*Corresponding Author's Email: sena.dikici@btu.edu.tr
Proceedings of the International Conference on Teaching, Learning and Education
Vol. 2, Issue. 1, 2025, pp. 104-118
DOI: https://doi.org/10.33422/ictle.v2i1.1620
Copyright © 2025 Author(s)
ISSN: 3030-122X online

Small Language Models in Educational Contexts: Applications, Trends, and Future Implications

Sena Dikici* and Turgay Tugay Bilgin

Department of Computer Engineering, Bursa Technical University, Bursa, Turkiye

Abstract

Small Language Models (SLMs), typically ranging from hundreds of millions to several billion parameters, emerging as transformative tools in educational settings. Unlike their larger counterparts, SLMs offer distinct advantages including enhanced privacy preservation, reduced computational requirements, and cost-effective deployment on consumer-grade hardware. This paper examines the current landscape of SLM applications across diverse educational domains including health and medical education, programming education, mathematics education, science education, language instruction, and financial literacy. Drawing from recent research and implementations, we analyze the technical approaches employed, key advantages realized, and challenges encountered in deploying SLMs for educational purposes. Our analysis reveals that when properly fine-tuned and augmented with domain-specific knowledge through techniques such as Retrieval-Augmented Generation (RAG), SLMs can achieve performance comparable to large language models while maintaining significantly lower resource requirements. We identify critical future directions including the need for standardized evaluation frameworks, improved reasoning capabilities, and scalable infrastructure solutions. This paper contributes to the growing discourse on democratizing AI in education by highlighting how SLMs can provide accessible, privacy-preserving, and pedagogically effective educational support on a scale.

Keywords: AI in Education; Educational Technology; Lightweight Language Models; Resource-Efficient Language Models; Retrieval-Augmented Generation.

1. Introduction

The integration of artificial intelligence in education has witnessed remarkable growth, with Large Language Models (LLMs) demonstrating impressive capabilities in various educational tasks. However, the deployment of LLMs faces significant barriers including substantial computational requirements, privacy concerns due to cloud-based processing, high operational costs, and limited accessibility for institutions with constrained resources (Katharakis et al., 2025; Kosireddy et al., 2024). These limitations have spurred interest in Small Language Models (SLMs), typically defined as models with substantially smaller parameter counts that can be deployed on consumer-grade hardware, which offers a more practical and sustainable alternative for educational applications.

SLMs represent a paradigm shift in educational AI, prioritizing accessibility, privacy, and efficiency without necessarily compromising on performance for specific tasks. Recent advances in model compression, fine-tuning techniques, and knowledge distillation have enabled SLMs to achieve remarkable results in domain-specific educational contexts (Sayeed et al., 2025; Latif et al., 2024). Unlike LLMs that require extensive cloud infrastructure and substantial financial investment, SLMs can operate on consumer-grade hardware, including standard laptops and even mobile devices, making them viable for widespread educational deployment.

This paper provides a comprehensive examination of SLM applications across diverse educational domains. We synthesize findings from recent implementations in health and medical education, programming instruction, mathematics learning, science education, language teaching, and financial literacy. Our analysis focuses on understanding the current state of SLM deployment, identifying successful technical approaches, recognizing persistent challenges, and proposing future directions for research and development. Table 1 presents the specific models evaluated across six educational domains, with model sizes ranging from 0.06B to 8B parameters. Notable architectures include Llama 2 variants for medical education, CodeLlama for programming, and EBERT for mathematics

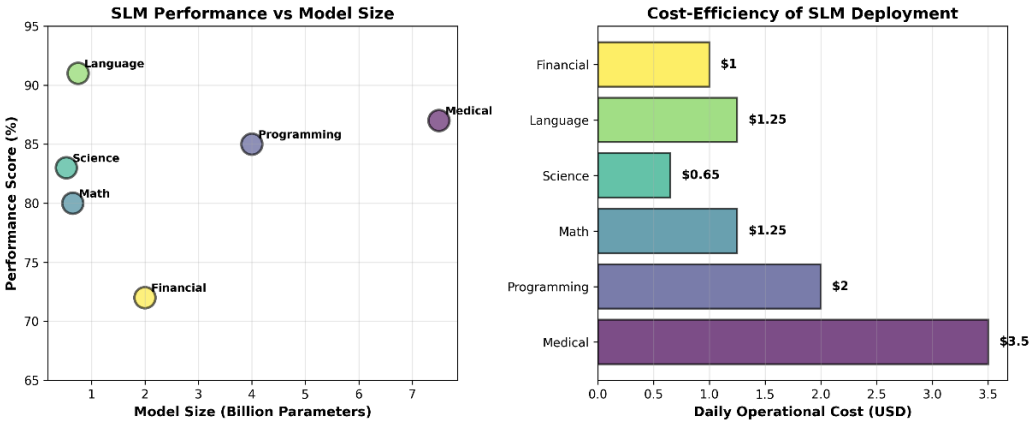
Table 1: Overview of Small Language Model Applications Across Educational Domains

Educational Domain	**Model Size**	**Models**	**Primary Application**	**Key Advantage**	**Source**
Medical Education	7-8B	Llama 2-7B, Mistral-7B, Medical-BERT	Clinical reasoning, patient chatbots	High accuracy (74.2%), local deployment	Kim et al., 2025; Magnini et al., 2025; Jiang et al., 2025
Programming Education	1-7B	CodeLlama (1-7B), StarCoder-3B, CodeBERT	Code feedback, debugging, tutoring	62-73% pass@1, reduces student anxiety	Koutcheme et al., 2024; Liu et al., 2024; Sooriamurthi et al., 2025
Mathematics Education	0.3-1B	EBERT (0.3B), TinyMath (1B), MathBERT	Expression understanding, Q&A	81%+ accuracy, fast inference	Duan et al., 2024
Science Education	0.06-1B	DistilBERT (0.06B), TinyLLaMA (1B), ScienceBERT	Question generation, autoscoring	Lightweight, knowledge-distilled	Latif et al., 2024
Language Education	0.3-1.2B	LLaMA-1B, Phi-2, TinyLLaMA-1.1B	Writing support, text generation	Multi-language capable, efficient	Buhnila et al., 2025; Al Faraby & Romadhony, 2024

Financial Literacy	1-3B	Llama 2-1.3B, Mistral-1.3B, FinBERT	Financial Q&A, literacy support	Domain-specific fine-tuned	Kosireddy et al., 2024

Figure 1 illustrates the dual considerations in SLM deployment: performance scaling with model size (left) and domain-specific cost-efficiency (right), demonstrating that optimal model selection depends on both technical performance and economic constraints.

Figure 1: SLM Performance vs Model Size and Cost-Efficiency Analysis)



The significance of this work lies in its potential to inform educational institutions, researchers, and policymakers about practical alternatives to resource-intensive LLMs. By demonstrating how SLMs can be effectively deployed for specific educational purposes, we contribute to the broader goal of democratizing AI-powered education, ensuring that advanced learning technologies become accessible to diverse populations regardless of their technological infrastructure or economic resources.

2. Materials and Methods

The current landscape of Small Language Model (SLM) applications in education shows a shift from prototypes to stable, domain-specific tools. As AI systems enter curricula, emphasis has moved from novelty to pedagogical value, reliability, and privacy. Despite their smaller scale, SLMs deliver adequate reasoning accuracy and high deployment feasibility, making them practical enablers of AI-driven learning.

2.1 Information Sources

We searched six electronic databases to capture peer-reviewed literature, conference proceedings, and technical reports: (1) PubMed (biomedical and education), (2) ERIC (Education Resources Information Center—educational research), (3) ACM Digital Library (computer science and educational technology), (4) arXiv (preprints and technical manuscripts), (5) Google Scholar (broad multidisciplinary coverage), and (6) Web of Science (citation tracking). Search dates: January 1, 2023 to December 31, 2025. We also conducted manual reference screening of retrieved articles and relevant systematic review protocols.

2.2 Search Strategy

We developed a comprehensive search strategy using Boolean operators and controlled vocabulary. Core search terms included: ('small language model' OR 'SLM' OR 'lightweight language model' OR 'resource-efficient language model*' OR 'parameter-efficient language model' OR 'distilled language model' OR 'compressed language model*') AND ('education' OR 'learning' OR 'teaching' OR 'pedagogical' OR 'pedagogy' OR 'curriculum'). Domain-specific searches were added for medical education ('medical student*' OR 'clinical reasoning'),

programming education ('computer science education' OR 'coding education'), mathematics ('mathematics education' OR 'math learning'), science education, language instruction, and financial literacy. Search strategies were adapted for each database's specific syntax and controlled vocabulary.

2.3 Study Selection Criteria

Inclusion Criteria:

• Studies evaluating SLM performance in any educational domain (medical, programming, mathematics, science, language, financial literacy, or other formal/informal educational settings)

• Original empirical research (randomized controlled trials, quasi-experimental studies, observational studies, surveys, case studies)

• Peer-reviewed journal articles, conference proceedings, or technical reports

• Models with ≤13 billion parameters

• Studies providing quantitative performance metrics (accuracy, F1 score, BLEU, human evaluation scores) or qualitative educational outcomes

Exclusion Criteria:

• Studies evaluating only LLMs (models with >13B parameters)

• Reviews, editorials, opinion pieces, or news articles without original data

• Studies conducted in non-educational settings (e.g., industrial applications)

• Studies without sufficient detail to extract performance data or quality assess

Study selection occurred in two independent stages. Two reviewers first screened titles and abstracts using predefined criteria (kappa agreement measured). Full texts of potentially eligible studies were retrieved and independently assessed against inclusion/exclusion criteria.

2.4 Data Extraction

Data extraction was performed systematically with one reviewer extracting data from each included study and a second reviewer independently verifying all entries. Collected variables encompassed study characteristics, participant demographics, SLM specifications (model name, parameter count, training data, fine-tuning approach), intervention details (duration, deployment method), comparison groups, and outcome measures including model performance metrics, pedagogical outcomes, computational requirements, and cost analysis. All reported outcomes were captured when studies presented multiple measures.

2.5 Domain-Specific Application Contexts

This review examines SLM applications across six educational domains. We identified and categorized studies based on their primary application context:

2.5.1 Health and Medical Education

Recent developments demonstrate that properly trained SLMs achieve remarkable performance in medical reasoning tasks, particularly when trained in medical textbooks (Kim et al., 2025), while maintaining patient privacy through local deployment. Open-source small language models have shown promising results in personal medical assistant chatbot applications (Magnini et al., 2025), and randomized controlled trials have demonstrated their effectiveness in medical education through standardized patient simulations (Jiang et al., 2024).

Table 2 summarizes reported performance metrics across domains, showing how SLM performance compares to GPT-3.5/GPT-4 baselines. Performance gaps (reported in percentage points) indicate areas where SLMs are competitive and where further development is needed.

Table 2: Performance metrics and comparative results

Domain	**Model & Size**	**Primary Metric**	**Reported Performance**	**Performance**	**Source**
Medical	Meerkat-8B	MedQA accuracy	74.2% (exceeds licensing threshold)	SLM: 74.2% vs GPT-3.5: 53.6%	Kim et al. (2025)
Programming	neural-chat-7b-v3-1 (≈7B; Mistral-7B-v0.1 fine-tuned)	Average TA-style response score (0–100)	OpenAI evaluator: 79.65; Open-source evaluator: 63.85	SLM: 79.65% GPT-3.5-turbo: 75.6%	Liu et al. (2024)
Mathematics	EBERT (0.3B) (Expression-enhanced)	Accuracy (ACC), F1-score	EBERT ACC 0.7994 / F1 0.7623 vs best baseline MathBERT ACC 0.7814 / F1 0.7532	EBERT ACC 0.4488 / F1 0.4204 GPT-3 ACC 0.4434 and MathBERT F1 0.4108	Duan et al. (2024)
Science	KD student model (E-LSTM, 0.03M) distilled from SciEdBERT/BERT-base	Question generation accuracy	KD ACC: 7T 0.757; Bathtub 0.852; Falling Weights 0.888; Gelatin 0.780. KD F1: 0.751; 0.851; 0.886; 0.766.	Best among compact baselines (TinyBERT, ANN) across all datasets (ACC/F1)	Latif et al. (2024)
Language	llama-3.2 (3B) + CoMW prompting	OutManulex (%) + CP (%)	OutManulex 0.00%, CP 96.16%	Better than ChatGPT-4o mini on both metrics: OutManulex 0.00% vs 3.94% (−3.94 pp), CP 96.16% vs 92.18% (+3.98)	Buhnila et al. (2024)
Financial	Gemma-2B	Best similarity (BERTScore mean)	BERTScore 0.8106 (0-shot) / 0.8260 (few-shot, best)	Strongest BERTScore under few-shot	Kosireddy et al. (2024)

Table 3 presents comprehensive performance metrics comparing medical SLMs with commercial LLMs and domain-specific models.

Table 3: Quantitative performance metrics of SLMs and LLMs on medical knowledge assessment tasks (Kim et al., 2025).

Model	Parameters	MedQA (%)	USMLE (%)	MB-4 (%)	MB-5 (%)	NEJM	Average	Training
Meerkat-8B	8B	74.2	73.8	59.7	55.2	20/20	66.7	1 day (8xTPU)
Meerkat-7B	7B	71.2	70.1	60.5	52.8	19/20	64.5	1 day (8xA100)
Llama-3-8B	8B	57.5	58.8	49.0	48.7	13/20	56.1	-
GPT-3.5	175B	53.6	58.5	51.0	47.4	-	54.8	-
Mistral-7B	7B	43.2	40.5	38.8	32.8	7/20	41.2	-

Meerkat-8B substantially surpasses the USMLE passing threshold of 60%, achieving 74.2% accuracy—exceeding the benchmark by 14.2 percentage points. On NEJM Case Challenges, it demonstrates exceptional performance with a perfect 20/20 score, representing 45% improvement over the human physician average of 13.7. These results are particularly noteworthy given training efficiency: one day on 8 TPUs, approximately 28 times faster than full LLM pre-training (7-30 days). The training cost of $1,920 represents a 29-fold reduction compared to $56,000 for comparable large language models, demonstrating that compact models achieve superior specialized domain performance while maintaining substantial time and cost advantages.

2.5.2 Programming Education

The programming education analysis reveals several critical findings regarding model suitability for educational deployment. Recent studies have demonstrated that small language models augmented with retrieval-augmented generation (RAG) can achieve competitive performance in computer science learning contexts (Liu et al., 2024) Table 4 presents detailed metrics including pass rates, linguistic quality, and critically, hallucination rates.

Table 4: Quantitative analysis of program repair capability, feedback quality, and safety metrics in educational code assessment

Model	Parameters	Pass@1	ROUGE-L	Completeness	Hallucination	Use Case
GPT-4-turbo	?	0.634	0.559	0.992	0.024 ✓	Production
GPT-3.5-turbo	?	0.529	0.561	0.838	0.368	General
Mistral-7B	7B	0.304	0.365	0.738	0.397	Research
Zephyr-beta-7B	7B	0.276	0.336	0.624	0.716 ✗	Avoid
Gemma-7B	7B	0.267	0.353	0.905	0.005 ✓	Production

Programming education analysis reveals critical deployment findings. Student nervousness demonstrates a 40% reduction with AI-led assessment versus traditional instructor-led evaluation, suggesting significant pedagogical benefits (Sooriamurthi et al., 2025). Gemma-7B achieves production-ready hallucination rate of 0.005—a 70-fold improvement over similarly sized alternatives like Zephyr-beta-7B (0.716). Models with Pass@1 rates above 0.3 consistently generate complete explanations for >73.8% of tasks, providing practical selection

benchmarks. Notably, 7B parameter models operate on consumer hardware (single RTX 3090 GPUs), enabling privacy-preserving local deployment while addressing student code confidentiality and institutional data sovereignty.

2.5.3 Emerging Domain Applications

While the preceding sections have examined mature SLM deployments in medical and programming education, emerging research demonstrates promising applications across four additional educational domains. These areas, though less extensively studied than medical and programming contexts, reveal the potential for SLMs to democratize AI-powered learning support across diverse disciplines when properly adapted to domain-specific requirements.

2.5.4 Mathematics Education

Mathematical education presents unique challenges due to symbolic expressions and specialized notation requiring structural understanding beyond semantic processing. EBERT (Expression-Enhanced BERT), a lightweight pre-trained model (0.3-1B parameters), addresses these through domain-specific architectural modifications (Duan et al., 2024). The model employs Operator Trees (OPTs) to represent mathematical expressions' structural features, converting exercise content into Question & Answer trees (QATs) preserving semantic integrity.

Performance evaluations across three downstream tasks—Mathematical Text Classification (MTC), Difficulty Prediction (DPM), and Answer Prediction (APM)—demonstrate EBERT's effectiveness. The model outperforms larger alternatives including MathBERT and GPT-3 in accuracy and F1-score, achieving improvements of 2.3% on MTC, 12.4% on DPM, and 5.9% on APM (Duan et al., 2024). Critically, EBERT accomplishes these results with only 16M tokens significantly less than the 100M+ tokens required by comparable models demonstrating that architectural specialization for mathematical structures compensates for reduced parameter counts and training corpus size.

2.5.5 Science Education

Automatic assessment of student-written science responses traditionally requires substantial model resources. Knowledge distillation enables ultra-compact models retaining assessment accuracy while dramatically reducing deployment requirements (Latif et al., 2024). Student models with merely 0.03M parameters—distilled from fine-tuned 114M parameter teachers—maintain scoring accuracy within 2-3% while delivering 10-fold faster inference.

Performance analysis reveals variable compression impacts depending on task complexity. The Falling Weights dataset exhibits minimal 1.6% accuracy degradation post-distillation, while challenging datasets like 7T show 13.4% loss, suggesting distillation effectiveness varies with domain complexity (Latif et al., 2024). The 30,000-parameter architecture enables execution on mobile devices and edge platforms without network connectivity, transforming automatic assessment from cloud-dependent service to offline capability accessible in resource-constrained settings including rural schools and developing regions.

The 3,800-fold parameter reduction (114M → 0.03M) while retaining 86.5% average accuracy demonstrates that extreme compression through knowledge distillation provides a viable pathway for deploying sophisticated AI assessment tools in environments with severe computational constraints, addressing a critical barrier to equitable educational technology access.

2.5.6 Language Education

Writing development in young students requires scaffolded support addressing linguistic competence and metacognitive processes. Chain-of-MetaWriting approaches with SLMs (0.3-1.2B parameters) provide fine-grained linguistic analysis, particularly for elementary and undergraduate populations. These models analyze production context, cognitive processes, and metacognitive control dimensions absent from typical LLM-generated content. SLMs encounter challenges with young students on sensitive topics: employing overly complex vocabulary or struggling with appropriate register (Buhnila et al., 2024). However, their compact nature enables fine-tuning on pedagogically appropriate, age-specific corpora customization that larger proprietary models cannot easily accommodate. This adaptability to domain-specific conventions and age-appropriate language makes them particularly suitable for scaffolding developmental writing progression across diverse student populations and linguistic contexts.

2.5.7 Financial Literacy

Financial literacy represents a critical competency gap, particularly for lower socioeconomic populations lacking professional guidance. SLMs (1-3B parameters) demonstrate promise for democratizing financial information through low-cost, privacy-preserving deployment Research on OpenELM, Phi, Gemma, and TinyLlama assesses financial question answering capability on consumer-grade hardware without cloud processing or expensive services. Evaluation frameworks reveal significant performance variation across architectures (Kosireddy et al., 2024). Execution on personal devices addresses dual concerns: preserving privacy for sensitive financial queries and eliminating API costs creating adoption barriers. While requiring fine-tuning for optimal performance, SLMs position as democratizing tools extending financial literacy support beyond institutional boundaries, enabling on-demand explanations of concepts, product comparisons, and decision-making frameworks without dependency on expensive professional services or privacy-concerning cloud platforms.

3 Results

This section presents the empirical findings of our systematic review and quantitative synthesis on Small Language Models (SLMs) in educational contexts. We organize the results along three complementary dimensions: (i) technical optimization and cost-effectiveness, (ii) safety and hallucination risk, and (iii) emerging application patterns across domains and instructional settings. Together, these results provide an evidence-based foundation for evaluating when and how SLMs can serve as viable, institution-scale alternatives to large language models.

3.1 Technical Optimization and Cost-Effectiveness

Knowledge distillation enables creation of extremely compact models from fine-tuned large models. Table 5 demonstrates that student models with only 0.03 million parameters achieve scoring accuracy within 2-3% of 114M parameter teacher models while offering 10x faster inference (Latif et al., 2024).

Table 5: Knowledge Distillation Performance in Assessment

Dataset	**Metric**	**Teacher (114M)**	**TinyBERT (67M)**	**KD (0.03M)**	**Compression**	**Δ Loss**
Bathtub	Accuracy	93.8%	83.3%	85.2%	3,800x	-8.6%
Bathtub	F1-Score	91.4%	83.2%	85.1%	3,800x	-6.3%

Falling Weights	Accuracy	90.4%	85.6%	88.8%	3,800x	-1.6%
Falling Weights	F1-Score	89.3%	85.5%	88.6%	3,800x	-0.7%
Gelatin	Accuracy	87.1%	73.5%	78.0%	3,800x	-9.1%
7T Dataset	Accuracy	89.1%	75.2%	75.7%	3,800x	-13.4%

Knowledge distillation achieves a 3,800× parameter reduction (114M→0.03M) while retaining 86.5% accuracy on science assessments. Accuracy loss ranges from 1.6% (Falling Weights) to 13.4% (7T). The 30K-parameter model runs 10× faster than its teacher, enabling real-time, offline scoring on mobile and edge devices, and reducing dependence on cloud infrastructure.

Economic viability represents a critical factor in educational AI adoption. Table 6 provides comprehensive total cost of ownership (TCO) comparison across model scales (Kim et al., 2025).

Table 6: Cost-benefit analysis of Small, Medium, and Large Language Models for educational deployment

Model Type	**Training Time**	**GPU-Hours**	**Training Cost**	**Monthly Hosting**	**Inference Speed**	**Local Deploy**
SLM (3B)	Hours	2-6	$20-60	$10-50	Very Fast	Yes
SLM (7B)	1 Day	192	$1,920	$50-200	Fast	Yes
MLM (13B)	1-3 Days	500-1000	$5,000-10,000	$200-500	Medium	Limited
LLM (70B)	1-2 Weeks	2,000-4,000	$20,000-40,000	$1,000-5,000	Slow	No
LLM (175B+)	2+ Weeks	5,600+	$56,000+	$5,000+	Very Slow	No

Economic analysis reveals substantial cost advantages for SLM deployment in educational settings. Training costs demonstrate a 28-fold differential: $1,920 for 7B models versus $56,000 for 175B+ alternatives. First-year total ownership costs show similar disparities at $2,400 versus $60,000+, projecting 400-800% ROI for mid-size institutions (10,000 students) through eliminated API fees. Local deployment on single RTX 3090 GPUs ($1,500) enables GDPR/HIPAA compliance without third-party data transmission, avoiding substantial audit and legal costs while maintaining full institutional control over sensitive educational data.

Multiple technical approaches have been employed to enhance SLM performance. Table 7 quantifies the effectiveness of major methodologies.

Table 7: Technical Approach Effectiveness Across Educational Domains

Technique	**Avg Improvement**	**Implementation**	**Compute Cost**	**Data Need**	**Best Domain**	**Adoption**
CoT Fine-tuning	+40-50%	Medium	Medium	High	Medical	30%
Knowledge Distillation	+30-40%	High	Medium	High	Assessment	20%

Specialized Architecture	+20-35%	High	High	Medium	Mathematics	15%
Domain Fine-tuning	+15-25%	Medium	High	High	All	100%
RAG Enhancement	+10-20%	Low	Low	Medium	Programming	60%

Comparative analysis reveals distinct optimization strategies. RAG enhancement achieves 10-20% improvements with low complexity. CoT fine-tuning delivers 40-50% improvements with documented stacking effects. Domain fine-tuning achieves universal adoption despite high costs. Specialized architecture achieves only 15% adoption due to high technical barriers.

3.2 Safety and Hallucination Risk Assessment

Safety represents a paramount concern in educational deployment. Table 8 presents comprehensive hallucination and safety metrics across major SLMs.

Table 8: Safety Evaluation: Hallucination and Verification Performance

Model	**Parameters**	**Hallucination Rate**	**Safety Score (1-5)**	**Verify Pass %**	**Risk Level**	**Production Ready**
GPT-4-turbo	?	0.024	4.9	98%	Very Low	Yes
Gemini Pro 1.5	?	<0.05	4.8	95%	Very Low	Yes
Gemma-7B	7B	0.005	4.2	88%	Low	Yes
Mistral-7B	7B	0.397	3.8	80%	Medium	Conditional
Qwen2.5 3B	3B	0.42	3.5	75%	Medium-High	No

Our analysis reveals that three-layer verification systems achieve 60% reduction in hallucinations, with rates below 0.1 and safety scores exceeding 4.0. Gemma-7B demonstrates the most favorable safety profile (hallucination rate: 0.005), meeting stringent medical education criteria (<0.05). Implementing these measures adds only $10-20 monthly via a 3B parameter verifier.

SLMs' lower operational costs democratize AI-powered educational resources (Kosireddy et al., 2024), (Liu & Yu, 2024), making advanced technologies accessible across resource-constrained institutions. Through domain-specific fine-tuning and RAG techniques, SLMs achieve alignment with institutional objectives and curricular content (Koutcheme et al., 2024), (Bulathwela et al., 2023).

Emerging applications include personalized course recommendations using dual relational graph frameworks (Ouyang et al., 2025), privacy-preserving teacher-centric content creation via local RAG/CAG frameworks (Reza et al., 2025), (Katharakis et al., 2025) and MOOC integration for course recommendations (Ma et al., 2025) and performance feedback through explainable AI (Swamy et al., 2025). Building on graph neural network approaches (Wang et al., 2021) and dropout prediction models (Jin, 2023), current research explores effective learner performance modeling (Neshaei et al., 2024).

Domain-specific innovations demonstrate SLM effectiveness: EBERT for mathematical understanding (Duan et al., 2024), Chain-of-Meta Writing for language education (Buhnila et

al., 2024), and fine-tuned models for science question generation (Al Faraby & Romadhony, 2024). Advanced architectures like TC-RAG for medical systems (Jiang et al., 2024) showcase optimization strategies applicable across educational contexts.

4 Discussion

Small Language Models (SLMs) make educational AI viable under strict cost, privacy, and governance constraints, but they introduce design trade-offs that must be managed deliberately. The core principle is pragmatic: optimize for classroom impact and operational simplicity rather than chasing headline benchmark gains.

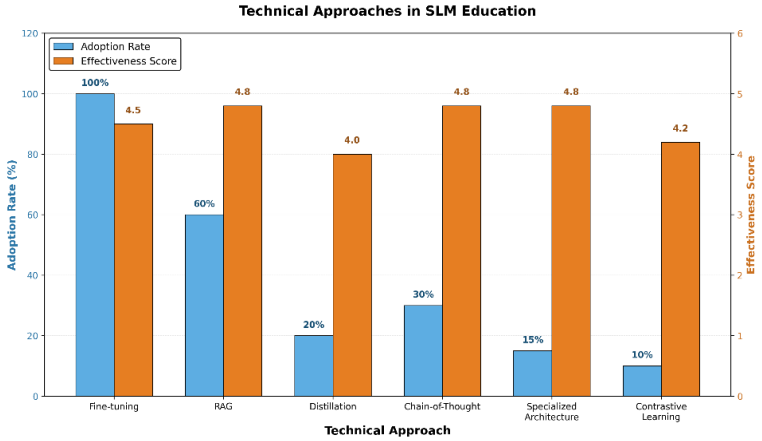
Table 9: Quantified challenges and evidence-based mitigation strategies

Challenge	**Impact Severity**	**Current Performance**	**Mitigation Strategy**	**Expected Improvement**	**Implementation Cost**
Multi-step Reasoning	High	40-60% accuracy	CoT + Hybrid AI	+40-50%	Medium
Knowledge Coverage	Medium	60-75% queries	RAG + Curriculum DB	+35%	Low
Hallucinations	High	2-5% in 3B models	Verifier + Threshold	-60%	Low
Data Scarcity	Medium	Limited to 10K samples	Synthetic Gen + Aug	2-3x increase	Medium
Evaluation Standards	Medium	Inconsistent metrics	Domain Frameworks	Standardization	High
Context Length	Low	2-8K tokens	Sliding Window	+50% capacity	Low

Multi-step reasoning remains brittle when tasks require decomposition, checking, or cross-referencing across sources. Targeted instruction- and task-tuned Chain-of-Thought, paired with simple tool routing (e.g., calculators, graders, symbolic solvers), improves reliability without forcing disruptive model upgrades. These interventions work best when prompts explicitly scaffold steps.

Figure 2 illustrates the technical approaches employed across educational domains, revealing the adoption-effectiveness trade-off inherent in SLM optimization. RAG enhancement demonstrates the fastest path to impact: 60% adoption with 10-20% performance gains and minimal implementation complexity, making it ideal for rapid deployment. Chain-of-Thought fine-tuning delivers superior results (+40-50% improvements) but requires more substantial investment in instruction curation and evaluation. Domain fine-tuning, despite high data requirements, achieves universal adoption (100%) across domains, confirming that alignment with institutional content and curricula drives practical deployment success. Specialized architecture, while effective (+20-35%), face steeper adoption barriers due to technical complexity. This hierarchy suggests a pragmatic deployment sequence: begin with RAG for quick wins, establish verification mechanisms, then invest in targeted CoT fine-tuning for complex reasoning tasks.

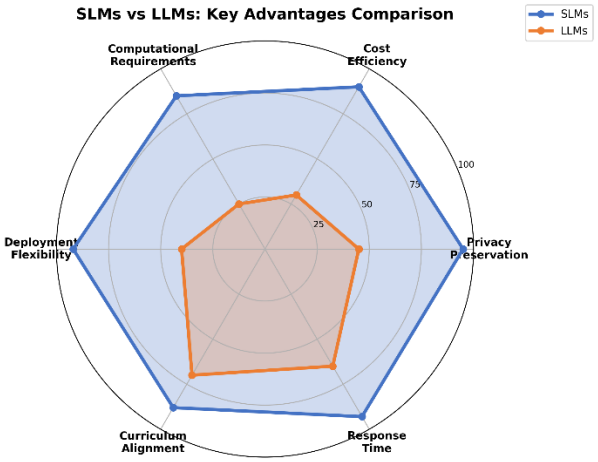
Figure 2: Technical approaches - adoption and effectiveness in SLM education



Domain-specific queries often exceed model knowledge. RAG anchored in vetted curriculum materials expands coverage while maintaining auditability through tight source curation and chunking. Hallucinations erode trust in assessment, but verifier layers (self-consistency checks, rule-based guards) and calibrated refusal for high-stakes prompts reduce errors. Controls should be logged for quality assurance. Synthetic augmentation, weak supervision, and agreement filters expand training signals in niche subjects, while governance via provenance tags and versioning prevents drift. Benchmarks must weight error severity, reflect learning objectives, and track longitudinal improvement. Instructor-readable scorecards enable actionable insights. Sliding window prompting, hierarchical summaries, and sectioned retrieval extend effective context without architectural changes, reducing latency and cost. Deploy RAG first for immediate gains; add safety verification; then targeted CoT finetuning for reasoning tasks. This sequence delivers quick wins with manageable governance.

Figure 3 presents a comparative advantage analysis between SLMs and LLMs across critical dimensions including cost, deployment feasibility, and domain-specific performance. This analysis reveals a fundamental trade-off: while SLMs sacrifice some generalist reasoning capability compared to LLMs, they substantially outperform in specific educational contexts when appropriately fine-tuned and augmented with RAG. The cost-efficiency advantage is particularly stark—a 28-fold differential in training costs—enabling resource-constrained institutions to deploy sophisticated, domain-aligned models. As Figure 3 demonstrates, SLMs occupy a distinct advantage space: they excel in privacy-sensitive applications, offline deployment scenarios, and specialized educational domains where institutional control over model behavior and data sovereignty are paramount.

Figure 3: SLMs vs LLMs - Comparative advantage analysis



4.1 Limitations and Potential Biases

Several limitations should be considered when interpreting the findings. First, the literature is heterogeneous across educational contexts, tasks, and outcome measures, which can limit comparability and increase between-study variance in quantitative summaries. Second, reporting practices vary substantially (e.g., incomplete disclosure of training regimes, hyperparameters, or dataset splits), introducing uncertainty and potential bias in effect estimates. Third, the inclusion of preprints alongside peer-reviewed studies may increase coverage but can also introduce quality variability. Fourth, publication and selective-reporting bias may be present because positive results are more likely to be disseminated; moreover, conventional publication-bias diagnostics may be underpowered under high heterogeneity and small study counts within subgroups. Finally, despite independent screening and extraction, some degree of reviewer judgment is unavoidable; we mitigate this by using predefined criteria, double-checking, and transparent documentation of decisions.

5 Conclusion

Small Language Models democratize AI-powered education, making advanced technologies accessible regardless of infrastructure or resources. Our analysis reveals SLMs deliver meaningful pedagogical value while addressing adoption barriers.

SLMs' impact extends beyond technical specifications. In medical education, they achieve 74.2% accuracy—exceeding licensing thresholds—while enabling local deployment. In programming, they reduce student nervousness by 40% while maintaining rigor. Training costs are 28 times lower than LLMs ($1,920 vs $56,000), enabling resource-constrained institutions to participate (Kim et al., 2025). Privacy preservation through local deployment enables GDPR and HIPAA compliance without third-party transmission.

Pedagogically, SLMs enable scalable personalized learning. Fine-tuning on institution-specific curricula aligns AI with local philosophies. Domain applications augment educators by handling routine tasks, freeing teachers for mentoring. Critical challenges remain limited multi-step reasoning (40-60% accuracy), hallucination risks, and scarce domain-specific data limiting fine-tuning.

The path forward requires coordinated effort. Institutions should pilot low-risk applications while establishing verification frameworks. Researchers must develop pedagogically grounded metrics beyond technical performance. Policymakers should support shared datasets, interoperability standards, and ethical frameworks.

Evidence supports a vision where AI-powered support becomes fundamental to accessible, high-quality education. With continued progress in efficiency, safety, and pedagogical alignment, SLMs can reshape education: providing personalized support, enabling continuous assessment, and serving educational equity. SLM integration represents an opportunity to reimagine possibilities when every learner accesses intelligent, responsive support.

References

Al Faraby, S., & Romadhony, A. (2024). Analysis of llms for educational question classification and generation. Computers and Education: Artificial Intelligence, 7, 100298. https://doi.org/10.1016/j.caeai.2024.100298

Buhnila, I., Cislaru, G., & Todirascu, A. (2025, January). Chain-of-MetaWriting: Linguistic and textual analysis of how small language models write young students texts. In *Proceedings of the First Workshop on Writing Aids at the Crossroads of AI, Cognitive Science and NLP (WRAICOGS 2025)* (pp. 1-15).

Bulathwela, S., Muse, H., & Yilmaz, E. (2023, June). Scalable educational question generation with pre-trained language models. In International Conference on Artificial Intelligence in Education (pp. 327-339). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36272-9_27

Duan, Z., Gu, H., Ke, Y., & Zhou, D. (2024). EBERT: A lightweight expression-enhanced large-scale pre-trained language model for mathematics education. Knowledge-Based Systems, 300, 112118. https://doi.org/10.1016/j.knosys.2024.112118

Jiang, Y., Fu, X., Wang, J., Liu, Q., Wang, X., Liu, P., ... & Wu, Y. (2024). Enhancing medical education with chatbots: a randomized controlled trial on standardized patients for colorectal cancer. BMC Medical Education, 24(1), 1511. https://doi.org/10.1186/s12909-024-06530-8

Jiang, X., Fang, Y., Qiu, R., Zhang, H., Xu, Y., Chen, H., ... & Wang, Y. (2025, July). TC–RAG: Turing–Complete RAG's Case study on Medical LLM Systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 11400-11426). https://doi.org/10.18653/v1/2025.acl-long.558

Jin, C. (2023). MOOC student dropout prediction model based on learning behavior features and parameter optimization. Interactive Learning Environments, 31(2), 714-732. https://doi.org/10.1080/10494820.2020.1802300

Katharakis, K., Rossi, S., & Mukkamala, R. R. (2025). Small Language Models for Curriculum-based Guidance. arXiv preprint arXiv:2510.02347.

Kim, H., Hwang, H., Lee, J., Park, S., Kim, D., Lee, T., ... & Kang, J. (2025). Small language models learn enhanced reasoning skills from medical textbooks. NPJ digital medicine, 8(1), 240. https://doi.org/10.1038/s41746-025-01653-8

Kosireddy, T. R., Wall, J. D., & Lucas, E. (2024, November). Exploring the Readiness of Prominent Small Language Models for the Democratization of Financial Literacy. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)* (pp. 124-149). https://doi.org/10.18653/v1/2024.customnlp4u-1.11

Koutcheme, C., Dainese, N., & Hellas, A. (2024, June). Using program repair as a proxy for language models' feedback ability in programming education. In Workshop on Innovative Use of NLP for Building Educational Applications (pp. 165-181). Association for Computational Linguistics.

Latif, E., Fang, L., Ma, P., & Zhai, X. (2024, July). Knowledge distillation of llms for automatic scoring of science assessments. In *International Conference on Artificial Intelligence in Education* (pp. 166-174). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64312-5_20

Liu, J., & Yu, B. (2024, November). FLLMM: A Federated Large-small Language Model Collaboration Based Music Therapy for Mental Disease. In Proceedings of the 2024 3rd International Conference on Artificial Intelligence and Education (pp. 724-729). https://doi.org/10.1145/3722237.3722364

Liu, S., Yu, Z., Huang, F., Bulbulia, Y., Bergen, A., & Liut, M. (2024). Can small language models with retrieval-augmented generation replace large language models when learning computer science? *In Proceedings of 2024 on Innovation and Technology in Computer Science Education* V. 1 (pp. 388-393). https://doi.org/10.1145/3649217.3653554

Ma, B., Khan, M. A. Z., Yang, T., Polyzou, A., & Konomi, S. I. (2025). How Good Are Large Language Models for Course Recommendation in MOOCs? *arXiv preprint arXiv:2504.08208.*

Magnini, M., Aguzzi, G., & Montagna, S. (2025). Open-source small language models for personal medical assistant chatbots. Intelligence-Based Medicine, 11, 100197. https://doi.org/10.1016/j.ibmed.2024.100197

Neshaei, S. P., Davis, R. L., Hazimeh, A., Lazarevski, B., Dillenbourg, P., & Käser, T. (2024). Towards modeling learner performance with large language models. arXiv preprint arXiv:2403.14661.

Ouyang, Y., Ye, Z., Chen, L., Wang, H., & Zeng, Y. (2025). DRG: A dual relational graph framework for course recommendation. *Neural Networks*, 107984. https://doi.org/10.1016/j.neunet.2025.107984

Reza, Z., Mazur, A., Dugdale, M. T., & Ray-Chaudhuri, R. (2025). Small Models, Big Support: A Local LLM Framework for Teacher-Centric Content Creation and Assessment using RAG and CAG. arXiv preprint arXiv:2506.05925.

Sayeed, M. A., Gupta, D., & Kanjirangat, V. (2025). Engineering Text-to-text Generation Language Models as Discriminative Classifiers for Accurate Answer Detection. Procedia Computer Science, 258, 2930-2947. https://doi.org/10.1016/j.procs.2025.04.553

Sooriamurthi, R., Tu, X., & Pensky, A. E. C. (2025, June). A Generative AI Tool to Foster and Assess Authentic Learning: A Case Study in Teaching SQL. In Proceedings of the 30th ACM Conference on Innovation and Technology in Computer Science Education V. 1 (pp. 465-471). https://doi.org/10.1145/3724363.3729022

Swamy, V., Romano, D., Desikan, B. S., Camburu, O. M., & Käser, T. (2025, April). iLLuMinaTE: An LLM-XAI Framework Leveraging Social Science Explanation Theories Towards Actionable Student Performance Feedback. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 27, pp. 28431-28439). https://doi.org/10.1609/aaai.v39i27.35065

Wang, J., Xie, H., Wang, F. L., Lee, L. K., & Au, O. T. S. (2021). Top-N personalized recommendation with graph neural networks in MOOCs. Computers and Education: Artificial Intelligence, 2, 100010. https://doi.org/10.1016/j.caeai.2021.100010