained smartphones [9]. MathAide complements this systems evidence with teacher-perceived value for unplugged planning, feedback, analytics, grouping, and personalization [7], motivating joint evaluation of performance and educational usefulness.

2.2 Curriculum-Grounded Educational Language Models

Lewis et al. establish generation conditioned on retrieved evidence [5]. Hevia et al. apply this mechanism to an offline textbook tutor with a lightweight model, while showing that very small models can struggle with long or noisy context [3].

RAGMan grounds programming tutors in instructor and assignment materials; in a 455-student deployment, 78% of surveyed users reported learning support and within-scope answers were highly accurate [6], although the system assumes heavier infrastructure than GroundEd.

EduChatQA combines curriculum graphs, hybrid semantic-BM25 retrieval, teacher-aligned prompting, and unsupported-question refusal [2]. Adenuga finds stronger expert-rated conceptual support and instructional quality from curriculum RAG with learner state and pedagogical policies [1]. In low-resource bilingual schools, Rugemalila et al. identify a grounded 0.5B model as a realistic quality-deployability compromise [8].

2.3 Research Gap

Prior work establishes constrained LLM deployment, course tutors, curriculum-aware retrieval, and small-model trade-offs, but does not isolate **model scale** $\times$ **curriculum grounding** $\times$ **offline resource cost**. GroundEd tests whether retrieved evidence supports material-specific questions beyond generic parametric knowledge and thereby shifts the minimum viable model size for offline AIED.

3 GroundEd System

3.1 Design Goals

GroundEd targets fully offline operation, educator-controlled knowledge, low computational cost, and bounded responses with source traceability and abstention when curricular evidence is insufficient.

3.2 Offline Curriculum Ingestion

Educator-provided documents are extracted and cleaned while retaining provenance, segmented into **TODO: XX-XX-token** chunks with **TODO: XX-token overlap**, bound to subject, grade, unit, objective, and page metadata, and indexed locally using **TODO: dense/BM25/hybrid retrieval**.

3.3 Curriculum-Grounded Generation

For a query q , the retriever ranks curriculum passages d_i and selects the top- k evidence set

$$C(q) = \text{TopK}_{d_i \in D} s(q, d_i), \quad (1)$$

where D is the local curriculum corpus and $s(\cdot)$ is the retrieval score.

The generator receives the student query, retrieved curriculum evidence, source metadata, and a pedagogical instruction template. The prompt explicitly instructs the model to (1) prioritize supplied curriculum evidence, (2) avoid unsupported claims, (3) provide age/grade-appropriate explanations, and (4) abstain when the curriculum does not contain sufficient evidence.

Figure 2 summarizes the corresponding student-query and local inference pipeline.

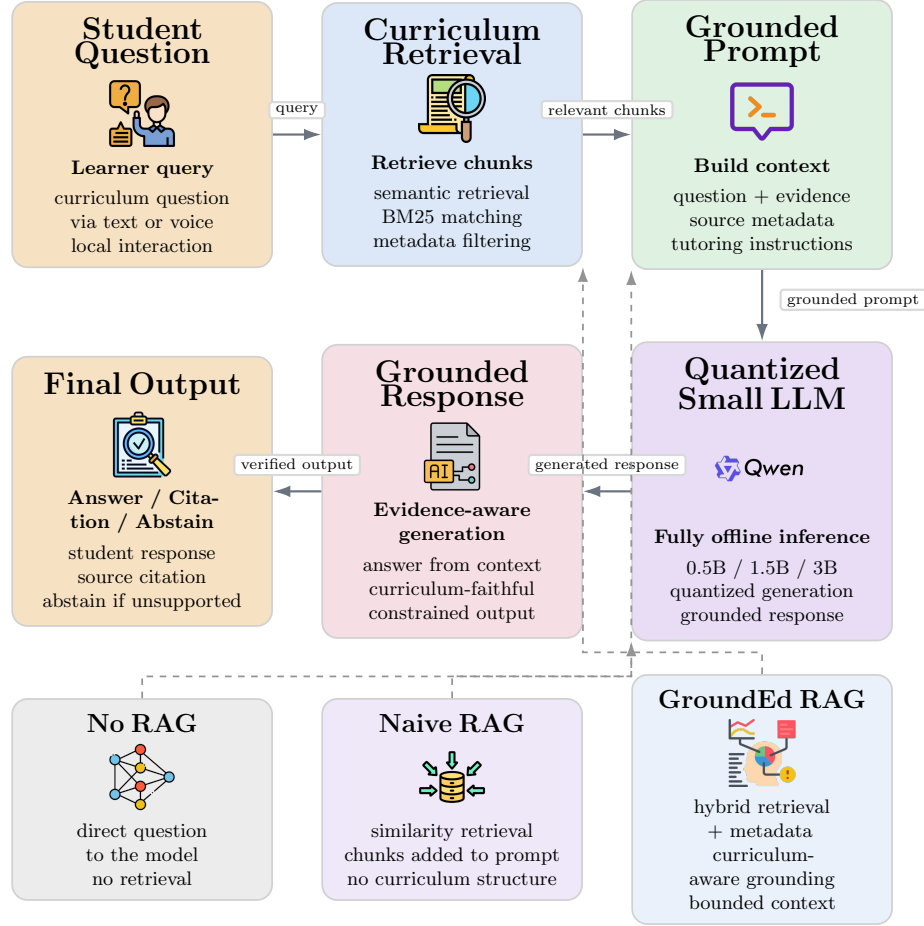


Fig. 2. GroundEd student query and inference pipeline. A student query is grounded using retrieved curriculum evidence before local generation, with No RAG, Naive RAG, and GroundEd RAG shown as the evaluated grounding conditions.

The same pedagogical prompt template is held constant across model conditions; the final specification will be reported with the experimental configuration.

3.4 Local Model Runtime and Hardware

Quantized models run locally using **TODO: runtime** on the Pi-class Linux platform summarized in Table 1; hardware is an experimental constraint, not the contribution.

4 Evaluation

4.1 Experimental Design

We use a factorial design varying **model scale** and **grounding strategy**. Every model receives the same questions under No RAG, Naive RAG, and GroundEd; the key asymmetric comparison pairs a smaller grounded model with a larger ungrounded model.

4.2 Curriculum Corpus

We use a bounded, openly licensed middle/high-school science curriculum. We report source documents, page and token counts, chunking configuration, embedding model, and index size.

4.3 Question and Task Set

The curriculum-aligned benchmark contains **TODO: N questions**, reviewed by a subject-matter or education expert. Items cover factual recall, conceptual explanation, application, scaffolding, misconception correction, deliberately obscure material-specific details, and out-of-curriculum queries.

4.4 Metrics

Educational metrics are correctness, evidence faithfulness, relevance, pedagogical appropriateness, hallucination, citation correctness, and abstention precision/recall. Systems metrics are time to first token, latency, tokens/s, peak RAM, model-plus-index storage, power, and energy per response. Teachers or subject experts blindly rate **TODO: N responses per condition** for correctness, usefulness, and age appropriateness.

4.5 Analysis

We report means or medians with 95% confidence intervals and paired tests because conditions share questions; subjective ratings include inter-rater agreement. The primary analysis estimates the quality–resource Pareto frontier across latency, memory, power, and cost.

Table 1. Target deployment platform. Replace placeholders with measured values.

Component	Configuration
Platform / memory	TODO: Pi-class board / 8 GB
CPU / storage	TODO
OS / runtime	TODO: 64-bit Linux / runtime
Quantization	TODO: e.g., GGUF Q4_K_M
Curriculum index	TODO: dense / BM25 / hybrid
Power measurement	TODO: inline meter / telemetry
Approx. cost	TODO: \$XX

Table 2. Planned model $\times$ grounding evaluation matrix.

Model	Params.	Quant.	Grounding conditions
TODO: Model A	$\sim 0.5\text{B}$	TODO	None / Naive RAG / GroundEd
TODO: Model B	$\sim 1\text{--}1.5\text{B}$	TODO	None / Naive RAG / GroundEd
TODO: Model C	$\sim 3\text{B}$	TODO	None / Naive RAG / GroundEd
Optional reference	$\sim 7\text{B}$ or cloud	TODO	None / RAG

5 Results

5.1 Effect of Curriculum Grounding

Results will compare No RAG, Naive RAG, and GroundEd within each model scale.

Table 3. Curriculum-grounding results. All values are placeholders.

Model	Condition	Corr.	Faith.	Pedag.	Halluc.	OOS F1
0.5B	No RAG	–	–	–	–	–
0.5B	Naive RAG	–	–	–	–	–
0.5B	GroundEd	–	–	–	–	–
1.5B	No RAG	–	–	–	–	–
1.5B	Naive RAG	–	–	–	–	–
1.5B	GroundEd	–	–	–	–	–
3B	No RAG	–	–	–	–	–
3B	Naive RAG	–	–	–	–	–
3B	GroundEd	–	–	–	–	–

5.2 Grounding Versus Model Scale

Cross-scale results will compare 0.5B GroundEd with 1.5B/3B ungrounded models and 1.5B GroundEd with the 3B ungrounded model, including quality per GB of RAM and second of latency.

5.3 Offline Deployment Cost

Table 4. Resource cost on the target offline device. All values are placeholders.

Configuration	TTFT (s)	tok/s	RAM (GB)	W	Index (MB)
0.5B + GroundEd	—	—	—	—	—
1.5B + GroundEd	—	—	—	—	—
3B + GroundEd	—	—	—	—	—

5.4 Out-of-Curriculum Behavior

Results will report unsupported-answer rate, abstention precision/recall, and representative failure cases.

6 Discussion

6.1 Grounding Versus Scale

Grounding can reduce the parametric knowledge required for a bounded domain, but it does not necessarily replace reasoning or instruction-following capacity. Because very small models may also use retrieved context less reliably, retrieval quality becomes increasingly important as scale decreases. GroundEd therefore tests whether grounding shifts the minimum viable model size rather than making scale irrelevant.

6.2 Implications for AIED Unplugged

GroundEd shifts the offline-first question from whether a model runs to the smallest model that remains educationally useful with appropriate knowledge. Smaller grounded models could trade generality for cost, responsiveness, privacy, and curricular control; retrieval quality, abstention, latency, and energy are consequently educational design constraints.

6.3 Limitations

The study is limited by subject and grade coverage, textbook completeness, extraction and retrieval errors, benchmark realism, evaluator validity, hardware specificity, and linguistic or cultural scope. It evaluates technical and pedagogical capability, not student learning outcomes, which require classroom study.

7 Ethical and Equity Considerations

Local inference can broaden access and reduce transmission of student queries, but does not remove hallucination, curricular bias, inappropriate guidance, or inequitable language performance. GroundEd therefore requires source traceability and abstention; the final study will document curriculum licensing, provenance, representation, accessibility, and any human-subject determination.

8 Conclusion

GroundEd evaluates whether curriculum grounding can shift the quality–resource frontier for offline AIED by allowing smaller quantized language models to operate effectively within a bounded educational domain. Comparing grounding quality, model scale, and deployment cost provides a practical basis for determining how much local model capacity curriculum-specific educational AI requires.

References

1. Adenuga, J.A.: Retrieval-augmented generation for pedagogically aware educational ai: An expert-rated comparison of a prompt-only llm tutor and an integrated, learner-state-aware rag tutor. *Frontiers in Education* **11**, 1896839 (2026). <https://doi.org/10.3389/educ.2026.1896839>
2. Guan, J., Li, J., Hou, K., Li, J.: Educhatqa: Curriculum-aware hybrid retrieval and teacher-aligned prompting for textbook-grounded educational qa. *Complex & Intelligent Systems* **12**, 226 (2026). <https://doi.org/10.1007/s40747-026-02342-z>
3. Hevia, J.S., Arredondo, F., Kumar, V.: Towards an efficient, customizable, and accessible ai tutor. In: *Proceedings of the Innovation and Responsibility in AI-Supported Education Workshop. Proceedings of Machine Learning Research*, vol. 273, pp. 250–254. PMLR (2025), <https://proceedings.mlr.press/v273/hevia25a.html>
4. Isotani, S., Pinto, I.I.B.S., Chalco, G.C., Dermeval, D., Mello, R.F.: Aied unplugged: Leapfrogging the digital divide to reach the underserved. In: *Artificial Intelligence in Education. Communications in Computer and Information Science*, vol. 1831, pp. 772–779. Springer (2023). https://doi.org/10.1007/978-3-031-36336-8_118
5. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: *Advances in Neural Information Processing Systems*. vol. 33, pp. 9459–9474 (2020)
6. Ma, I., Krone-Martins, A., Lopes, C.V.: Integrating ai tutors in a programming course. *arXiv preprint arXiv:2407.15718* (2024), <https://arxiv.org/abs/2407.15718>
7. Rodrigues, L.A.L., Guerino, G., Silva, T.E.V., Chalco, G.C., Oliveira, L., Penha, R.S.d., Mello, R.F., Vieira, T., Marinho, M.L.M., Macario, V., Pinto, I.I.B.S., Dermeval, D., Isotani, S.: Mathaide: A qualitative study of teachers’ perceptions of

- an its unplugged for underserved regions. *International Journal of Artificial Intelligence in Education* **35**(1), 2–30 (2025). <https://doi.org/10.1007/s40593-024-00397-y>
8. Rugemalila, I.E., Cai, W., Zhang, X., Liu, C., Wang, B.: Curriculum-aware retrieval-augmented generation for bilingual tutoring in low-resource swahili–english secondary schools. *Technologies* **14**(2), 129 (2026). <https://doi.org/10.3390/technologies14020129>
 9. Santos, M.M., Barros, A., Rodrigues, L., Dermeval, D., Primo, T., Bittencourt, I.I., Isotani, S.: Near feasibility, distant practicality: Empirical analysis of deploying and using llms on resource-constrained smartphones. In: *Proceedings of the International Conference on Information and Communication Technologies and Development (ICTD '24)*. ACM (2024). <https://doi.org/10.1145/3700794.3700817>
 10. Uema, M.A., de Souza, T.d.P.C., Rodrigues, L.A.L., Dermeval, D., Bittencourt, I.I., Isotani, S.: Designing pedagogical elasticity for aied unplugged with local llms. In: *Artificial Intelligence in Education*. pp. 138–152. Springer (2026). https://doi.org/10.1007/978-3-032-29791-4_10
 11. Uema, M.A., de Souza, T.d.P.C., Rodrigues, L.A.L., Dermeval, D., Bittencourt, I.I., Isotani, S.: Towards elastic offline-first applications for aied unplugged. In: *Artificial Intelligence in Education. Lecture Notes in Computer Science*, vol. 16586, pp. 415–423. Springer (2026). https://doi.org/10.1007/978-3-032-29773-0_33