



White Paper

Small Language Models

Democratizing Generative AI for Social Good Innovations in the Low-Resourced Global South

EqualyzAI

www.equalyz.ai

Nigeria. United Kingdom. United States

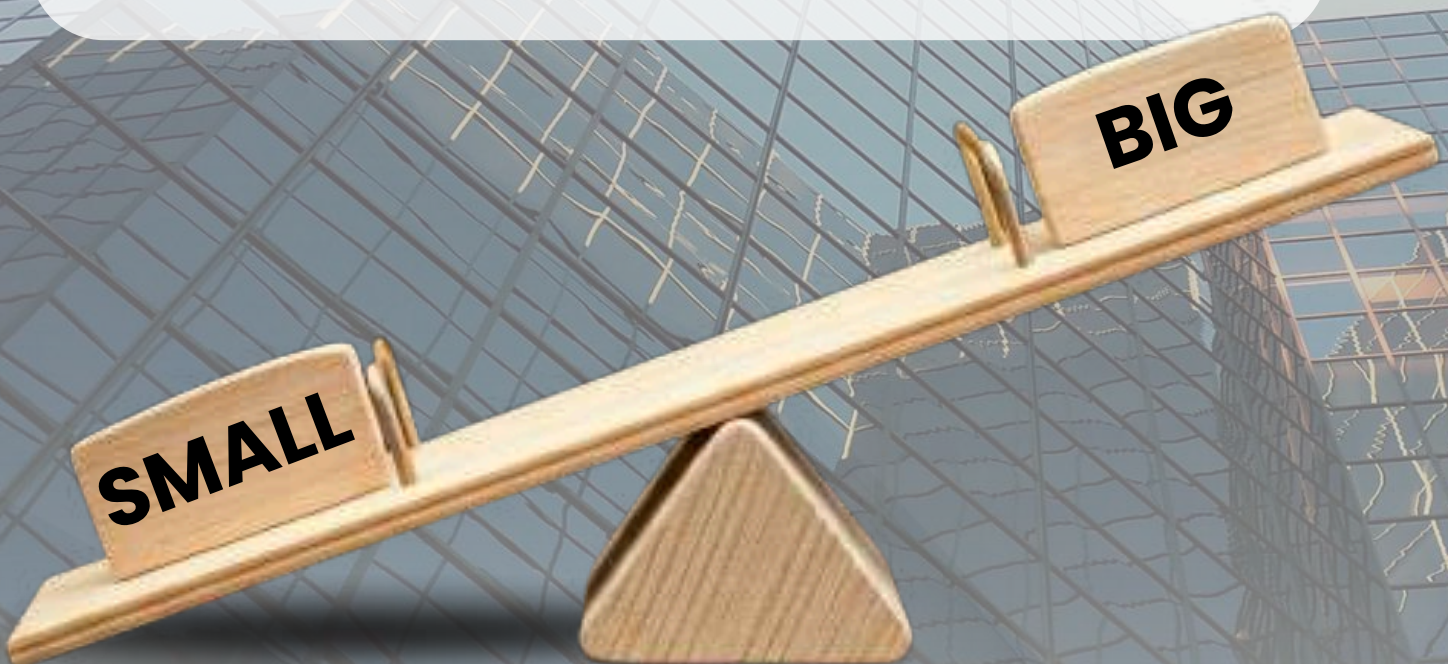


Table Of Contents



	Table of Contents	2
	Table of Figures	3
	i. Disclaimer	4
	ii. Context	5
	iii. About the Authors	7
	iv. Executive Summary	9
1	Introduction: Why Small Language Models?	10
	1.1 Choosing the Right Engine for the Journey	13
	1.2 Challenges with Large Language Models	14
	1.3 Why Small Language Models for The Global South	15
	1.4 Advantages of SLMs	16
2	Applications of Small Language Models	18
	2.1 Domain-Specific Use Cases	19
	2.2 Localized Deployment	20
3	How SLMs Work	22
	3.1 Steps to Build and Deploy an SLM	26
	3.2 Optimized Model Design	26
	3.3 Hardware Efficiency	27
	3.4 Sustainability	27
4	Emerging Trends and Innovations in Access and Inclusion	29
	4.1 Localized Deployment for Access and Inclusion.	30
	4.2 Hardware Possibilities with SLM	31
5	SLMs: An Ethical, Safe, Transparent, and Compliant Approach	33
	5.1 SLMs for Ethical and Regulatory Compliance	34
	5.2 Transparency and Risk Mitigation	35
6	Future Directions	37
	6.1 SLMs' Bright Future	38
	6.2 Limitations and Key Considerations	39
7	EqualyzAI: Enabling SLM Strategies	41
	7.1 Key Offerings	42
	7.2 Demo and Contact	42
8	Conclusion	44

Table Of Contents



v. Appendix	46
1. Small Language Models (SLMs)	
2. Large Language Models (LLMs)	
3. Llama	
4. Claude 3	
5. Moremi AI	
6. Bayer or Microsoft's E.L.Y.	
7. GPUs	
8. Nvidia	
9. Mistral AI	
10. Edge AI	
11. GDPR	
12. HIPAA	

vi. References	48
-----------------------	-----------

Table of Figures

Figure 1. Challenges in Emerging Markets	9
Figure 2: Choosing between SLM or LLM Deployment	11
Figure 3. LLMs or SLMs for Specific Needs	13
Figure 4.SLMs Overview	16
Figure 5. SLM Applications	19
Figure 6. Building an Effective SLM	26
Figure 7. Future Impact of SLMs	38



Disclaimer

This white paper is for informational purposes only. Although efforts have been made to ensure accuracy, the authors and publishers assume no liability for errors or omissions. Moreover, concepts, technologies, and case studies discussed are based on current developments and may evolve.

The content aims to explore cutting-edge AI models and their potential implementation across industries. It also does not constitute legal, financial, or technical advice. Readers should consult professionals for tailored guidance.

References to specific companies, technologies, or initiatives do not imply an endorsement unless explicitly stated.

The use of this document is at the reader's discretion. The authors and publishers are not liable for consequences arising from its use or implementation.

Bridging the AI Divide—Unlocking the Potential of Small Language Models in Emerging Markets for High-Impact Social Innovations: The EqualyzAI Mandate

The past two years have witnessed an unprecedented surge in the development and adoption of Large Language Models (LLMs). This has introduced numerous new possibilities in the AI landscape. From OpenAI's launch of ChatGPT to the development of Llama, Claude 3, and Gemini (see Appendix), the global AI innovation ecosystem has been abuzz with innovations. These models have, notably, inspired diverse use cases, including multimodal applications that generate text, speech, images, and videos through intuitive user-centric interfaces, such as chatbots and conversational assistants.

While LLMs have demonstrated unparalleled prowess across several domains in their abilities for text generation, question answering, and reasoning, they present several limitations in implementation due to the requirements of computational infrastructure and large datasets, especially in emerging market regions with limited technological infrastructure. This situation is inadvertently widening the 'digital divide' because the opportunities related to the LLM revolution are inaccessible in areas where technological inclusion is often most critical. A lack of inclusivity in training and evaluation datasets and model design has also hindered the development of nuanced, localized solutions that address the unique challenges of low-income countries. Insights from various studies indicate that most available LLMs underperform in specialized domains in low-resourced languages. This is due to insufficient domain-specific knowledge, especially in high-impact areas like healthcare, education, financial inclusion, agriculture, and governance.

It is well known that LLMs hold immense promise for improving the quality of life in the Global South (countries having a relatively low level of economic and industrial development), specifically when it comes to healthcare, education, small business operations, and public service delivery. It follows that there is a need to explore paths to democratizing LLM applications in the world's resource-constrained regions. Small Language Models (SLMs) might address this gap. They offer a compelling and practical solution to sustainable development in emerging markets where limited infrastructure, undigitized datasets, offline access points, and constrained budgets are everyday challenges.

SLMs offer reduced computational and resource requirements, low inference latency, cost-effectiveness, efficient development, and easy customization and adaptability. This, in turn, lowers the barrier to entry for governments, small businesses, and individuals seeking to integrate generative AI into their workflows. SLMs are particularly well-suited for resource-limited environments and domain-specific customization. They can address some of LLMs' challenges. They also seem ideal for applications that require localized data handling for privacy, minimal inference latency for efficiency, offline last-mile access, and domain knowledge acquisitions through lightweight fine-tuning [28; 30].

SLMs therefore represent a revolutionary approach to bridging the digital divide and making AI accessible to those who need it most. From localized agricultural advisory systems to AI-driven healthcare personalized diagnostics in underserved regions or tailored educational tools in native languages—all powered by lightweight, adaptable, and nuanced SLMs. These models also provide easy access to real-time and complex computing capabilities for startups, small- and medium-sized businesses, and governments in developing economies. SLMs are also well-positioned to accelerate the digital transformation of various use cases across industries, thereby driving innovation and delivering social good.

In this white paper, we explore how AI can be democratized with SLMs for high-impact social good innovations in the world's low-resourced regions. The integrated synthesis of knowledge captured in this piece provides an actionable blueprint for bridging the digital divide, fostering inclusion, and driving a meaningful impact across underserved regions. This can, in turn, create a future where technology works for everyone. Such an agenda is consistent with our mandate at EqualyzAI to democratize AI with SLMs in unlocking AI's full potential, specifically when it comes to sustainable development through an accessible, scalable, and inclusive alternative to LLMs. We believe that embracing SLMs will be a game-changer. They can allow us to create an ecosystem where emerging markets are active participants rather than mere spectators in the AI revolution. Indeed, we believe that AI's power can be harnessed to solve pressing challenges and improve the well-being of the next billion.



About the Authors ▼



**Olubayo
Adekanmbi, PhD**
Co-Founder, EqualyzAI

- Olubayo has 23 years of experience as a C-level technical and business leader. He has led high-impact innovations, transformational strategies, and data science and analytics projects for Africa's two largest telecommunication companies (with oversight functions covering over 300 million customers in 30 countries).
- Olubayo is also an award-winning solution developer, who has led the development of globally acclaimed solutions for the Bill and Melinda Gates Foundation, the Mastercard Foundation, the World Bank, and others.
- Five of the products Olubayo has led have been listed on the IRCAI/UNESCO Top 100 AI products for sustainable development. He also recently won the Global Grand Challenge on building LLMs for social good.
- Olubayo also served in global initiatives, such as the UK Government-led International Scientific Report on the safety of advanced AI, GSMA AI for Africa, Meta Community Forum Advisory Board member, Gates Foundation Geospatial Insight Advisory Board, DataDotOrg/Rockefeller Foundation Inclusive Growth and Recovery Challenge and many others.
- Olubayo has a PhD from the University of London (City-Cass Business School).

About the Authors ▼



Ife Adebara, PhD
Co-Founder, EqualyzAI

- Ife is a globally recognized researcher with over 7 years of experience in Natural Language Processing (NLP), linguistics, and language policy.
- She has served as a member of the Deep Learning and Natural Language Processing Group at the University of British Columbia (UBC) and as an associate member of the African Languages Technology Initiative in Nigeria.
- Ife has presented her research work at top NLP conferences, including ACL, EMNLP, COLING, and the LT4ALL conference organized by UNESCO.
- Her work has been recognized beyond the academic sphere, including media coverage by CBC News, Global News Canada, AMD, and City News Vancouver.
- Ife also won the Top 10 Outstanding Global AI Solutions Award from the IRCAI under the auspices of UNESCO for AfroLID and Serengeti—a language identification model and natural language understanding model for 517 African languages.
- Ife’s PhD is from UBC Canada. Her dissertation focused on the development of deep learning technologies for 517 African Languages and endeavoured to make “computers usable in African languages.”

Executive Summary



SLMs represent a paradigm shift in AI's utility for social good, particularly in emerging markets like Africa due to the following factors:

- Limited data resources
- Infrastructure limitations
- High cost

SLMs offer the opportunity to break these barriers by being cost-effective, scalable, and tailored to emerging markets' unique needs. Small Language Models (SLMs) provide a cost-effective, efficient, and accessible alternative to large language models (LLMs), particularly for social good in the Global South. They reduce computational demands, enable localization for underserved communities, and ensure low latency with offline capabilities, enhancing data privacy and usability in regions with limited connectivity. Emerging compact hardware allows SLMs to operate on affordable devices like smartphones, supporting scalable real-time applications in healthcare, education, and governance.

We will (a) explore how SLMs can transform sectors like healthcare, education, finance, and governance and (b) emphasize SLMs' technical, economic, and environmental advantages [1; 25]. Here is a summary of the key themes we will discuss:

- **Efficient Alternative:** SLMs are cost-effective and accessible. This reduces computational requirements and enables localization for underserved communities.
- **Improved Functionality:** On-device deployment ensures low latency, offline capabilities, and enhanced data privacy for regions with limited connectivity.
- **Advancing Accessibility:** Compact hardware allows SLMs to run on affordable devices like smartphones. This supports scalable and real-time solutions.
- **Applications for Social Good:** SLMs are ideal for healthcare, agriculture, education, and governance because they have rapid prototyping and deployment potential.
- **Ethical and Regulatory Alignment:** SLMs simplify compliance with legal, ethical, and governance standards while ensuring data security, supporting data sovereignty laws, and reducing bias.
- **Challenges and Innovation:** Although facing limits like reduced generalization and reliance on high-quality data, ongoing innovations are expanding SLMs' capabilities.
- **EqualyzAI's Role:** EqualyzAI offers tailored end-to-end SLM strategies, deployment expertise, and ethical frameworks. These can drive impactful change in the Global South.



Figure 1. Challenges in Emerging Markets



01

Introduction

Small Language Models (SLMs)

SLMs are compact, efficient, and designed for environments with limited resources. They also require few computational resources and offer quick training times.

Large Language Models (LLMs)

LLMs are resource-intensive, require massive computational power and infrastructure. Their cost and energy demand also make them less suitable for low-resource settings.

Which type of language model to deploy?

Small Language Models



Ideal for resource-constrained environments like smartphones, offering cost-effectiveness and quick training times.



Large Language Models



Suitable for complex tasks in high-performance environments, though costly and resource-intensive.

Figure 2: Choosing between SLM or LLM Deployment

Key Differences between SLMs & LLMs

The table below offers a better understanding of the differences between SLMs and LLMs across various dimensions.

Small Language Models (SLM)

SLMs are designed for accessibility in low-income and resource-constrained environments.

Compact, streamlined, and efficient. Requires few computational resources and training data.

Fast to train, allowing rapid iteration and customization for specific tasks or domains.

Ideal for deployment on devices with limited processing power like smartphones or edge computing systems.

Low energy consumption due to reduced memory use and inference times.

VS

Large Language Models (LLM)

LLMs [1] requires massive computational resources and infrastructure, which limits accessibility.

Computationally expensive, with high resource demands for training and deployment.

Training can take days or even months due to the size and complexity.

Best suited for centralized, high-performance environments.

High energy consumption means a lack of sustainability for resource-limited setups.

ACCESSIBILITY



EFFICIENCY



TRAINING TIME



ENERGY CONSUMPTION



DEPLOYMENT SUITABILITY



1.1 Choosing the Right Engine for the Journey

Imagine cutting a piece of paper—you instinctively reach for scissors: simple, efficient, and perfectly suited for the task. Now picture using a chainsaw instead. Despite its power, the chainsaw creates unnecessary complexity, waste, and inefficiency. This analogy illustrates a fundamental principle: the right tool makes all the difference.

In the rapidly evolving world of AI, this principle holds true. While Large Language Models (LLMs) are powerful and impressive—akin to Lamborghinis for AI—they often bring complexity, high costs, and resource demands that limit practical applications. For many real-world scenarios, they simply become overkill.

Enter Small Language Models (SLMs), the Toyotas of the AI landscape: reliable, scalable, and cost-effective [6]. SLMs are designed for balance—delivering impactful results without the bloated infrastructure and expense of larger models. Their simplicity and efficiency make them particularly valuable for emerging markets and industries that demand accessible, agile solutions [3; 4; 5].

This paper showcases the transformative potential of SLMs, highlighting their ability to address diverse challenges, drive innovation, and bring AI within reach of everyday applications. With a focus on practical usability and scalable performance, SLMs are redefining what's possible in AI-driven problem-solving.

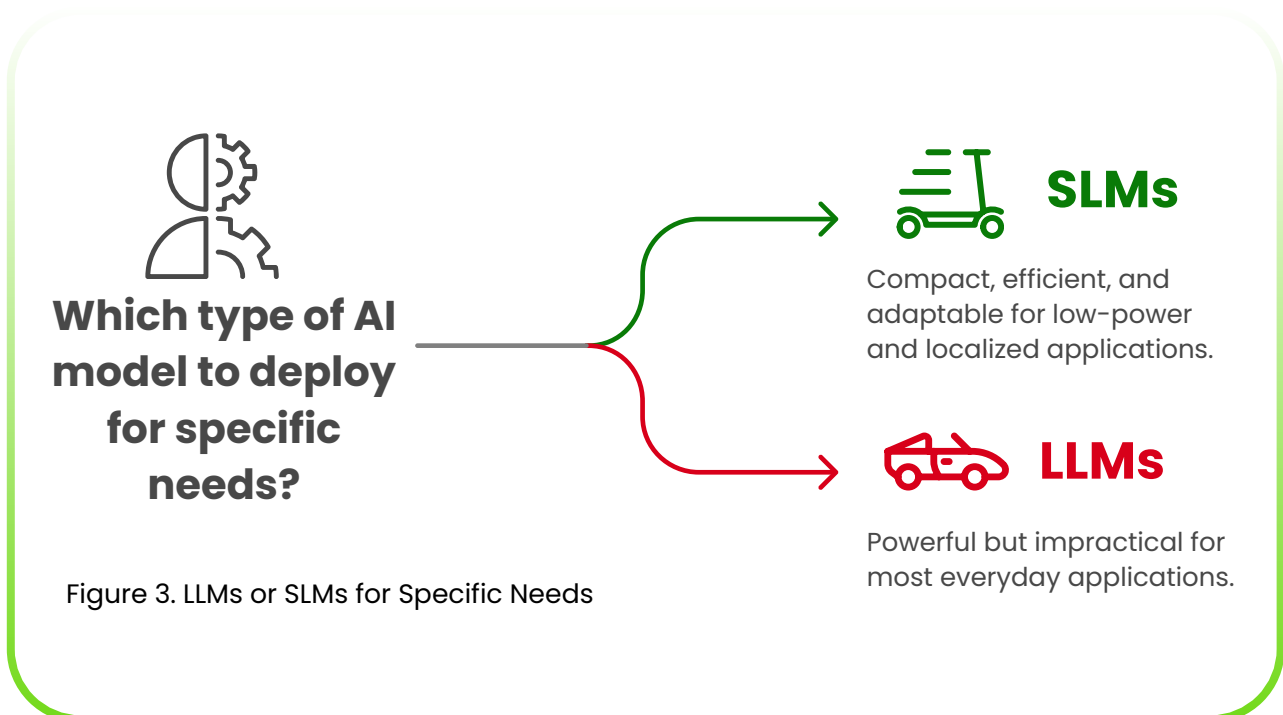


Figure 3. LLMs or SLMs for Specific Needs

1.2 Challenges with Large Language Models

In regions with limited infrastructure, the high computational and energy demands of Large Language Models (LLMs) make them impractical. These challenges are magnified by socio-economic concerns and the prohibitive costs of hardware and cloud services, leaving low-income areas unable to benefit from advanced AI.

Here are some of LLMs' disadvantages in the Global South context:

- **High Computational Demands:** LLMs require significant computational power, making them impractical for use in regions with limited infrastructure.
- **Energy Intensity:** LLMs' energy demands contribute to environmental concerns, which are particularly relevant in countries that are vulnerable to climate change.
- **Limited Accessibility:** High hardware and cloud service costs hinder LLMs' adoption in low-income regions.
- **Data Hungry:** LLMs typically require hundreds of billions, or even trillions, of tokens for pretraining (equivalent to millions of books worth of text). These are inaccessible in low-resource regions, where most languages are not currently documented in digital formats.



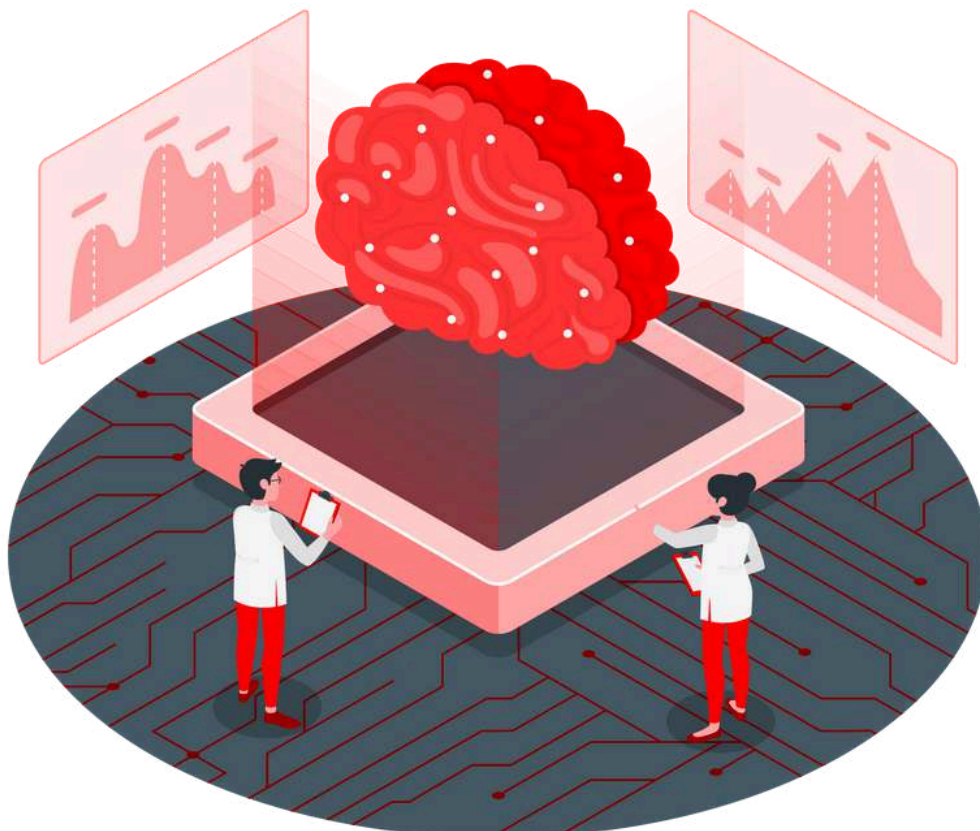
1.3 Why Small Language Models for The Global South

The Global South encompasses regions with immense potential but also faces systemic challenges when it comes to adopting cutting-edge technology. Issues like limited computational resources, unreliable internet connectivity, and constrained financial resources have created barriers to leveraging traditional Large Language Models (LLMs) [2; 20; 25; 26; 27; 44].

The alternative presented by SLM has also been validated by the fact that size is not always the defining factor in model performance [26]. The following examples highlight how Small Language Models (SLMs) can achieve competitive, if not superior, results in key benchmarks.

- SLMs are increasingly demonstrating superior performance in various tasks and even outperforming much larger models. For example, SLMs excel in content moderation, offering more precise and efficient solutions than LLMs [24].
- Arcee AI's SuperNova is a 70-billion-parameter SLM. It can, however, outperform GPT-4's massive 1.8 trillion-parameter model in instruction-following tasks.
- Microsoft's Phi-2 has 2.7 billion parameters. It showcases how a small model can deliver performance that rivals, or even surpasses, models that are up to 25 times its size [25].

SLMs are emerging as a viable option for democratizing AI by providing scalable and resource-efficient solutions that cater to these unique contexts.



1.4 Advantages of SLMs

Three key advantages of SLMs in the Global South context are as follows:



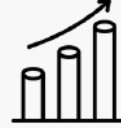
Compact and Efficient:

SLMs operate on minimal computational resources. This makes them suitable for low-power devices like smartphones and Internet of Things (IoT) systems [7].



Adaptable

SLMs can be rapidly fine-tuned and deployed for localized use cases. This enables fast response times for diverse applications [8].



Affordable and Scalable

SLMs' low cost allows development agencies, non-profit organizations, startups, and governments to experiment and scale AI solutions across sectors [8].

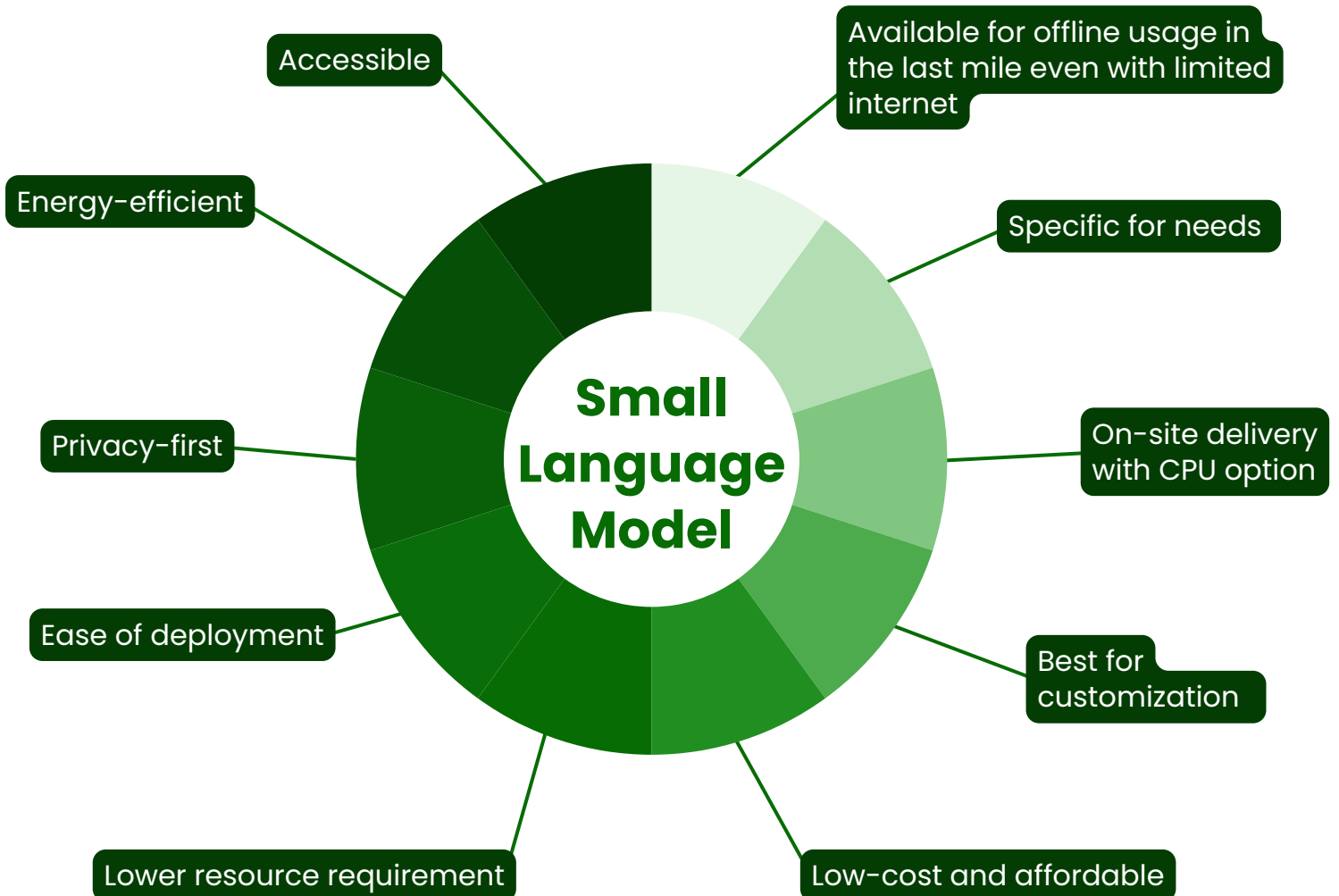







Figure 4. SLMs Overview

Small Language Models for High Impact Innovations



Nigeria • United Kingdom • United States

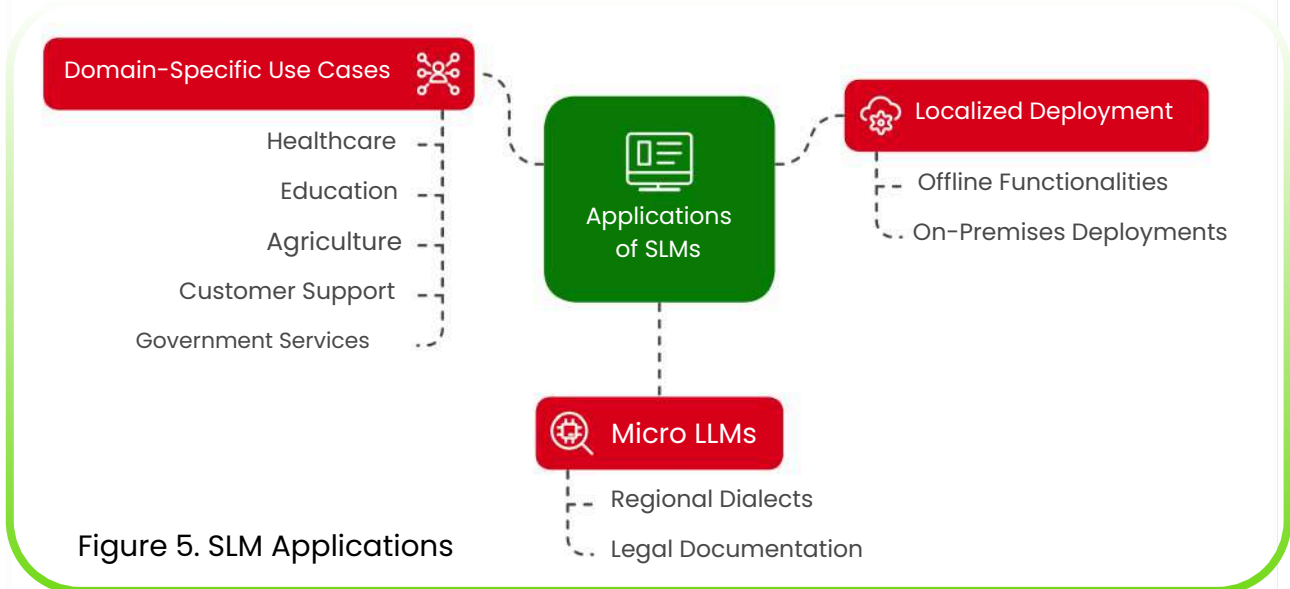
     equalyz_ai www.equalyz.ai



2 Applications of SLMs

Small Language Models (SLMs) power applications like chatbots, language translation, sentiment analysis, and local language processing.

Their efficiency and accessibility make them ideal for healthcare, education, and resource-limited environments, addressing real-world challenges with practicality [22] and [29].



2.1 Domain-Specific Use Cases

Here are some SLM use cases across key sectors:

- **Agriculture:** Bayer/Microsoft Expert Learning for You (E.L.Y.) Crop Protection system, an SLM-powered AI system based on Microsoft’s Phi-3 SLM, helps farmers approach crop protection and sustainable agriculture in a more precise and intelligent-guided manner [34].



- **Healthcare:** Tools like Moremi AI [45] assist with diagnostics and medical reporting in underserved clinics. For example, SLM-powered mobile applications can offer disease diagnosis suggestions based on identifiable symptoms, even in areas with limited internet connectivity.
- **Education:** Customized AI tutors can cater to underserved schools, helping students in remote areas access personalized learning experiences. An SLM-based learning assistant could also help bridge language barriers in multilingual classrooms. In fact, EqualyzAI recently released uLearn to develop science courses in local Nigerian languages targeted at rural students.
- **Customer Support:** SLMs enable real-time localization for query handling. This can improve customer satisfaction for businesses operating in diverse linguistic regions.
- **Government Services:** SLMs can power citizen service processes (such as e-governance platforms) for improved responsiveness and accuracy. The Singapore government’s PAIR is an example of this.

2.2 Localized Deployment

SLMs are ideal for localized deployments that allow the use of AI without cloud-based infrastructure. Two specific applications come to mind:

- **Offline Functionalities:** SLMs can operate without internet access, making them indispensable for remote and rural areas [9].
- **On-Premises Deployments:** By ensuring data security and adhering to local regulations, on-premises SLMs are ideal for sensitive applications like healthcare and governance [10].






**Personalised health diagnostics
in every rural community.
Powered in everyone's
local language.**



Nigeria • United Kingdom • United States

 equalyz_ai www.equalyz.ai



03

How SLMs Work

SLMs use compression techniques to reduce model size without sacrificing accuracy, ensuring high performance on low-specification hardware. As mentioned, SLMs run efficiently on low-power devices like smartphones and IoT systems. This reduces dependency on expensive graphics processing units (GPUs).

SLMs are also energy efficient. They consume less power than LLMs, which makes them a sustainable alternative for environmentally friendly AI applications, especially in energy-poor regions. The seven different types of model compression approaches that can be used by SLMs are explained below:

Pruning

What it is

This reduces the size of a language model while maintaining its performance. Imagine a tree with many branches. Pruning cuts away some of the nonessential branches, making the tree smaller but still healthy and useful.

What it does

It creates models that are smaller, faster, use less memory, and have improved inference speed. This is because pruned models can process inputs fast, which makes them suitable for real-time applications.

Quantization

What it is

Language models use numbers called “weights” to process and understand texts. These weights are usually represented as floating-point numbers, which can be very precise but also take up a lot of space.

What it does

It reduces the precision of these weights, representing them as simpler numbers that take up less space, for example, by rounding off the weights to the nearest whole number.

Knowledge distillation

What it is

This transfers knowledge from a large, complex language model (the “teacher”) to a smaller, simpler language model (the “student”).

What it does

This process helps the student model learn from the teacher’s expertise and perform better on relevant tasks. The result is improved performance and faster inference at a reduced size.

Low-rank factorization

What it is

This reduces the size and complexity of large matrices in language models. Think of a matrix as a big spreadsheet with many rows and columns.

What it does

It can reduce the size of these large matrices by breaking them down into smaller, more manageable pieces. It does so by identifying patterns, reducing them to smaller matrices, and then reconstructing them.

Weight sharing

What it is

This reduces the number of parameters (weights) needed to train the model. This is done by sharing the same weights across different parts of the model.

What it does

This process eliminates redundancy by identifying similar tasks, sharing the same weights across these similar tasks or patterns and reducing the number of parameters needed.

Token skipping

What it is

This reduces the number of tokens (words or characters) that must be processed. This is done by skipping over tokens that are nonessential for understanding the text's meaning.

What it does

It works by identifying tokens that can be skipped (e.g. punctuation or common function words). This means only processing the most important tokens and then adjusting the model to account for the skipped tokens. The result is that the text's meaning is still accurately captured.

Early exit

What it is

This helps to minimize computational costs and improve efficiency. This is done by allowing the model to exit the processing pipeline early, thereby skipping unnecessary computations.

What it does

This can reduce computational costs through fast processing times and improve efficiency by evaluating the model's confidence in the output after each layer. If the model is confident enough, then it exits the pipeline early, skipping the remaining layers. If necessary, the model also refines its output before returning the result.

3.1 Steps to Build and Deploy an SLM

Successful deployment of Small Language Models can be broken down into four phases.

Management & Learning

Monitor performance, gather feedback, and refine the model for improvements.

Development

Build and train the model, ensuring efficiency and compatibility with devices.

Deployment

Integrate the trained model into the target environment, optimize its performance, and make it accessible for real-world use.

Design

Define objectives, target applications, and resource constraints for the model.

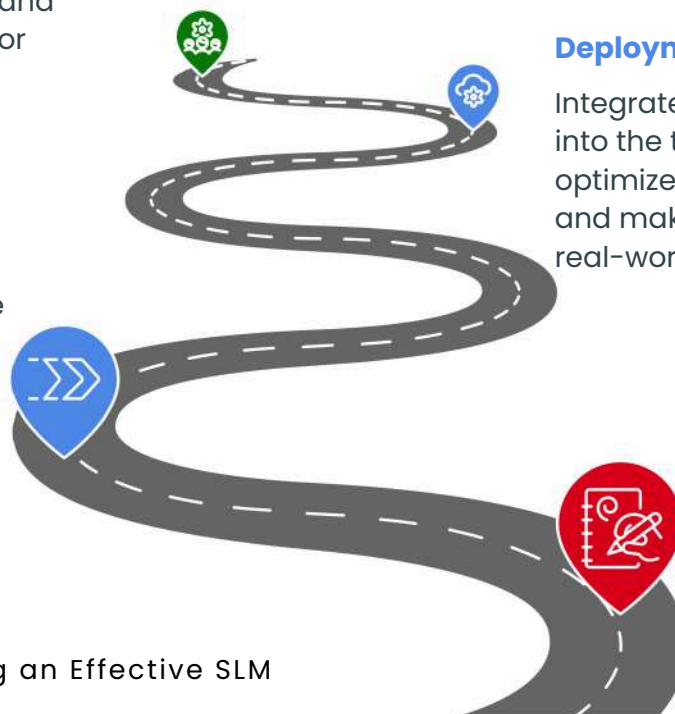


Figure 6. Building an Effective SLM

3.2 Optimized Model Design

As suggested, SLMs are ideal in use cases where domain-specific knowledge acquisition is a challenge. They present a unique solution, ensuring accuracy with a lower operational burden. Two things are worth noting when it comes to optimizing SLM model designs:

- **Compression Techniques:** Methods like pruning, quantization, and knowledge distillation reduce model size without compromising accuracy. This ensures that SLMs can deliver high performance, even on low-specification hardware.
- **Accuracy Retention:** Techniques like low-rank factorization maintain model precision. This ensures reliable outputs for critical applications (e.g., healthcare diagnostics) [2].

3.3 Hardware Efficiency

The reduction in hardware requirements to run SLMs is significant when compared with LLMs. It also allows faster implementation with specialized models across legacy systems and enables precise fine-tuning without the heavy lifting of diverse infrastructure elements. Compatibility and reduced GPU dependency are significant in this regard:

- **Compatibility:** SLMs' ability to run on low-power devices (e.g., smartphones and IoT devices) reduces dependency on expensive GPUs and/or high-end servers.
- **Reduced GPU Dependency:** SLMs enable cost-efficient operations, making AI accessible to small businesses and NGOs with limited budgets.

3.4 Sustainability






SLMs consume significantly less energy when compared to LLMs. This makes them an environmentally friendly choice. Indeed, their adoption can drive greener AI applications, especially in regions facing climate-related challenges.



**We collect, augment
and enhance text, audio,
image and video datasets
to build localised AI
models for **Agentic AI**.**



Nigeria • United Kingdom • United States

     equalyz_ai www.equalyz.ai



04

Emerging Trends and Innovations in Access and Inclusion

Big tech companies are adopting SLM innovations to improve efficiency and accessibility. Google's on-device speech recognition enhances privacy and Apple's OpenELM optimizes processes. In emerging markets, LelapaAI's Inkuba focuses on local language processing in Africa. Shakti—a 2.5B-parameter SLM—has demonstrated impactful results in resource-constrained environments. Hardware advancements like CPU-friendly models from Nvidia, Mistral AI, and Edge AI are expanding accessibility, improving connectivity, and bridging digital divides for last-mile access and deployment. The possibility of cheaper on-premises deployment will be made easier with Nvidia's \$3,000 personal AI supercomputer, dubbed Digits, which may replace the need for Data Centre for low-resourced use cases in many emerging markets.

Other recent examples of note are described below:

- [Moondream2](#) is a small vision language model (VLM) designed to run efficiently on edge devices with very little memory and a remarkably small footprint.
- [OuteTTS-0.1-350m](#) has demonstrated how a relatively small language model can learn to generate high-quality text-to-speech through a simple yet effective approach.
- Microsoft's Phi-4, the new 14B model, which performs on par with OpenAI's GPT-4o-mini, is now available as fully open source, and organization like [Unsloth AI](#) has improved it.
- Kyutai Labs's [Helium-1](#) Preview, a 2-billion parameter multilingual base LLM is a lightweight language model optimized for edge and mobile environments.
- [MiniCPM-o-2.6](#) from Open Lab for Big Model Base is an advanced multimodal model that combines vision, speech, and streaming capabilities. It has only 8B parameters that make it possible to run locally [36; 43].

4.1 Localized deployment for access and inclusion

Unlike large language models (LLMs), which often rely on data centers for processing due to their size, SLMs have the unique ability to run locally on devices like computers and smartphones. This local deployment enhances latency, preserves data privacy, and reduces dependency on internet connectivity, making them a more practical choice for diverse applications. For example, Shakti, a 2.5-billion-parameter SLM, is designed for low-resource settings, and the LLM Ware Model Depot provides accessible AI solutions for Intel PCs [36].

SLMs deliver impactful, tailored solutions for driving innovation and addressing challenges in underserved communities. SLM addresses key issues on data sovereignty, tailored use cases, and seamless integration with existing systems, as explored below:

Data Sovereignty and On-Premises

Hardware: In emerging markets [16], where data sovereignty is a critical concern, SLMs can be deployed on-premises, ensuring that sensitive data remains within local boundaries. This capability aligns with regulatory requirements while fostering trust in AI applications.

Tailored Use Cases with Focused Data: SLMs excel in delivering necessary and manageable insights tailored to specific use cases. While they may lack the breadth of larger models, they compensate with depth and precision, making them indispensable for specialized tasks.

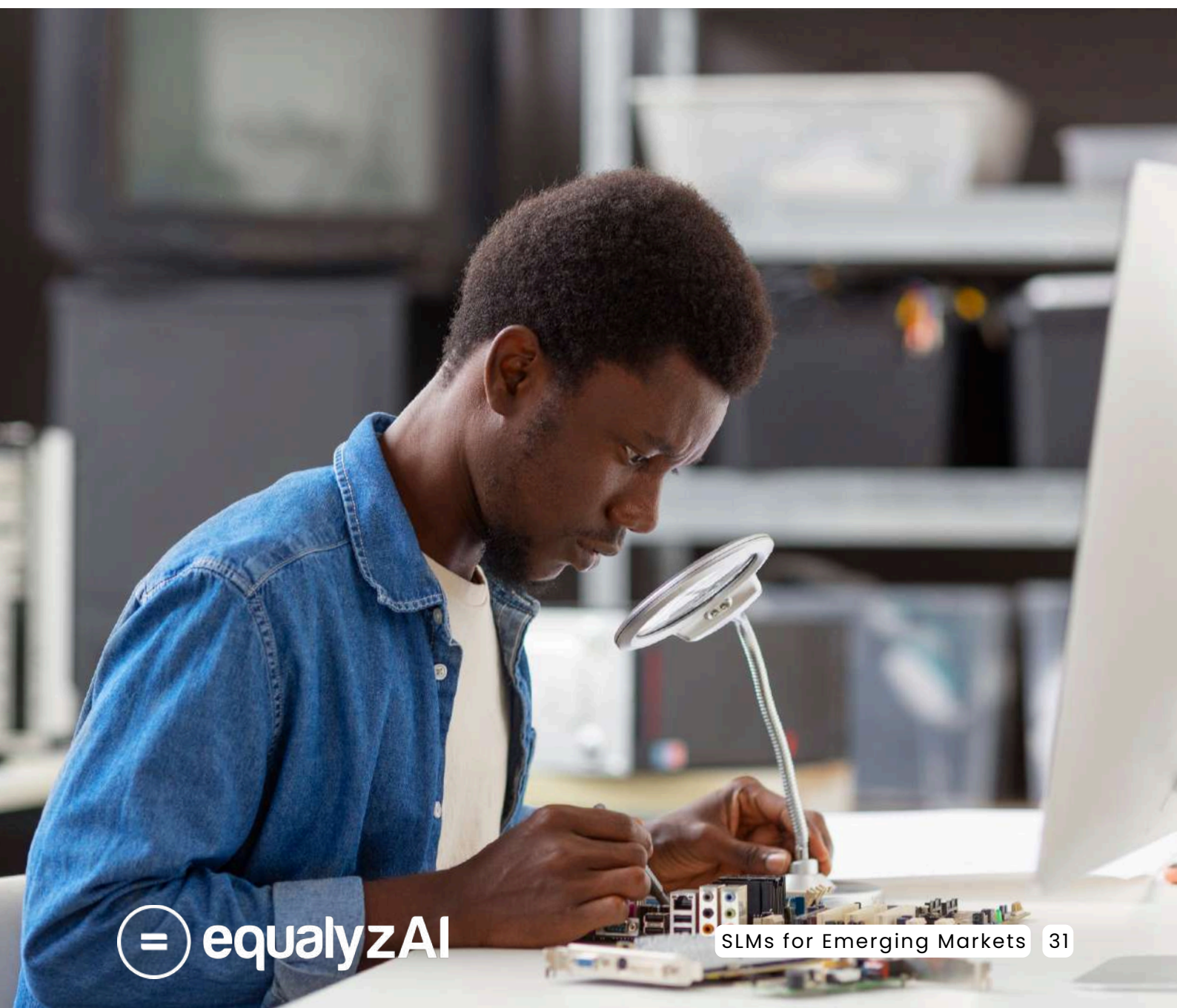
Seamless Integration: SLM can also be perfectly integrated with organizational legacy technology infrastructure, industrial control systems, and IoT systems. It is also becoming more available on low-cost hardware platforms through CPU-friendly models and Edge AI:

-
- **CPU-Friendly Models:** Nvidia and Mistral AI are developing models optimized for CPU operations, thereby expanding accessibility [14].
 - **Edge AI:** SLMs enhance connectivity and computing power for last-mile users. This can help bridge digital divides [37; 38; 39; 40].
 - **Legacy System Integration:** SLMs' lightweight nature allows easy integration with existing systems. This enhances workflow automation and productivity [37].
 - **Smart Device and IoT Integration:** SLMs can seamlessly integrate with smart devices and IoT. This enables real-time decisions in last-mile scenarios, even in rural areas.

4.2 Hardware Possibilities with SLM

There are, at least, three SLM hardware deployment and integration trends worth mentioning:

- **Cheaper and Simpler Maintenance:** Small models are easier and more cost-effective to maintain, update, and debug than larger alternatives.



Agentic AI that reasons,
acts and solves problems
in **everyone's language**
for the next 1 billion in Africa.

Moremi AI



Bayer



uLearn




PAIR




finance



Nigeria • United Kingdom • United States

equalyz_ai www.equalyz.ai

A woman with voluminous curly hair is sitting at a white desk, holding a white mug with both hands. She is wearing a light purple button-down shirt. In the background, there is a computer monitor, a keyboard, a mouse, and a vase with yellow tulips. The scene is framed by a large circular vignette.

05

SLMs: An Ethical, Safe, Transparent, and Compliant Approach

SLMs prioritize transparency, privacy, and risk mitigation. This makes them a robust choice for ethical AI deployment. Their simplified architectures ensure easier oversight and localized data processing helps meet global data protection standards, which are crucial for sensitive sectors (e.g., healthcare and governance). These include GDPR (General Data Protection Regulation) and HIPAA (Health Insurance Portability and Accountability Act).

By using focused training data, SLMs reduce the risk of inaccuracies and ‘hallucinations,’ and their task-specific design minimizes misuse. This, in turn, fosters more secure and ethical AI applications.

5.1 SLMs for Ethical and Regulatory Compliance

Seven points stand out when it comes to ethical and regulatory compliance in the SLM context. These are manageable size, security, regulatory compliance, safeguarding outputs, ease of troubleshooting, cost-effectiveness, and energy efficiency.

- **Manageable Size**

SLMs’ small size compared to larger models makes them easier to manage. This facilitates close monitoring of their operations, which helps when identifying and mitigating potential risks. This capability is especially valuable in critical applications, where precision and control are paramount [3].

- **Security**

SLMs’ limited size and scope make them less susceptible to attacks. Their reduced complexity minimizes vulnerabilities, which provides a more secure solution for sensitive applications. Their localized deployment also reduces the exposure of data to external threats. This naturally further enhances security [3].

- **Regulatory Compliance**

In certain industries, regulations require data processing to be done locally rather than in cloud environments. Because they are deployable on local devices, SLMs are well-suited to comply with these requirements. This makes them an ideal choice for applications in healthcare, finance, and other heavily regulated sectors [44].

- **Safeguarding Outputs**

SLMs can serve as intermediaries or “guardians” by monitoring LLM outputs. This functionality helps ensure that the outputs adhere to ethical standards, meet compliance requirements, and do not contain harmful or sensitive information. Indeed, SLMs can enhance AI systems’ overall reliability by acting as a safeguard layer [46].

- **Ease of Troubleshooting**

Due to their smaller scale, SLMs are simpler to debug and troubleshoot compared to larger models. This ease of maintenance ensures faster problem resolutions, leading to improved operational efficiency. Developers can quickly identify and rectify issues, which is crucial in real-time or mission-critical applications [17].

- **Cost-Effectiveness**

SLMs require fewer resources for training, deployment, and maintenance. This makes them a cost-effective and budget-friendly solution. Their ability to process data locally also eliminates the need for extensive cloud infrastructure. This ensures privacy while reducing expenses related to data transfer and storage [47].

- **Energy Efficiency**

As mentioned, SLMs’ lower computational requirements lead to reduced energy consumption, making them an environmentally friendly choice. By using SLMs, organizations can contribute to reducing their carbon footprint while still leveraging powerful AI capabilities [48].

5.2 Transparency and Risk Mitigation

Smaller targeted models make it easier to promptly gauge and address risks in, at least, two ways:

- **Reduced Hallucinations:** SLMs' focused training data minimizes the risk of generating inaccurate information.
- **Enhanced Safeguards:** Smaller, task-specific models reduce the likelihood of misuse. This, in turn, promotes ethical AI deployments.



Localize your **Digital Public Infrastructure** with **Agentic AI** that reasons in everyone's **local language.**


Health

Education

Agriculture



Nigeria • United Kingdom • United States

equalyz_ai www.equalyz.ai



06

Future Directions

6.1 SLMs' Bright Future

SLMs hold immense promise for shaping the future of AI, particularly in scenarios where accessibility, efficiency, and affordability are critical. The future impact is explored around the six themes below [37; 38; 39; 40; 41].

- **On-Device AI for Critical Interventions:** SLMs will continue to enable on-device AI solutions for sectors like healthcare and agriculture. This will address gaps in human resources and expertise in last-mile scenarios where intelligent-based and personalised decisions are required.
- **Ethical Industry Applications:** SLMs' use in ethical industries like healthcare will see more growth as it provides verifiable basis for end-to-end transparency and trust.
- **Technology Synergy:** Combining SLMs with telecom infrastructures, such as 5G networks and edge computing, opens new avenues for real-time AI applications. These integrations will open a new vista for fast and responsive AI services, which then drive innovation in various sectors.
- **Efficient and Affordable AI:** SLMs represent the future of cost-effective and accessible AI. They provide emerging markets with the opportunity to adopt AI technologies in a focused and impactful manner, even with limited datasets and constrained government budget provision.
- **Advances in Training and Architecture:** Continuous improvements in training techniques and model architectures are expanding SLMs' capabilities [18]. This will blur the distinction between SLMs and LLMs. Such developments position SLMs to power the kinds of smart devices and intuitive interfaces that address everyday challenges in the Global South.
- **Enterprise Use Cases:** Beyond social good applications, SLMs are increasingly relevant for enterprises. Specifically, they will enable more domain-specific use cases, thereby increasing innovations, improving productivity, and creating commercial value with broader socio-economic opportunities.

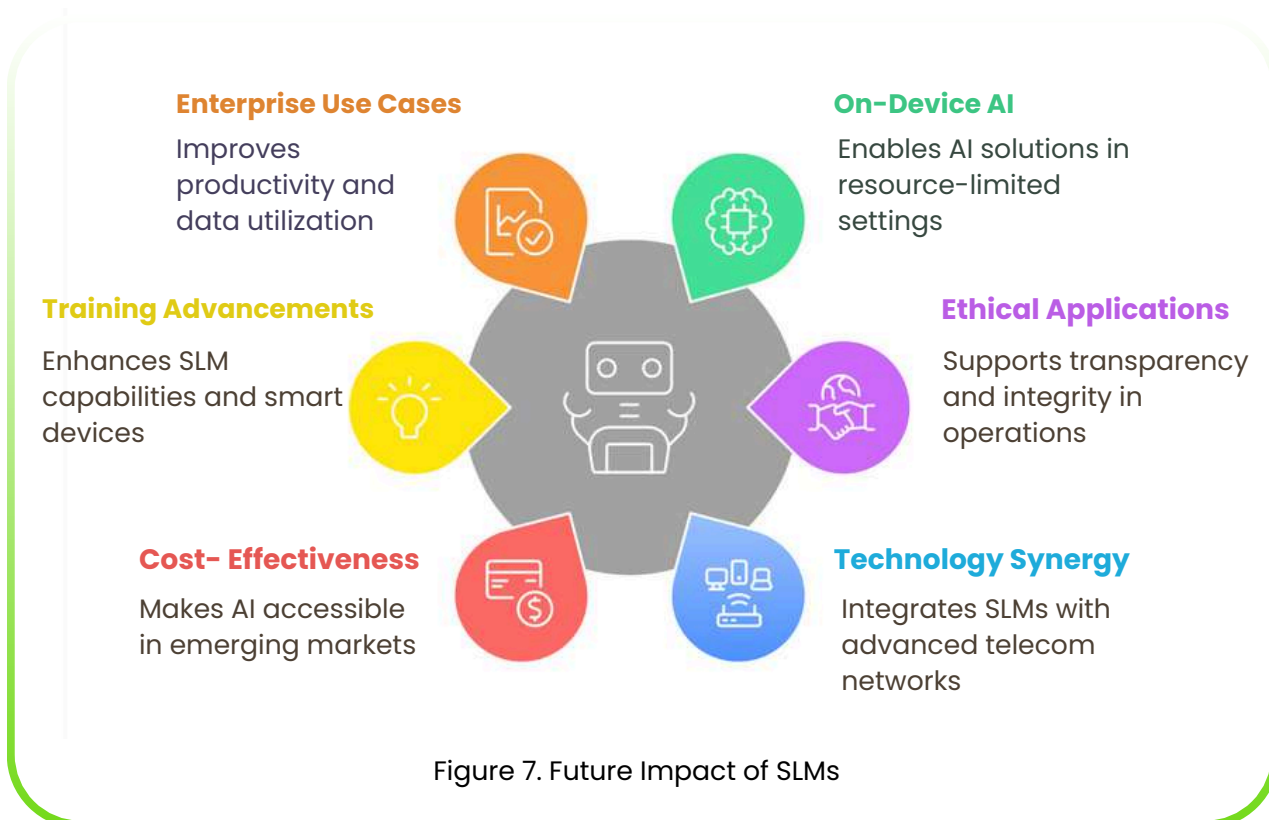


Figure 7. Future Impact of SLMs

6.2 Limitations and Key Considerations

Although SLMs offer significant advantages, we must acknowledge their limitations if we are going to ensure realistic expectations and effective implementation [19; 42]. Three such limitations stand out:

- **Limited Generalizability:** Due to a reduced parameter count, SLMs may struggle with contextual understanding and capturing complex linguistic nuances. This limitation can impact their performance in tasks requiring deep comprehension.
- **Reduced Generative Capabilities:** Compared to larger models, SLMs may find it challenging to produce diverse or widely representative output. This could restrict their utility in creative or narrative-driven applications.
- **Additional Datasets for Domain-Specific Adaptation:** Fine-tuning SLMs for niche or highly specialized domains often demands additional data and resources. This requires bespoke data collection, augmentation, and enrichment, especially in areas where dataset do not exist at all.










From Conversational Chatbots to Intelligent AgentAI. Powered in everyone's local language.



Nigeria • United Kingdom • United States

     equalyz_ai www.equalyz.ai



07

EqualyzAI: Enabling SLM Strategies

EqualyzaI specializes in developing Small Language Models (SLMs) tailored for hyperlocal and emerging market needs. By leveraging high-quality, domain-specific data and expert language enrichment, we create custom solutions for sectors like healthcare, agriculture, education, financial inclusion, and governance. EqualyzaI's resource-efficient and real-time models promote AI adoption in underserved regions, fostering social good.

Specifically, the mission at EqualyzaI is to democratize AI using SLMs. Doing so can bridge the digital divide, foster inclusion, and drive meaningful impact across underserved regions. We envision a future where technology works for everyone, everywhere.

7.1 Key Offerings

Three of EqualyzaI's key offerings are as follows:

1 Data Collection and Enrichment

Ecosystem: We have built an ecosystem for hyperlocal, multimodal, and domain-specific data collection and enrichment. Such an ecosystem can facilitate rapid, sustained, and purpose-driven data collection, annotation, review, enrichment, and augmentation. This is supported by (a) our specialized on-body audio recording devices for spontaneous speech by natural speakers and (b) an incentivized, enrichment, and augmentation app for continuously collecting data and updating existing datasets.

2 Purpose-built, Iterative, and Nuanced SLM Developments:

We have a unique ability to collect the most original, representative, hyperlocal, and largely undocumented datasets (including text, speech, image, and video). This empowers us to build nuanced models that perform better than the existing models.

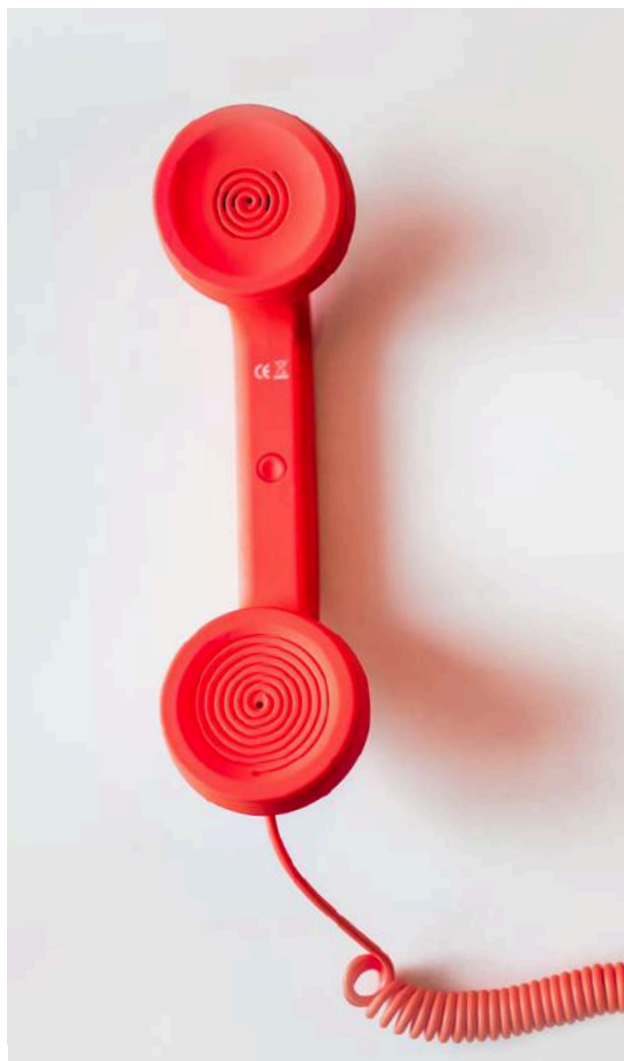
3 Innovative Agentic AI Products for Enterprise and Development Agencies:

We leverage our unique datasets and nuanced models to build tailor-made and high-impact solutions in people's local languages. This is especially important in social good domains related to healthcare, agriculture, financial inclusion, education, and governance.

7.2 Demo and Contact

Experience EqualyzaI's Financial Inclusion SLM and engage our bespoke solution team to work with your specific use case for social good, enterprise, or government.

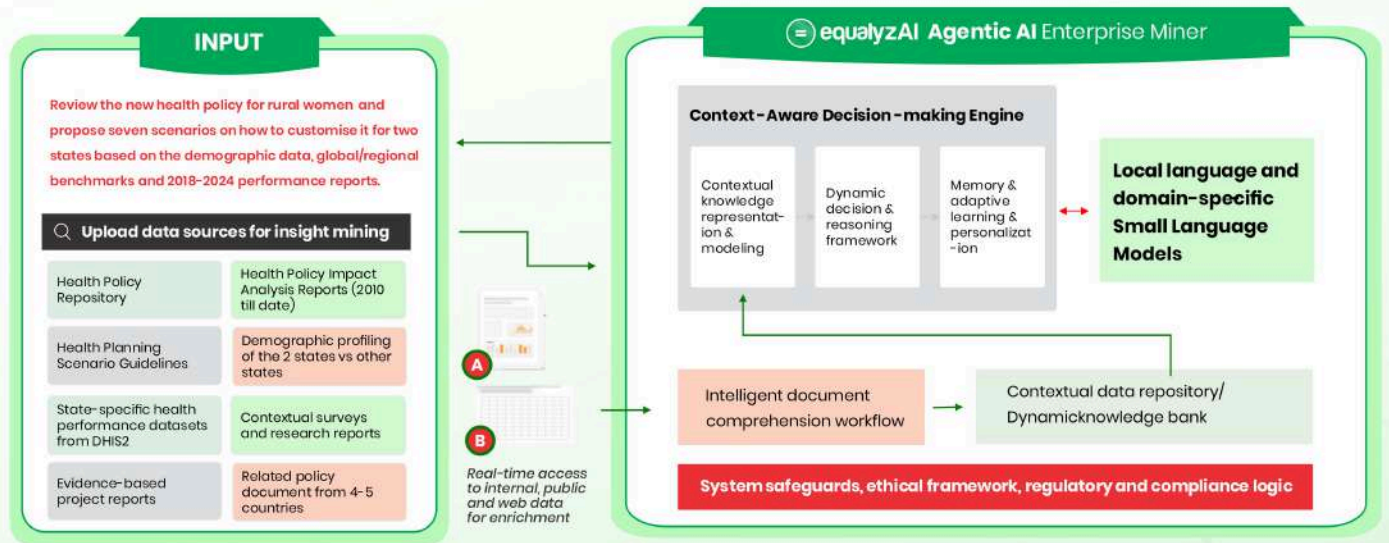
[Click to watch the demo video](#)





EqualyzAI Agentic AI Enterprise Miner

for localised and context - aware decision making
for enterprise and development agencies





08

Conclusion

We have discussed how SLMs hold immense potential to revolutionize sectors like healthcare, education, finance, and governance. To unlock SLMs' capabilities, governments, development agencies, multilateral organizations, non-profit organizations, and tech innovators must find common ground and collaborate [35].

By democratizing AI through SLMs, we can bridge the digital divide and drive meaningful impact in underserved regions. This can create a future where technology serves the many rather than the few.



Appendix



V

- **Small Language Models (SLMs):** Streamlined versions of language models specifically designed to be lightweight and efficient. With fewer parameters, they are versatile, cost-effective, and accessible. SLMs excel in addressing practical, real-world challenges where simplicity and reliability are prioritized.
- **Large Language Models (LLMs):** Advanced AI systems are built with billions of parameters, which allows them to understand and generate highly complex and nuanced text. They are powerful, sophisticated, and ideal for tackling intricate and demanding tasks. However, their size and computational requirements make them resource-intensive and less practical for everyday use.
- **Llama (Large Language Model Meta AI):** A family of advanced language models developed by Meta for natural language tasks like text generation, translation, and AI applications. Llama is designed to be efficient, fine-tuneable, and accessible for research and non-commercial use. Llama models compete with state-of-the-art models like GPT and offer open access to researchers. This fosters collaboration and innovation.
- **Claude 3:** The latest AI model by Anthropic, designed for safe and effective text-based interactions. Claude 3 excels at understanding complex queries, handling long contexts, and performing tasks like writing, coding, and summarization. With a focus on safety and alignment, the model is widely used in customer support, content creation, and education.
- **Moremi AI:** A state-of-the-art generative AI system driving research and innovation in biology, biochemistry, and drug discovery. Trained on diverse biomedical data from multiple continents, Moremi AI serves as a key tool for advancing global scientific research.
- **Bayer/Microsoft's E.L.Y. (Expert Learning for You):** A collaborative initiative between Bayer and Microsoft, one that combines their expertise in life sciences and cloud technology to accelerate innovation in healthcare and agriculture. This platform leverages advanced AI, data analytics, and cloud computing to address global challenges, including improving access to healthcare, enhancing agricultural productivity, and promoting sustainable practices. By harnessing the power of technology and data, E.L.Y. aims to make a meaningful impact in improving people's lives across the globe.
- **GPUs (Graphics Processing Units):** Traditionally used for high-performance computing tasks like training and running large AI models (e.g., LLMs), GPUs are powerful. This is because they can process many calculations simultaneously, which makes them essential for large-scale, deep-learning tasks.
- **Nvidia:** Known for its GPUs and specialized hardware like the Nvidia Jetson platform, which is designed to accelerate AI computations on edge devices (e.g., smartphones and IoT systems). Nvidia also develops CPU-friendly models, which means that their hardware advancements allow AI models like SLMs to run efficiently on devices with low power requirements. This expands accessibility and supports low-cost, energy-efficient AI operations, particularly in resource-constrained environments.

-
- **Mistral AI:** A company specializing in developing AI models and hardware that focus on maximizing computational efficiency, especially for edge AI applications. Mistral AI is contributing to the development of hardware or models optimized for resource-constrained environments where small, energy-efficient models like SLMs can effectively perform their tasks. This supports last-mile users in emerging markets who may not have access to expensive or powerful computing infrastructure.
 - **Edge AI:** Deploying AI models directly on local devices (e.g., smartphones and IoT devices), enabling real-time data processing without relying on central servers. This reduces latency, minimizes bandwidth use, enhances privacy, and improves energy efficiency. Edge AI is particularly valuable in resource-constrained environments. It offers applications like on-device speech recognition, smart sensors, and real-time decision-making for autonomous systems while ensuring that sensitive data remains secure and that processing occurs locally.
 - **GDPR (General Data Protection Regulation):** A law that protects EU residents' privacy and personal data. GDPR requires businesses to obtain consent before processing personal data, mandates strong data protection measures, and gives individuals the right to access and delete their data. Non-compliance can result in heavy fines.
 - **HIPAA (Health Insurance Portability and Accountability Act):** A US law that ensures the privacy and security of health data. HIPAA regulates how healthcare providers and insurers handle protected health information. It also requires safeguards to prevent breaches and gives patients access to their health records. Non-compliance can lead to significant penalties.



References

The logo consists of the lowercase letters 'vi' in a bold, sans-serif font, centered within a white circle. This circle is surrounded by a thick green ring.

- [1] Microsoft blog exploring AI models and differences between LLMs and SLMs <https://www.microsoft.com/en-us/microsoft-cloud/blog/2024/11/11/explore-ai-models-key-differences-between-small-language-models-and-large-language-models/>
- [2] IBM Think <https://www.ibm.com/think/topics/small-language-models>
- [3] Ataccama blog <https://www.ataccama.com/blog/small-language-models>
- [4] Salesforce blog <https://www.salesforce.com/blog/small-language-models/>
- [5] UNESCO <https://www.salesforce.com/blog/small-language-models/>
- [6] Hatchworks blog <https://hatchworks.com/blog/gen-ai/small-language-models/>
- [7] Mesh Digital Insights <https://insights.meshdigital.io/the-power-of-going-small-how-small-language-models-are-driving-competitive-advantage-in-ai/>
- [8] Capgemini Insights <https://www.capgemini.com/insights/expert-perspectives/small-is-the-new-big-the-rise-of-small-language-models/>
- [9] Deep Sense AI blog <https://deepsense.ai/implementing-small-language-models-slms-with-rag-on-embedded-devices-leading-to-cost-reduction-data-privacy-and-offline-use/#:~:text=efficient%20fine%2Dtuning,-Offline%20Functionality,larger%20LLM%20in%20a%20cloud.>
- [10] National Crowdfunding and Fintech Association (NCFA) Canada <https://ncfacanada.org/small-language-models-prioritize-privacy-and-efficiency/>
- [11] LinkedIn Article by Alex Velinov published in the “In AI We Trust” community <https://www.linkedin.com/pulse/less-more-future-small-language-models-alex-velinov-3felf/>
- [12] Arxiv-InkubaLM: A small language model for low-resource African languages <https://arxiv.org/html/2408.17024v1>
- [13] Arxiv-Shakti: A 2.5 billion parameter small language model optimized for Edge AI and low-resource environments <https://arxiv.org/html/2410.11331v1>
- [14] Analytics Vidhya blog <https://www.analyticsvidhya.com/blog/2023/12/a-step-by-step-guide-for-small-language-models-on-local-cpus/>
- [15] Data Science Dojo blog <https://datasciencedojo.com/blog/small-language-models-slms/>
- [16] ISG One research article <https://isg-one.com/research/articles/full-article/the-big-benefits-of-small-language-models>
- [17] Medium article by Arman Kamran <https://medium.com/@armankamran/slms-small-language-models-and-the-7-key-scenarios-where-they-surpass-llms-b548d73de85e#:~:text=Large%20models%20are%20often%20described,overkill%20for%20simpler%20NLP%20tasks>

[18] Foundation Educol <https://funeducol.org/2024/08/12/llms-vs-sllms-the-differences-in-large-small/>

[19] Aisera blog <https://aisera.com/blog/small-language-models/#:~:text=Limitations%20of%20Small%20Language%20Models&text=While%20the%20specialized%20focus%20of,a%20wide%20range%20of%20topics>

[20] University of Toronto-Schwartz Reisman Institute for Technology and Science <https://srinstitute.utoronto.ca/news/what-are-small-super-tiny-language-models>

[21] PSC Council https://www.pscouncil.org/_p/cr/p/Service_Contractor_Magazines/Fall_2024_Service_Contractor_Magazine/Trustworthy_and_Secure_AI_How_Small_Language_Models_Strengthen_Data_Security.aspx#:~:text=Their%20smaller%20size%20allows%20for,ethical%20standards%20and%20regulatory%20compliance.

[22] Aisera blog <https://aisera.com/blog/small-language-models/#:~:text=Benefits%20of%20Small%20Language%20Models&text=Unlike%20their%20larger%20counterparts%2C%20SLMs,relevant%20outputs%20for%20legal%20professionals.>

[23] PSC Council https://www.pscouncil.org/_p/cr/p/Service_Contractor_Magazines/Fall_2024_Service_Contractor_Magazine/Trustworthy_and_Secure_AI_How_Small_Language_Models_Strengthen_Data_Security.aspx

[24] Arxiv-SLM-Mod: Small language models surpass LLMs at content moderation <https://arxiv.org/html/2410.13155v1>

[25] Microsoft-blog <https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/>

[26] Archee AI blog <https://www.arcee.ai/blog/everything-you-need-to-know-about-small-language-models>

[27] Aisera blog <https://aisera.com/blog/small-language-models/>

[28] Arxiv-A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with LLMs, and trustworthiness <https://arxiv.org/html/2411.03350v1>

[29] Data Camp blog <https://www.datacamp.com/blog/small-language-models>

[30] Annotation box <https://annotationbox.com/small-language-models/>

[31] The Drum Publication <https://www.thedrum.com/news/2024/08/07/move-over-llms-why-microsoft-salesforce-others-are-developing-small-language-models>

[32] Coin Telegraph news publication <https://cointelegraph.com/news/david-v-goliath-small-language-models-challenge-big-techs-ai-giants>

-
- [33] ARS technical publication <https://arstechnica.com/information-technology/2024/04/apple-releases-eight-small-ai-language-models-aimed-at-on-device-use/>
- [34] Bayer <https://www.bayer.com/en/agriculture/article/genai-for-good>
- [35] Elder Research blog <https://www.elderresearch.com/blog/government-ai-starting-small-with-small-language-models-slms/>
- [36] Mark Tech Post publication <https://www.marktechpost.com/2024/10/28/llmware-introduces-model-depot-an-extensive-collection-of-small-language-models-slms-for-intel-pcs/>
- [37] Reworked <https://www.reworked.co/digital-workplace/the-small-language-model-advantage-in-todays-digital-workplace/>
- [38] LinkedIn article by Ramachandran Muhae <https://www.linkedin.com/pulse/empowering-edge-ai-small-language-models-challenges-ramachandran-muhae/>
- [39] The Wall Street Journal-Deloitte <https://deloitte.wsj.com/cio/small-language-models-bringing-generative-ai-to-the-edge-0db24f8d>
- [40] Forbes publication <https://www.forbes.com/councils/forbestechcouncil/2024/11/15/scaling-small-language-models-slms-for-edge-devices-a-new-frontier-in-ai/>
- [41] Sanket Daru blog <https://sanketdaru.com/blog/small-language-models-to-solve-big-business-problems/>
- [42] AI Business publication <https://aibusiness.com/nlp/3-most-common-problems-with-small-language-models>
- [43] Microsoft News publication <https://news.microsoft.com/source/features/ai/the-phi-3-small-language-models-with-big-potential/>
- [44] Arion Research <https://www.arionresearch.com/blog/what-you-need-to-know-about-small-and-narrow-language-models>
- [45] Sanger Institute blog <https://sangerinstitute.blog/2024/09/02/join-us-for-the-explainable-ai-in-biology-conference/#:~:text=In%20his%20talk%2C%20Darlington%20will,tool%20for%20advancing%20global%20research>
- [46] SLM as guardian: Pioneering AI safety with small language models <https://www.aimodels.fyi/papers/arxiv/slm-as-guardian-pioneering-ai-safety-small>
- [47] Small language models: The future of secure and cost-effective AI in the public sector <https://siliconangle.com/2024/06/28/small-language-models-cost-effective-ai-public-sector-awssummit/>
- [48] SLMs-A cheaper, greener route into AI <https://www.unesco.org/en/articles/small-language-models-slms-cheaper-greener-route-ai>
-