

Small language models offer a pathway for physics and photonics education in low-resource regions.

Hanieh Fattahi and
Asghar Ghorbani

Bridging the Digital Divide

Schoolgirls working together
with a digital tablet.
Getty Images



Small language models can help bridge the digital divide by enabling interactive on-site science education, bringing AI-powered learning to students without the need for internet access or advanced infrastructure.

S. Cook-Ordóñez

and infrastructure. For instance, approximately 50% of secondary schools in Africa do not have access to electricity, and over 90% lack adequately equipped laboratories. This shortfall severely hinders hands-on learning, reduces student engagement and negatively impacts educational outcomes.

Another challenge is the shortage of qualified STEM educators, leading to poor academic performance and a lack of confidence in pursuing STEM careers. Furthermore, economic hardship often forces students to prioritize work over education, leading to lower enrollment and retention in STEM disciplines. Gender disparities further reduce participation. Overcoming these societal barriers requires targeted policies, mentorship programs and initiatives that foster gender inclusivity in STEM.

Foundational literacy and numeracy skills are also frequently lacking in students in under-resourced countries despite increasing school enrollment numbers. In Kenya, Tanzania and Uganda, 75% of third-grade students cannot read a basic sentence.

Addressing these challenges requires comprehensive educational reform—spanning curriculum design, teacher training, infrastructure investment and targeted interventions to create equitable opportunities for students in underserved regions—all of which are costly and resource-intensive.

AI and machine learning offer a promising, low-cost solution. Large language models (LLMs) such as ChatGPT have emerged as powerful tools for personalized learning and tutoring. However, their dependence on cloud computing and stable internet limits their effectiveness. According to UNESCO, approximately 89% of students in sub-Saharan Africa lack access to a household computer, and more than 80% do not have internet access at home.

SLMs provide a compelling alternative. They are lightweight, domain-specific and designed for local deployment. They can run efficiently on smartphones or computers, enabling interactive, personalized learning without requiring internet access. In this article, we explore the current state of SLM development and highlight the models' potential to advance STEM education in underserved settings.

Limited infrastructure, scarce educational resources and unreliable internet access hinder physics and photonics education in under-resourced regions, perpetuating deep inequities in science, technology, engineering and mathematics (STEM) education. This article explores how small language models (SLMs)—compact, AI-powered tools capable of running offline on low-power devices—offer a scalable and practical solution.

By functioning as virtual tutors, supporting native-language instruction and enabling interactive learning experiences, SLMs can help mitigate the shortage of trained educators and lack of access to laboratory facilities. With targeted investment in AI technologies, SLMs can narrow the digital divide and foster scientific empowerment in marginalized communities.

The digital divide in STEM education

STEM education is a driving force behind innovation, economic development and technological progress. Yet access to quality STEM education remains limited in low-resource regions. The state of physics education in Africa exemplifies this disparity: According to the African Development Bank, fewer than 5% of African students pursue tertiary education in STEM fields, with physics ranking among the least popular subjects.

Several factors contribute to the underrepresentation of students in STEM disciplines in low-resource regions. Many schools lack essential learning resources

STEM education is a driving force behind innovation, economic development and technological progress. Yet access to quality STEM education remains limited in low-resource regions.

Limitations of LLMs: The case for SLMs

LLMs rely on transformer architectures that address the critical limitations of earlier sequential models. A key concept in language modeling is the token, which refers to the fundamental units of text that the model processes. Tokens can be individual words, subword units or even characters, depending on how the text is tokenized. For example, in a phrase like “machine learning,” the model might treat it as two separate tokens (“machine” and “learning”) or as a single unit, depending on how it is parsed.

A fundamental breakthrough of transformer models is the self-attention mechanism. Unlike previous models that processed tokens sequentially, self-attention computes relationships among all tokens simultaneously. This allows the model to capture long-range dependencies, enhancing its ability to understand complex linguistic patterns and improving performance across a wide range of language tasks.

Consider the sentence: “The physicist who discovered the new star, which had been hidden for centuries, received an award.” Self-attention effectively would resolve dependencies across varying distances, including long-range associations that traditional models struggled

with. For instance, the model correctly associates the verb “received” with “physicist” despite intervening clauses like “which had been hidden ...”.

When generating “received,” the self-attention model specifically focuses on “physicist” rather than distracting words like “star” or “centuries.” Similarly, when resolving “who,” it directly links it back to “physicist,” something that traditional models struggle with.

Self-attention is especially effective when processing long texts, such as answering questions based on a research paper. In these scenarios, the model processes the entire text—both the questions and the document content simultaneously—allowing it to efficiently identify relevant information across the full context and establish meaningful connections between distant parts of the text.

Another crucial component of transformers is multi-head attention, which allows the model to analyze text from different perspectives simultaneously. Intuitively, each attention head specializes in different linguistic aspects—one may focus on syntax, another on semantic meaning and another on contextual relationships—potentially capturing different aspects of a sentence's meaning and developing a more comprehensive and nuanced understanding of the text.



Visualization of GPT-2 attention for the sentence: “The physicist who discovered the new star, which had been hidden for centuries, received an award.” Top: Top five attention weights from the token “received” to preceding tokens. Bottom: Heatmap of three attention heads in the final layer: Head 1 focuses on “The physicist,” Head 3 on “The” and punctuation, and Head 8 on “who” and punctuation. Values are normalized as percentages.

A. Ghorbani and H. Fattahi

By prioritizing the use of high-quality, domain-specific datasets, SLMs maximize learning efficiency within tight computational budgets, allowing them to deliver competitive performance.

Although LLMs have transformed natural language processing, they require significant power and energy. For instance, training a 70-billion-parameter model like Llama 2 emits 291 tons of CO₂—equivalent to the annual emissions of 65 gasoline-powered cars. Even larger models, such as Llama 3.1, consume exponentially more resources.

While proprietary LLMs continue to push the boundaries of AI, their immense computational demands and dependence on cloud-based infrastructure limit their widespread use in many real-world applications. These models are particularly valuable in high-stakes research, advanced natural language processing and enterprise AI solutions, where performance takes precedence over cost. Yet their cost and reliance on internet connectivity are significant barriers. To overcome these challenges, the AI community has increasingly shifted its focus toward SLMs—compact yet powerful architectures that strike a balance between performance, efficiency and adaptability. Unlike their larger counterparts, SLMs represent a scalable, cost-effective alternative that supports inclusive and practical AI innovation across a wide range of environments.

SLM design, adaptability and deployment

SLMs are designed to operate efficiently on resource-constrained devices while maintaining strong performance on domain-specific tasks. This progress has been fueled by improved data curation, refined training methodologies, optimized architectural designs, enhanced fine-tuning strategies and inference-time efficiency optimizations.

Training-time techniques

Pretraining is the foundational stage where language models learn general linguistic patterns and world knowledge. LLMs typically rely on massive web-scale datasets, often containing trillions of tokens from heterogeneous sources like websites, books and online forums—an approach that introduces significant noise and redundancy.

SLMs adopt a more targeted approach. By prioritizing the use of high-quality, domain-specific datasets, they maximize learning efficiency within tight computational

budgets, allowing them to deliver competitive performance despite their smaller scale. This strategy has proven particularly effective in domains like mathematical reasoning, where data quality often outweighs model size. Strategic data curation—deduplication, domain selection and dataset balancing—can reduce data volume while preserving generalization capabilities.

Fine-tuning techniques

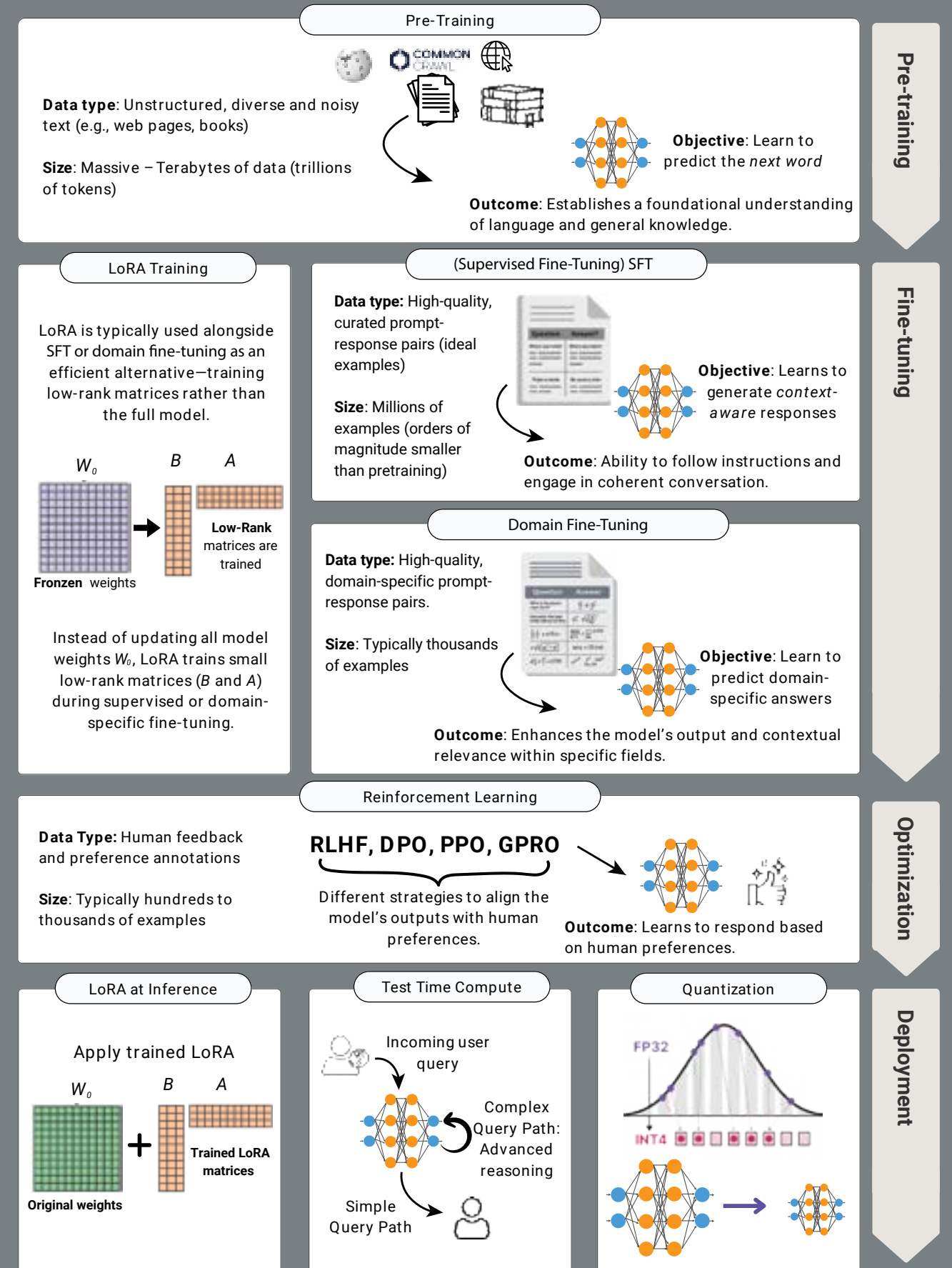
Fine-tuning adapts pretrained models to real-world applications. While pretraining provides broad linguistic and domain knowledge, fine-tuning tailors models with specialized applications.

Instruction tuning transforms models into responsive and interactive systems by training on conversational datasets using human-provided instructions. This is well-suited for applications such as AI-driven tutoring systems and specialized research assistants.

Domain adaptation refines SLMs for specialized domains such as science, mathematics, medicine and legal reasoning by training on curated domain-specific datasets. For example, models like DeepSeekMath, pretrained on math-related corpora, demonstrate that relevance often surpasses larger models in achieving high performance.

Low-rank adaptation (LoRA) is a fine-tuning technique that adapts a pretrained model for a specific task by modifying only a small portion of its overall knowledge. Instead of updating all model parameters, LoRA uses low-rank matrices to efficiently capture task-specific changes. Its modularity means multiple modules can be trained independently for different tasks and seamlessly integrated into the same base model—just like swapping lenses on a camera to suit different environments. This makes LoRA especially useful for multi-domain adaptation, enabling multiple compact, task-specific models to operate efficiently on limited hardware.

Reinforcement learning-based tuning, which uses techniques like Group Relative Policy Optimization, has gained particular attention for its effectiveness in mathematical reasoning tasks, demonstrating substantial improvements in problem-solving accuracy and model adaptability.



Overview of the training and deployment pipeline for LLMs, illustrating the stages from pretraining to real-world application.

A. Ghorbani and H. Fattahi

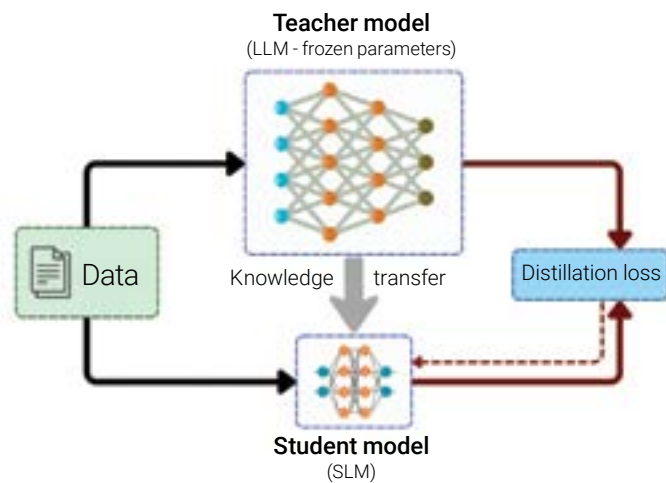


Illustration of the knowledge distillation process for training a student language model. A teacher LLM is guided by seed knowledge and skill-specific prompts to generate domain-relevant content to train the student model according to a defined learning objective.

A. Ghorbani and H. Fattahi

Test-time compute

Test-time compute (TTC) dynamically adjusts computational effort during inference. This concept distinguishes between two modes of thinking: System 1, which is fast, intuitive and effortless, and System 2, which is slower, deliberate and cognitively demanding. For example, answering factual questions like “What is the capital of France?” might only require low TTC processing, whereas solving multi-step mathematical proofs demands high-TTC deliberation.

This adaptive strategy allows SLMs to adjust their resource usage based on task complexity. For example, a 7-billion-parameter SLM leveraging TTC strategies has matched or surpassed leading commercial models like GPT-4o in specialized domains—up to 1000× reduction in compute cost—supporting sustainable AI on low-resource devices.

Knowledge distillation

Knowledge distillation enables the transfer of knowledge from large teacher models to smaller, more efficient student models. Beyond simple output imitation, it teaches small models nuanced reasoning patterns and task-specific expertise. Recent studies show that compute-efficient distillation strategies can enable a 3-billion-parameter (3B) model to outperform LLMs 100 times their size. Distilled variants of DeepSeek-R1-Distill outperform much larger non-reasoning models such as GPT-4o on a range of benchmarks.

Quantization

Quantization reduces the computer memory and processing power required to run LLMs by using

low-precision numerical formats. Switching from a 32- or 16-bit floating point to lower-precision formats such as 8-bit integers (INT8) or smaller can shrink model size by 4×, 8× or 16×, often without major performance loss. This enables large models to run efficiently on smaller, less-powerful devices like smartphones, laptops or single-GPU systems.

Advancements in on-device deployment

SLMs are increasingly being deployed directly on devices like smartphones, laptops and Internet of Things (IoT) systems thanks to innovations in three critical layers of the technology stack: hardware, frameworks and applications.

Hardware: Specialized chips for everyday AI

Modern systems-on-chip (SoCs) designs include three types of compute units—central processing units (CPUs), graphics processing units (GPUs) and neural processing units (NPUs)—purpose-built for high-throughput, low-power AI tasks. Prominent examples include Apple’s Neural Engine, Google’s Tensor SoC and Qualcomm’s Snapdragon series.

Framework: Bridging models to hardware

Inference frameworks are essential for running AI models efficiently on edge devices by translating models into hardware-optimized code. They reduce memory and compute demands through methods like quantization and apply device-specific optimizations for CPUs, GPUs and NPUs. Lightweight tools such as llama.cpp, MLC-LLM, ExecuTorch (Meta), LiteRT (Google) and MNN enable fast, portable inference across platforms. Newer solutions—including PowerInfer-2, HeteroLLM and llm.npu—further improve on-device performance through advanced scheduling and hardware-aware design. Together, these toolchains lower the resource barriers for deploying powerful AI models directly on mobile and embedded systems.

Applications: Real-world on-device SLM use cases

Tools like Gemini Nano, Apple Intelligence or PocketPal AI can operate entirely offline and deliver features like summarization, text rewriting and smart replies. Tools like LM Studio also make it possible for users to run models locally on desktops and laptops. These examples represent just a fraction of a rapidly expanding ecosystem of on-device language model applications that deliver private, low-latency AI capabilities without relying on the cloud.

The on-device deployment of SLMs is made possible by a multi-layered system architecture. This stack

In physics and photonics, SLMs can serve as on-demand, offline virtual tutors, explaining complex topics and supporting interactive, localized learning.

begins with hardware acceleration, builds upon efficient software frameworks and ultimately enables real-world applications. Each layer plays a vital role in making edge-based AI more accessible, private and efficient.

Scientific and technical applications

The recent advancements in SLMs have opened up numerous opportunities for scientific applications by enhancing efficiency, reducing computational costs and improving adaptability across specialized domains.

In mathematical reasoning, SLMs like DeepSeek-Math have been fine-tuned using mathematical datasets such as Proof-Pile-2, enabling strong performance on benchmarks like MATH and SAT, surpassing previous open-weight models. In science, SciGLM supports collegiate-level reasoning. AstroLlaMA is improving tasks in astronomy, like automated paper summarization.

Notably, recent benchmarks show SLMs can rival or outperform earlier LLMs in STEM domains. On the massive multitask language understanding (MMLU) benchmark, covering 57 academic and professional subjects, models like Qwen 2.5 and DeepSeekMath outperform many legacy LLMs—all while being efficient enough for deployment on edge devices. This progress holds promise for education and professional use in resource-limited settings.

Equitable and inclusive education

SLMs are uniquely suited to educational settings in underserved communities. They operate efficiently on low-power devices with limited computational capacity, removing the dependency on stable internet connectivity or costly infrastructure.

In physics and photonics, SLMs can serve as on-demand, offline virtual tutors, explaining complex topics and supporting interactive, localized learning. SLMs can be fine-tuned for language localization, allowing them to operate in students’ native languages, which increases accessibility and fosters linguistically inclusive learning environments.


Educators also stand to benefit. SLMs can generate lesson plans, create problem sets and translate dense academic texts into simpler language. By offering culturally and ethically adaptive outputs, SLMs can align

educational content with local values and social contexts—an essential consideration for meaningful and sustainable learning.

SLMs also promote collaborative and project-based learning. Students can use shared devices to work together on science problems, run simulations or conduct guided discussions, with the model acting as a moderator or knowledge source. In settings where lab access is limited or nonexistent, this kind of digitally mediated collaboration becomes a valuable substitute for hands-on experimentation.

In sectors like health care, defense and field research, SLMs could also address broader structural challenges associated with cloud-based AI. They reduce reliance on expensive, centralized infrastructure, mitigate latency issues and environmental costs and enhance data privacy by keeping sensitive information on-device.

Of course, challenges remain, many of which are active areas of research and must be considered when implementing real-world solutions. Hallucination remains a key concern, particularly in educational contexts where factual accuracy is critical. Additionally, multilingual performance still lags in many low-resource languages, limiting accessibility for linguistically diverse communities.

Realizing this vision will require investments in affordable hardware essential to ensure that these AI tools reach the classrooms and communities that need them most. Small language models hold immense potential to bridge the educational gap and digital divide in underserved regions, providing students and educators with tools to explore the wonders of physics and photonics, engage with cutting-edge knowledge and pursue their academic dreams. As educators and researchers, it is our collective responsibility to ensure that no one is left behind in the quest for scientific discovery and innovation. 

The authors gratefully acknowledge Loubna Ben Allal from Hugging Face for her valuable feedback on the manuscript.

Hanieh Fattahi (hanieh.fattahi@mpl.mpg.de) is with the Max Planck Institute for the Science of Light, Germany. Asghar Ghorbani (ghorban59@gmail.com) is with LLM Ventures.

References and resources: optica-opn.org/link/0925-slm.

References and Resources

- V. M. Talisayon. "[Physics teaching in developing countries](#)," Phys. Educ. 19, 105 (2002).
- C. N. R. Rao. "[Physics in the developing world](#)," Europhys. News 35, 8 (2004).
- N. R. Council. "[Optics and photonics: Essential technologies for our nation](#)." National Academies Press, Washington, DC, 2012.
- World Bank. "[Education in Africa: Challenges and opportunities](#)" (2020).
- United Nations Educational, Scientific and Cultural Organization (UNESCO). "[Global education monitoring report 2020: Inclusion and education—All means all](#)," (2020).
- T. Warren. "[Amazon web services outage is taking down a big chunk of the internet](#)," The Verge (2020).
- A. Bamgbose. "[Mother-tongue education in Africa: Context, policy, and practice](#)," Int. J. Educ. Dev. 81, 102358 (2021).
- African Development Bank. "[Africa education report 2021](#)" (2021).
- E. J. Hu et al. "[LoRA: Low-rank adaptation of large language models](#)," arXiv (2021).
- L. Ouyang et al. "[Training language models to follow instructions with human feedback](#)," arXiv (2022).
- A. Vaswani et al. "[Attention is all you need](#)," arXiv (2023).
- H. Touvron et al. "[Llama 2: Open foundation and fine-tuned chat models](#)," arXiv (2023).
- T. D. Nguyen et al. "[AstroLLaMA: Towards specialized foundation models in astronomy](#)," arXiv (2023).
- World Bank. "[Empowering Africa's future: Prioritizing STEM skills for youth and economic prosperity](#)" (2024).
- Gemma Team, "[Gemma: Open models based on Gemini research and technology](#)," arXiv (2024).
- F. Wang et al. "[A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with LLMs, and trustworthiness](#)," arXiv (2024).
- Z. Shao at al. "[DeepSeekMath: Pushing the limits of mathematical reasoning in open language models](#)," arXiv (2024).
- C. Snell et al. "[Scaling LLM test-time compute optimally can be more effective than scaling model parameters](#)," arXiv (2024).
- J. Lang et al. "[A comprehensive study on quantization techniques for large language models](#)," arXiv (2024).
- J. Xu et al. "[On-device language models: A comprehensive review](#)," arXiv (2024).
- A. Ghorbani, "[PocketPal AI - An app that brings language models directly to smart phones](#)," (2024).
- LM Studio team. [LM Studio](#). (2024).
- D. Zhang et al. "[Scilnstruct: A self-reflective instruction annotated dataset for training scientific language models](#)," arXiv (2024).
- H. Choi et al. "[Can large language models support middle school math teachers? A case study of curriculum-aligned warmups](#)," Br. J. Educ. Technol. n/a, 1 (2024).
- S. E. Huber et al. "[Leveraging the potential of large language models in education through playful and game-based learning](#)," Educ. Psychol. Rev. 36, (2024).
- World Bank Group. "[Artificial intelligence revolution in education: What you need to know](#)," (2024).
- Qwen Team. "[Qwen2.5: A Party of foundation models!](#)" Last modified September 19, 2024.
- A. Ghorbani and H. Fattahi. "[Bridging the digital divide: Small language models as a pathway for physics and photonics education in underdeveloped regions](#)," arXiv (2025).
- L. B. Allal et al. "[SmoLLM2: When smol goes big – data-centric training of a small language model](#)" (2025).
- Qwen.. "[Qwen2.5 technical report](#)," arXiv (2025).
- X. Guan et al. "[rStar-Math: Small LLMs can master math reasoning with self-evolved deep thinking](#)," arXiv (2025).
- DeepSeek-AI et al. "Deepseek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning," arXiv (2025).
- United States Environmental Protection Agency, "[Greenhouse gas emissions from a typical passenger vehicle](#)" (n.d.).