Navigating the landscape of language models with comparative insights - large language models (LLM) vs small language models (SLM)

Swetha Sistla
FinTech, USA

The performance and applications of Large Language Models and Smaller Language Models in NLP. LLMs, usually characterized by their enormously large number of parameters-running into billions and trillions of parameters-have revolutionized NLP by showing state-of-the-art performances in language generation, translation, summarization, sentiment analysis, and question answering. Large-scale training of LLMs on vast datasets has allowed them to learn very complex patterns of human languages and generate responses highly accurate and contextually relevant. However, given the enormous computational requirement in terms of large volumes of processing power, memory, and storage, LLMs are resource-intensive, slow to deploy, and difficult to scale, especially in environments with limited infrastructures.

In contrast, SLMs are designed to carry out NLP tasks with much fewer parameters and a lighter-weight architecture. While generally less accurate than larger models, SLMs have significant benefits due to their lower computational cost, faster inference time, and easier deployment. They are suitable for real-time use and resource-constrained environments such as mobile devices or edge computing. However, despite all these limitations, SLMs have gone through a radical development with techniques such as model pruning, quantization, and distillation that allow small models to preserve a substantial portion of their performance at a significantly lower size and lesser resource usage.

This work emphasizes the trade-offs between LLMs and SLMs regarding aspects such as model accuracy, computational efficiency, energy consumption, latency, and deployment flexibility. We further investigate recent advances in model compression and optimization techniques that allow SLMs to be competitive while avoiding the heavy resource costs of LLMs. These results are expected to help researchers and developers choose an appropriate model type with respect to task requirements, deployment scenarios, and available computational resources for a better trade-off between performance and practicability.

Biography

Swetha Sistla is a senior Lead Software Engineer and Technology Architect with experience of over 20 years, driving software systems transformation into FinTech. She is expertly versed in scalable system design, cloud migration, and real-time processing architecture using technologies such as Apache Kafka, Redis, MongoDB, and Microsoft Azure. Her leadership has played an instrumental role in the modernization of enterprise applications while enhancing performance and ensuring the delivery of secure, compliant solutions.

Swetha is a very energetic thought leader and author with a plethora of publications on Agile methodologies, cybersecurity, Generative AI, and microservices. She authors topics like responsible AI governance, the evolving large language model space, and leveraging AI to fuel enterprise innovation. Alongside her technical work, Swetha has deep engagement in fostering diversity and inclusion. She won the prestigious Americas DE&I Pacesetter Award for this, along with several other national recognitions for her mentorship activities.

An active contributor to the professional community, Swetha has led globally distributed teams that have enhanced their technical acumen to deliver results that make a difference. She holds a Bachelor of Technology and is certified in Generative AI, Python, and cloud technologies.

Swetha's passion lies in exploring the intersection of AI, cybersecurity, and scalable design to address modern enterprise challenges, making her a sought-after speaker and panelist at technology-focused conferences.

Received Date: November 25, 2024; Accepted Date: November 27, 2024; Published Date: January 03, 2025