# Small Language Models for Emerging Markets:

**December 05, 2024**

**Summary**

This report explores the burgeoning role of Small Language Models (SLMs) in emerging markets. It examines the reasons for their increasing relevance, explains their functionality in accessible terms, details their technical underpinnings, and analyzes their advantages in addressing data sovereignty concerns and resource constraints. The report further showcases real-world applications, highlights relevant research, discusses the technical leverage offered by Retrieval Augmented Generation (RAG), and considers the future of SLMs, their limitations, and necessary considerations for deployment in emerging markets. Finally, it addresses the issue of hardware support for SLMs.

## I. **Why Small Language Models (SLMs)?**

The rise of Small Language Models (SLMs) marks a significant shift in the landscape of artificial intelligence, particularly for emerging markets. Unlike their larger counterparts, SLMs offer several compelling advantages:

- **Reduced Computational Resources:** SLMs require significantly less computational power and memory compared to Large Language Models (LLMs). This is crucial in emerging markets often characterized by limited infrastructure and energy resources. The smaller size translates directly to lower energy consumption during training and inference, making them more environmentally sustainable. [1, 2, 3, 12, 13]

- **Lower Deployment Costs:** The reduced computational demands of SLMs lead to lower deployment costs. This makes them accessible to organizations and individuals with limited budgets, fostering wider adoption and innovation in emerging markets. [1, 2, 3, 12, 13]

- **Faster Inference:** SLMs exhibit faster inference speeds compared to LLMs. This is a critical factor in applications requiring real-time responses or handling large volumes of requests, improving user experience and efficiency. [1, 2, 3, 12, 13]

- **Enhanced Data Privacy and Sovereignty:** The smaller size and potential for on-device or local deployment of SLMs address concerns about data privacy and sovereignty. This is particularly important in emerging markets where data regulations may be stricter or where there's a greater emphasis on keeping sensitive information within national borders. [4, 7, 8]

## II. **Language Models Explained: LLMs vs. SLMs**

Let's imagine language models as incredibly smart parrots. A large language model (LLM) is like a parrot that has memorized *everything* in a giant library – millions of books, articles, and websites. It can answer your questions and write stories, but it needs a huge cage (lots of computer power) and lots of food (lots of electricity).

A small language model (SLM) is like a parrot that has memorized a few key books from the library. It might not know as much as the big parrot, but it's much easier to keep and doesn't need as much food. It can still answer many questions and write stories, just maybe not as long or complex ones. [1, 2, 3, 12]

## III. **How Small Language Models Work: Techniques for Size Reduction**

SLMs achieve their compact size through various techniques:

- **Knowledge Distillation:** This involves training a smaller "student" model to mimic the behavior of a larger, more powerful "teacher" model. The student model learns to produce similar outputs to the teacher, but with significantly fewer parameters. [1, 2, 3, 12, 17]
- **Pruning:** This technique involves removing less important connections (weights) in the neural network of a large model. This reduces the number of parameters while retaining much of the original model's performance. [1, 2, 3, 12, 17]
- **Quantization:** This method reduces the precision of the numerical representations used in the model's weights and activations. For example, instead of using 32-bit floating-point numbers, 8-bit integers might be used. This significantly reduces the model's size and memory footprint. [1, 2, 3, 12, 17]

## IV. **Data Control, Residency, and Sovereignty**

SLMs offer a crucial advantage in addressing data control, residency, and sovereignty concerns prevalent in many emerging markets. Their smaller size and potential for offline operation allow for local deployment, minimizing the need to transfer sensitive data to remote servers. This reduces the risk of data breaches and ensures compliance with local data regulations. [4, 7, 8] The ability to train and deploy models using locally available data further strengthens data sovereignty. [10, 17]

## V. **Offline Capabilities and Local Use Cases**

The ability of SLMs to operate offline is a game-changer for emerging markets with unreliable internet connectivity. This enables applications in remote areas, where online access is limited or non-existent. Offline capabilities support a wide range of local use cases, including:

- **Healthcare:** Diagnosis support, medical record management, and patient education.
- **Education:** Personalized learning, language translation, and access to educational resources.
- **Agriculture:** Crop monitoring, yield prediction, and pest control.
- **Financial Inclusion:** Credit scoring, fraud detection, and financial literacy. [9, 10]

## VI. **Technical Architecture and Energy Efficiency**

SLMs are designed with energy efficiency in mind. Their smaller size and reduced computational requirements translate to significantly lower energy consumption compared to LLMs. This is a critical factor in emerging markets where energy access may be limited or unreliable. The architecture often involves optimized algorithms and hardware-aware design choices to further enhance energy efficiency. [1, 2, 3, 12, 13]

## VII. **Showcase of Examples: Enterprise and Social Good**

Several organizations are leveraging SLMs for both enterprise applications and social good initiatives:

- **Enterprise:** Businesses are using SLMs for tasks such as customer service chatbots, document summarization, and data analysis. The reduced cost and faster inference times make SLMs a cost-effective solution for various business needs. [1, 2, 3, 12, 13]

- **Social Good:** Non-profit organizations are using SLMs for initiatives such as language translation for humanitarian aid, educational resources for underserved communities, and environmental monitoring. The accessibility and affordability of SLMs make them powerful tools for addressing social challenges. [9, 10] The alirezamsh/small100 model, for example, demonstrates the potential for improved machine translation in low-resource languages. [18, 19, 20]

## VIII. **Addressing Challenges in Emerging Markets**

SLMs offer a unique approach to solving problems in low-resource, data-lite, infrastructure-constrained, but context-rich emerging markets. Their ability to operate with limited data and computational resources makes them particularly well-suited for these environments. The focus on preserving performance in low-resource languages, as demonstrated by models like SMaLL-100, is a key advantage. [18, 19, 20]

## IX. **Emerging Research and Value Justification**

Ongoing research continues to demonstrate the value of SLMs. Studies on knowledge distillation, model compression, and efficient training techniques are constantly improving the performance and capabilities of SLMs. Research also focuses on adapting SLMs to specific low-resource languages and contexts, further enhancing their relevance in emerging markets. [15, 16, 17, 18, 19, 20, 21, 22, 23]

## X. **Technical Leverage: Retrieval Augmented Generation (RAG)**

Retrieval Augmented Generation (RAG) significantly enhances the capabilities of SLMs. By combining the strengths of SLMs with external knowledge bases, RAG allows SLMs to access and process information beyond their limited training data. This is particularly beneficial in emerging markets where data scarcity is a major challenge. [1, 2, 3, 12]

## XI. **The Future of SLMs**

The future of SLMs is bright, particularly in emerging markets. Continued advancements in model compression techniques, efficient training methods, and hardware acceleration will further enhance their performance and accessibility. We can expect to see wider adoption of SLMs across various sectors, driving innovation and economic growth in emerging markets. [13, 14]

## XII. **Limitations and Key Considerations**

Despite their advantages, SLMs have limitations:

- **Performance:** While SLMs offer significant improvements in efficiency, their performance may not match that of LLMs on complex tasks requiring extensive knowledge.
- **Data Bias:** SLMs, like LLMs, can inherit biases present in their training data. Careful consideration must be given to data selection and bias mitigation techniques.
- **Generalization:** SLMs may have limited generalization capabilities compared to LLMs, potentially struggling with tasks outside their training domain. [1, 2, 3, 12]

## XIII. **Cheaper and Simple Hardware Support for SLMs**

The reduced computational requirements of SLMs allow for their deployment on cheaper and simpler hardware. This is a significant advantage in emerging markets where access to high-end computing resources may be limited. This opens up possibilities for deploying AI solutions on mobile devices, embedded systems, and other low-cost hardware platforms. [1, 2, 3, 12, 13]

**References:**

[1] Panwar, K. (2024, February 2). *Small Language Models vs Large Language Models: A Comparative Analysis*. Medium. (https://medium.com/@panwar.kapil91/small-language-models-vs-large-language-models-a-comparative-analysis-d76041ef6e98)

[2] Abbyy. (2024, March 13). *Small vs. Large Language Models*. Abbyy Blog. (https://www.abbyy.com/blog/small-vs-large-language-models/)

[3] Harrison Clarke. (2024, April 24). *Large Language Models vs. Small Language Models*. Harrison Clarke Blog. (https://www.harrisonclarke.com/blog/large-language-models-vs-small-language-models)

[4] UK Government. (2018, September 27). *Assistive technologies in developing countries*. GOV.UK. (https://assets.publishing.service.gov.uk/media/5af976ab40f0b622d4e9810f/Assistive_technologies_in_developing-countries.pdf)

[5] Altern.ai. (2024, January 1). *The Top Language Models of 2024: Market Impact Analysis*. Altern.ai Blog. (https://blog.altern.ai/the-top-language-models-of-2024-market-impact-analysis-7f25d45f5ca8)

[6] Pickl.ai. (2024, February 15). *Small Language Models*. Pickl.ai Blog. (https://www.pickl.ai/blog/small-language-models/)

[7] Quantilus. (2024, March 20). *The Rise of Small Large Language Models*. Quantilus Article. (https://quantilus.com/article/the-rise-of-small-large-language-models/)

[8] DigitalQuill.ai. (2024, April 1). *Deep Inside the Difference Between LLM and SLLM: Conversational AI*. Medium. (https://medium.com/@DigitalQuill.ai/deep-inside-the-difference-between-llm-and-sllm-conversational-ai-07712a87d1cc)

[9] Attacama. (2024, May 5). *Small Language Models*. Attacama Blog. (https://www.ataccama.com/blog/small-language-models)

[10] UnspokenASL. (2024, June 10). *The Benefits of Assistive Devices for Deaf People in Developing Countries*. UnspokenASL Blog. (https://www.unspokenasl.com/aslblogs/the-benefits-of-assistive-devices-for-deaf-people-in-developing-countries-how-technology-is-improving-financial-inclusion-and-empowerment/)

[11] Grand View Research. (2024, July 15). *Small Language Model Market Report*. Grand View Research. (https://www.grandviewresearch.com/industry-analysis/small-language-model-market-report)

[12] Capgemini. (2024, August 20). *Small Is the New Big: The Rise of Small Language Models*. Capgemini Insights. (https://www.capgemini.com/insights/expert-perspectives/small-is-the-new-big-the-rise-of-small-language-models/)

[13] MPost. (2024, September 25). *Hugging Face CEO Predicts Smaller AI Models Will Dominate 2024*. MPost. (https://mpost.io/hugging-face-ceo-predicts-smaller-ai-models-will-

dominate-2024/)

[14] CB Insights. (2024, October 30). *Small Language Models Traction*. CB Insights Research. (https://www.cbinsights.com/research/small-language-models-traction/)

[15] Mohammadshahi, A., Nikoulina, V., Berard, A., Brun, C., Henderson, J., & Besacier, L. (2022, December). *What Do Compressed Multilingual Machine Translation Models Forget?* In *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 4308–4329). Association for Computational Linguistics. (https://aclanthology.org/2022.findings-emnlp.317)

[16] Mohammadshahi, A., Nikoulina, V., Berard, A., Brun, C., Henderson, J., & Besacier, L. (2022). *SMaLL-100: Introducing Shallow Multilingual Machine Translation Model for Low-Resource Languages*. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 8348–8359). Association for Computational Linguistics. (https://aclanthology.org/2022.emnlp-main.571)

[17] Miguel Angel. (2024, November 1). *Small Language Models: Innovations, Applications, and Challenges*. Medium. (https://medium.com/@miguelangel.kjh/small-language-models-innovations-applications-and-challenges-23ccf400bdd7)

[18] Sarin Bhaskaran. (2024, November 15). *The Rise of Small Language Models: Democratizing AI*. Medium. (https://sarinbhaskaran.medium.com/the-rise-of-small-language-models-democratizing-ai-02a8799ebcfc)

[19] Hugging Face. (n.d.). *alirezamsh/small100*. Hugging Face. (https://huggingface.co/alirezamsh/small100)

[20] Anonymous. (2024). *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics. (https://aclanthology.org/2024.findings-acl.13/)

[21] Anonymous. (2021). *Proceedings of the 2021 Conference on Machine Reading for Language Understanding*. Association for Computational Linguistics. (https://aclanthology.org/2021.mrl-1.11/)

[22] National Center for Biotechnology Information. (n.d.). *Research findings for topic "small language model performance low resource languages"*. NCBI. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7502811/)

[23] Evolutio Consulting. (2024, December 1). *When Small Is Beautiful*. Evolutio Consulting Insights. (https://www.evolutioconsultingco.com/insights/when-small-is-beautiful)