**The Rise of Small Language Models**

As language models evolve to become more versatile and powerful, it seems that going small may be the best way to go.

Feb 16th, 2024 3:00am by [Kimberley Mok](https://thenewstack.io/author/kimberleymok/). <https://thenewstack.io/the-rise-of-small-language-models/>

The impressive power of [large language models](https://thenewstack.io/what-is-a-large-language-model/) (LLMs) has evolved substantially during the last couple of years. These versatile AI-powered tools are in fact deep learning artificial neural networks that are trained with massively large datasets, capable of leveraging billions of parameters (or machine learning variables) in order to perform various [natural language processing](https://thenewstack.io/recent-advances-deep-learning-natural-language-processing/) (NLP) tasks.

These can run the gamut from [generating, analyzing and classifying text](https://thenewstack.io/openais-gpt-4-can-analyze-visual-images-pass-bar-exam/), all the way to generating rather convincing [images from a text prompt](https://thenewstack.io/the-power-and-ethical-dilemma-of-ai-image-generation-models/), to translating content into different languages, or [chatbots that can hold human-like conversations.](https://thenewstack.io/large-language-models-arent-the-silver-bullet-for-conversational-ai/) Well-known LLMs include proprietary models like OpenAI’s [GPT-4](https://thenewstack.io/openais-gpt-4-can-analyze-visual-images-pass-bar-exam/), as well as a growing roster of [open source contenders](https://thenewstack.io/large-language-models-open-source-llms-in-2023/) like Meta’s [LLaMA](https://thenewstack.io/why-open-source-developers-are-using-llama-metas-ai-model/" \t "_blank).

But despite their considerable capabilities, LLMs can nevertheless present some significant disadvantages. Their sheer size often means that they require hefty computational resources and energy to run, which can preclude them from being used by smaller organizations that might not have the deep pockets to bankroll such operations. With larger models there is also the risk of [algorithmic bias](https://thenewstack.io/uncovering-biases-the-importance-of-data-diversity-in-speech-recognition/) being introduced via [datasets that are not sufficiently diverse](https://thenewstack.io/removing-bias-from-ai-is-a-human-endeavor/), leading to faulty or inaccurate outputs — or the dreaded “[hallucination](https://thenewstack.io/3-ways-to-stop-llm-hallucinations/)” as it’s called in the industry.

**How Small Language Models Stack up Next to LLMs**

These issues might be one of the many that are behind the recent rise of small language models or SLMs. These models are slimmed-down versions of their larger cousins, and for smaller enterprises with tighter budgets, SLMs are becoming a more attractive option, because they are generally easier to train, fine-tune and deploy, and also cheaper to run.

Small language models are essentially more streamlined versions of LLMs, in regards to the size of their neural networks, and simpler architectures. Compared to LLMs, SLMs have fewer parameters and don’t need as much data and time to be trained — think minutes or a few hours of training time, versus many hours to even days to train a LLM. Because of their smaller size, SLMs are therefore generally more efficient and more straightforward to implement on-site, or on smaller devices.

Moreover, because SLMs can be tailored to more narrow and specific applications, that makes them more practical for companies that require a language model that is trained on more limited datasets, and can be fine-tuned for a particular domain.

Additionally, SLMs can be customized to meet an organization’s specific requirements for security and privacy. Thanks to their smaller codebases, the relative simplicity of SLMs also reduces their vulnerability to malicious attacks by minimizing potential surfaces for security breaches.

**TRENDING STORIES**

1. [**What ChatGPT and Claude Can See on Your Screen**](https://thenewstack.io/what-chatgpt-and-claude-can-see-on-your-screen/)
2. [**How to Set up and Run a Local LLM with Ollama and Llama 2**](https://thenewstack.io/how-to-set-up-and-run-a-local-llm-with-ollama-and-llama-2/)
3. [**The Building Blocks of LLMs: Vectors, Tokens and Embeddings**](https://thenewstack.io/the-building-blocks-of-llms-vectors-tokens-and-embeddings/)
4. [**A Comprehensive Guide to Function Calling in LLMs**](https://thenewstack.io/a-comprehensive-guide-to-function-calling-in-llms/)
5. [**Semantic Router and Its Role in Designing Agentic Workflows**](https://thenewstack.io/semantic-router-and-its-role-in-designing-agentic-workflows/)

On the flip side, the increased efficiency and agility of SLMs may translate to slightly reduced language processing abilities, depending on the benchmarks the model is being measured against.

Nevertheless, some SLMs like Microsoft’s recently introduced 2.7 billion-parameter [Phi-2](https://dev-kit.io/blog/ai/microsoft-phi-2), demonstrate state-of-the-art performance in mathematical reasoning, common sense, language understanding, and logical reasoning that is remarkably comparable to — and in some cases, exceed — that of much heftier LLMs. According to Microsoft, the efficiency of the [transformer](https://thenewstack.io/how-to-get-the-right-vector-embeddings/)-based Phi-2 makes it an ideal choice for researchers who want to improve safety, interpretability and ethical development of AI models.

Other SLMs of note include:

* [DistilBERT](https://medium.com/huggingface/distilbert-8cf3380435b5): a lighter and faster version of Google’s [BERT](https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html) (Bidirectional Encoder Representations Transformer), the pioneering deep learning NLP AI model introduced back in 2018. There are also Mini, Small, Medium and Tiny versions of BERT, which are scaled-down and optimized for varying constraints, and range in size from 4.4 million parameters in the Mini, 14.5 million in the Tiny, to 41 million parameters in the Medium version. There is also MobileBERT, a version designed for mobile devices.
* [Orca 2](https://www.microsoft.com/en-us/research/blog/orca-2-teaching-small-language-models-how-to-reason/): Developed by Microsoft by fine-tuning Meta’s LLaMA 2 by using [synthetic data](https://www.techopedia.com/definition/33305/synthetic-data) that is generated from a statistical model, rather than from real life. This results in enhanced reasoning abilities, and higher performance in reasoning, reading comprehension, math problem solving and text summarization that can overtake that of larger models that are ten times larger.
* [GPT-Neo](https://github.com/EleutherAI/gpt-neo) and [GPT-J](https://huggingface.co/docs/transformers/model_doc/gptj): With 125 million and 6 billion parameters respectively, these alternatives were designed by the open source AI research consortium [EleutherAI](https://www.eleuther.ai/" \t "_blank) to be smaller and open source versions of OpenAI’s GPT model. These SLMs can be run on cheaper cloud computing resources from CoreWeave and TensorFlow Research Cloud.

Ultimately, the emergence of small language models signals a potential shift from expensive and resource-heavy LLMs to more streamlined and efficient language models, arguably making it easier for more businesses and organizations to adopt and tailor generative AI technology to their specific needs. As language models evolve to become more versatile and powerful, it seems that going small may be the best way to go.