Argumentative Essay

Amogh Patankar

March 22, 2025

The future of enterprise solutions is set for an obvious shift in the ever-changing AI landscape: I argue that multimodal small language models (SLMs) under 20B parameters will replace traditional large language models (LLMs) in at least 80% of enterprise applications by 2028. I believe this to be true based on measurable trends in hardware improvements, algorithmic innovations in inference and development, and the evolving demands of industry-specific use cases.

LLMs are widely used in customer service, analytics, and natural language processing. However, scaling laws show that beyond a threshold, increasing model parameters reduces performance gains while increasing training and inference costs. Multimodal SLMs integrate text, images, audio, and video in a compact architecture tailored to specific tasks. Companies like Cohere are shifting to domain-specific models, which are more cost-effective and less prone to generating "hallucinations." Building on SLMs will continue, with improvements in hallucination reduction.

Additionally, cloud vendors will prioritize SLMs with fine-tuning capabilities. Hardware-aligned optimization reduces inference costs by about 90% through sparsity techniques. By 2027, over 80% of enterprise deployments will adopt SLMs, reducing redundancy and resource overhead. Advancements in hardware, such as improved GPUs and AI accelerators, will decrease energy and computational costs. SLMs like Microsoft's Phi-3, achieve comparable task accuracies to GPT-4 Turbo at a lower cost, reducing query costs from \$0.02 to \$0.0019.

Within the next few years, enterprises will enhance their capabilities in safeguarding language models, leading to a reduction in hallucinations. Retrieval-augmented SLMs, such as IBM WatsonX, have already demonstrated the potential to reduce errors to approximately 40% through the integration of knowledge bases. Consequently, an increasing number of companies will adopt SLMs with reduced hallucinations (IBM, 2024). These advancements will prompt cloud vendors like AWS and Azure to prioritize SLMs over LLMs, potentially transitioning towards SLM-as-a-service models.

Furthermore, algorithmic breakthroughs, like various quantization, pruning, and mixture-of-experts (MoE) techniques, have shown that inference can be performed using significantly less compute compared to LLMs. Studies on scaling laws leads us to believe that training costs are increasing, yet AI inference, requiring 1–2 FLOPs per parameter, can be drastically reduced through focused optimization. By combining these improvements with specialized fine-tuning, SLMs offer enhanced speed and reliability, tailored to meet the specific needs of enterprises.

One could argue that LLMs' have more advanced general-purpose capabilities that let it handle unstructured tasks and that enterprises benefit from LLMs' generalized knowledge. Yet, industry surveys suggest that over 60% of enterprise applications prioritize precision, regulatory compliance, and real-time performance over broad general intelligence, proving the need for smaller, domain-specific models. In sensitive sectors like finance, healthcare, and legal services, customized, smaller AI systems improve accuracy and reduce liability, aiding adoption.

Through the convergence of hardware innovations, algorithmic efficiency, and specialized solutions all support my argument that by 2027, multimodal SLMs will replace traditional LLMs in over 80% of enterprise applications. This shift will reduce costs, decrease resource consumption, and ensure that AI is more reliable and secure- qualities which are essential in today's competitive and regulated business environments.

Works Cited

Abdin, Marah, et al. "Phi-3 technical report: A highly capable language model locally on your phone." arXiv preprint arXiv:2404.14219 (2024).

Li, Beibin, et al. "Small language models for application interactions: A case study." arXiv preprint arXiv:2405.20347 (2024).

Marshall, Matt. "The Enterprise Verdict on AI Models: Why Open Source Will Win." VentureBeat, VentureBeat, 24 Oct. 2024, venturebeat.com/ai/the-enterprise-verdict-on-ai-models-why-open-source-will-win/.

Stauber, Doug/ "Watsonx.Governance 2.0- Here's What's New." Medium, IBM Data Science in Practice, 11 June 2024, medium.com/ibm-data-ai/watsonx-governance-2-0-heres-what-s-new-8cf0889109e1.