**A Guide to Using Small Language Models for Business Applications**

<https://kili-technology.com/large-language-models-llms/a-guide-to-using-small-language-models>

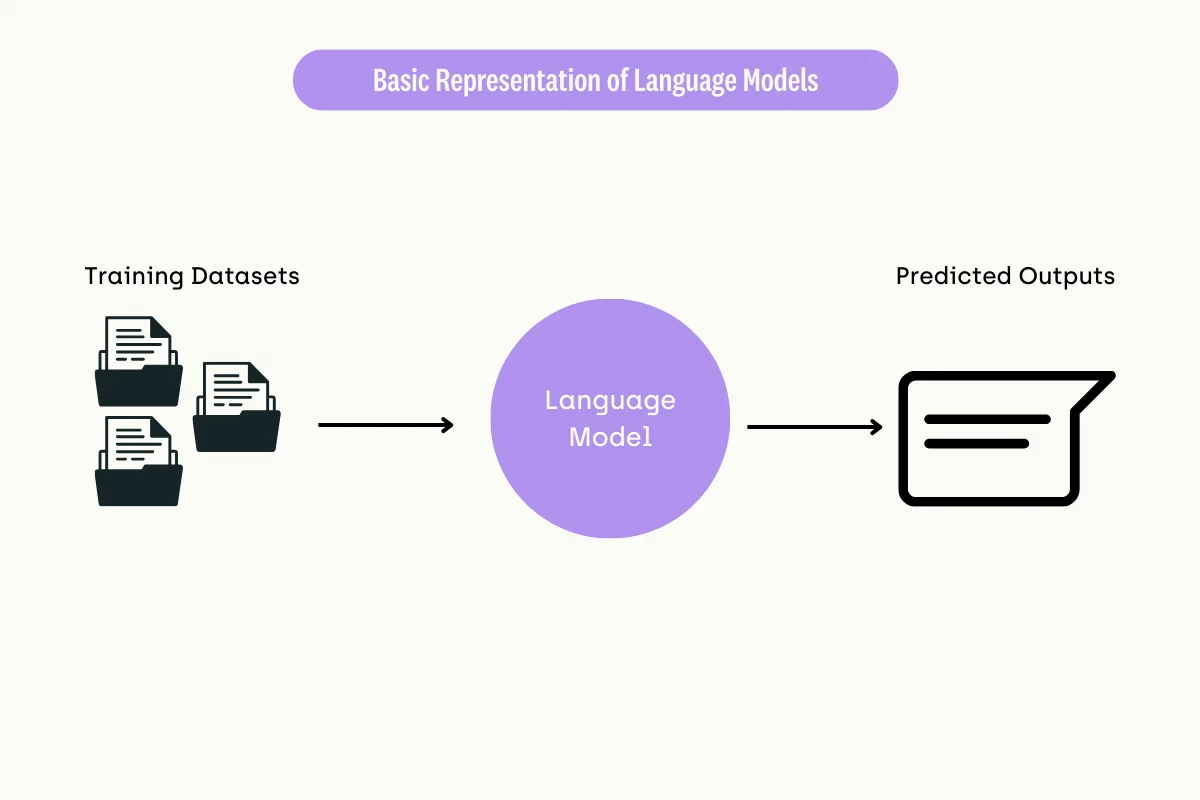
The successful implementation of SLMs requires careful planning, from dataset preparation to continuous evaluation and monitoring. Organizations must adopt a proactive approach to manage these models, ensuring they continually meet the evolving demands and maintain alignment with business goals.

Table of Contents

1. [The basics: What is a language model?](https://kili-technology.com/large-language-models-llms/a-guide-to-using-small-language-models#the-basics--what-is-a-language-model?)
2. [The arrival of the large language model](https://kili-technology.com/large-language-models-llms/a-guide-to-using-small-language-models#the-arrival-of-the-large-language-model)
3. [The small language model: what is it?](https://kili-technology.com/large-language-models-llms/a-guide-to-using-small-language-models#the-small-language-model--what-is-it?)
   * [The case for small language models](https://kili-technology.com/large-language-models-llms/a-guide-to-using-small-language-models#the-case-for-small-language-models)
   * [When should we use small language models?](https://kili-technology.com/large-language-models-llms/a-guide-to-using-small-language-models#when-should-we-use-small-language-models?)
4. [Potential use cases](https://kili-technology.com/large-language-models-llms/a-guide-to-using-small-language-models#potential-use-cases)
5. [Examples of small language models](https://kili-technology.com/large-language-models-llms/a-guide-to-using-small-language-models#examples-of-small-language-models)
   * [Smol LM](https://kili-technology.com/large-language-models-llms/a-guide-to-using-small-language-models#smol-lm)
   * [GPT-4o Mini](https://kili-technology.com/large-language-models-llms/a-guide-to-using-small-language-models#gpt-4o-mini)
   * [OpenELM Model by Apple](https://kili-technology.com/large-language-models-llms/a-guide-to-using-small-language-models#openelm-model-by-apple)
   * [Microsoft's Phi-3](https://kili-technology.com/large-language-models-llms/a-guide-to-using-small-language-models#microsoft-s-phi-3)
   * [Mistral 7B](https://kili-technology.com/large-language-models-llms/a-guide-to-using-small-language-models#mistral-7b)
   * [Google's Gemma 7B](https://kili-technology.com/large-language-models-llms/a-guide-to-using-small-language-models#google-s-gemma-7b)
6. [Leveraging small language models](https://kili-technology.com/large-language-models-llms/a-guide-to-using-small-language-models#leveraging-small-language-models)
   * [Fine-tuning small language models](https://kili-technology.com/large-language-models-llms/a-guide-to-using-small-language-models#fine-tuning-small-language-models)
   * [RAG for SLMs](https://kili-technology.com/large-language-models-llms/a-guide-to-using-small-language-models#rag-for-slms)
7. [What's next? Using small language models for strategic business applications](https://kili-technology.com/large-language-models-llms/a-guide-to-using-small-language-models#what-s-next?-using-small-language-models-for-strategic-business-applications)

The evolution of artificial intelligence (AI) in the business world has brought numerous tools and technologies into the spotlight, one of which is language models. These models of varying sizes and complexity have opened new avenues for enhancing business operations, customer interactions, and content generation. This guide focuses on understanding small language models, their advantages, practical use cases, and how businesses can leverage them efficiently.

**The basics: What is a language model?**

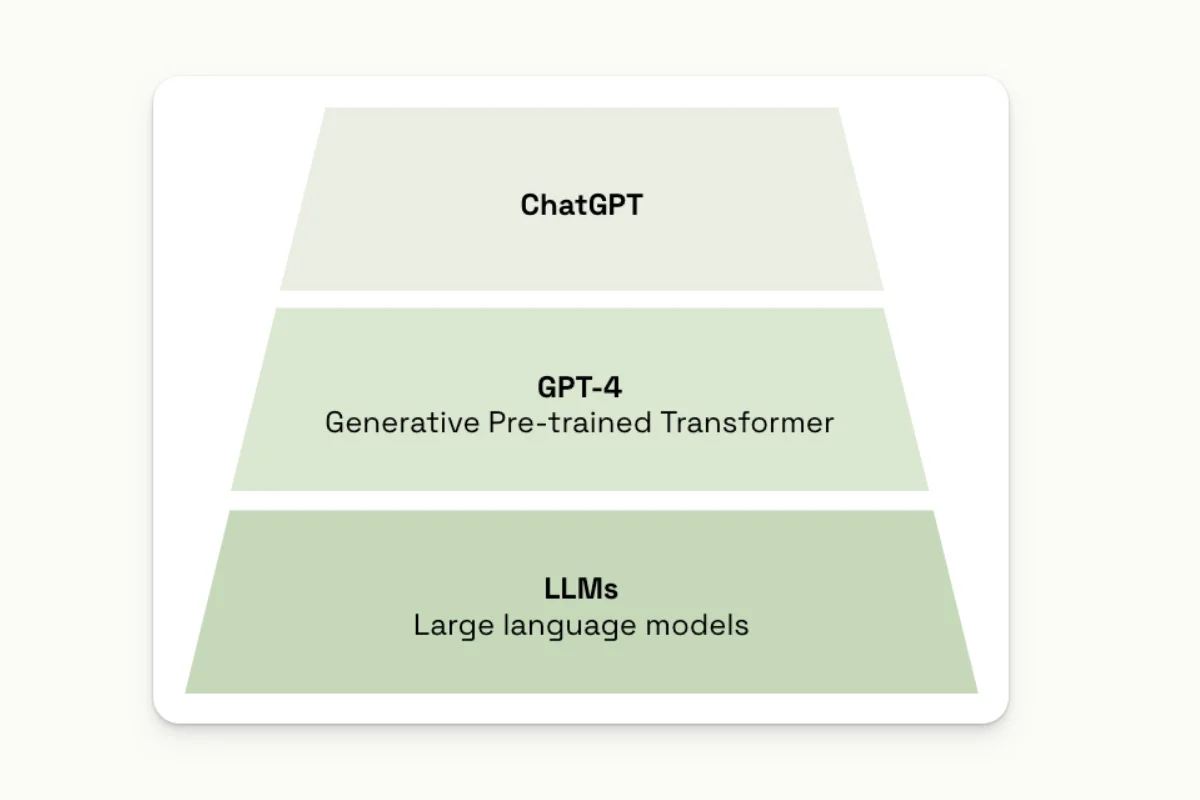


A [language model](https://kili-technology.com/data-labeling/nlp/) is an artificial intelligence that understands, generates, and manipulates human language. It works by predicting the likelihood of a sequence of words, helping it to generate coherent and contextually relevant text based on the input it receives.

Language models are typically trained on diverse and expansive datasets that include a broad range of texts such as books, articles, websites, and other digital content. This training enables the models to capture and learn language's structural patterns, nuances, and variability, which are essential for generating intelligent responses.

The core technology underpinning modern language models is neural networks, particularly types known as transformers. These neural networks are designed to process sequences of data (like text) and learn contextual relationships between words over long distances, enhancing the model's ability to generate coherent and contextually appropriate responses.

The arrival of the large language model



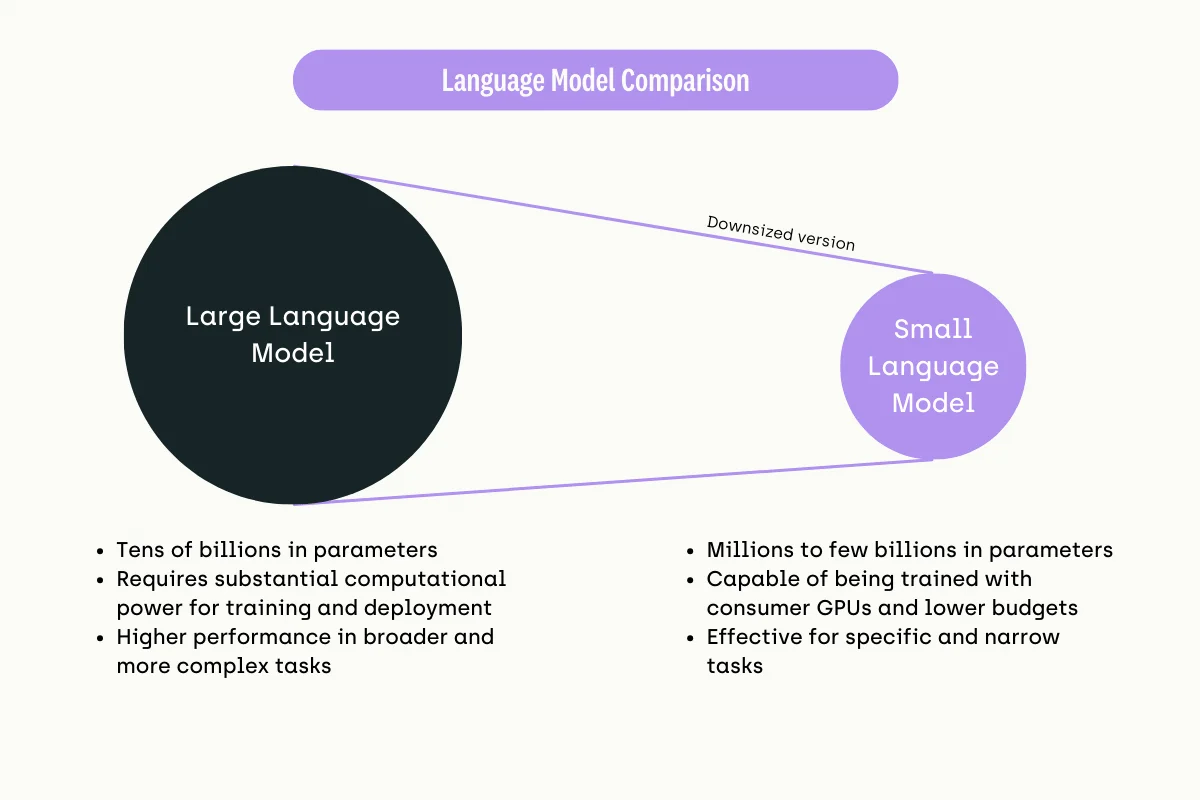
In recent years,[large language models (LLMs)](https://kili-technology.com/large-language-models-llms/) like GPT-3, BERT, LaMDA, Llama, and Mistral have gained prominence, driven by significant advances in deep learning technologies, the availability of massive datasets, and increased computational power. The public introduction of models such as ChatGPT in 2022 brought LLMs to widespread attention, highlighting their capacity to handle diverse applications—ranging from machine translation and text summarization to creating conversational agents like chatbots.

These models are constructed using billions of parameters—weights within the model that are adjusted during the training process to minimize errors in prediction. The large number of parameters allows these models to capture complex patterns and perform a wide variety of language-based tasks with high proficiency.

However, the scale and complexity of LLMs require substantial computational resources, which can pose accessibility challenges for many businesses. The extensive training necessary to develop these models often demands high-powered GPUs and significant electricity consumption, leading to greater operational costs and a larger carbon footprint.

Consequently, while LLMs are powerful tools for handling complex and diverse language tasks, their deployment is generally more suited to organizations that can meet these high resource demands.

The small language model: what is it?



Small language models are scaled-down versions of their larger counterparts. They are designed to offer a more accessible and resource-efficient option for organizations that require natural language processing capabilities but do not need or cannot support the heavyweight infrastructure that large models demand.

The key difference between small language models (SLMs) and large language models (LLMs) lies in the number of parameters they possess:

* SLMs typically have a parameter count ranging from thousands to a few million. They have a more modest scale compared to LLMs.
* In contrast, LLMs have a significantly higher number of parameters, often ranging from millions to billions or even trillions. For example, GPT-3 has 175 billion parameters, while GPT-4 has an even larger parameter count.

Parameters refer to the components of the model that are learned from the training data. They are the aspects of the model that are adjusted through the learning process to better predict or generate outcomes.

The large number of parameters in LLMs allows them to capture more complex patterns and relationships in language, enabling better performance on a wide range of natural language processing tasks. However, this comes at the cost of higher computational requirements for training and deploying these models.

SLMs, with their smaller parameter count, are more efficient and accessible, particularly for organizations with limited computational resources or budgets. They can still perform well on specific tasks while being faster to train and deploy than their larger counterparts.

The case for small language models

Small language models have become increasingly popular as businesses seek efficient, cost-effective solutions for incorporating artificial intelligence into their operations. These models offer several advantages:

* **Cost-effectiveness**: Small language models require significantly less computational power due to their fewer parameters and simpler architectures. This translates into lower hardware expenses. They are also cheaper to set up and maintain. With less demand for computational resources, ongoing operational costs such as data storage, processing power, and cooling are minimized.
* **Agility**: Their smaller size allows for quicker iterations and deployment cycles. Due to their smaller size, the training and tuning of small language models are considerably quicker. This allows for more rapid experimentation and iteration, which is crucial in a business environment where adapting to new data or changing requirements quickly can be a competitive advantage.
* **Simplicity**: Small language models are easier to integrate into existing business systems without significant infrastructure changes. Due to their smaller size and less demanding infrastructure needs, companies can add AI capabilities without overhauling their current IT setup.

Businesses with limited AI expertise can also more readily adopt small language models compared to large language models. Their simplicity makes them more approachable for teams with less specialized knowledge in machine learning. Small models can be quickly retrained or adapted to new tasks and datasets, making them particularly suitable for dynamic industries where consumer preferences and business needs evolve rapidly.

When should we use small language models?

Small language models (SLMs) are particularly well-suited for specific business contexts where efficiency, cost, and data privacy are major considerations. Here’s a detailed look at when and why businesses should consider deploying small language models:

**Specific, limited-scope language tasks**

* **Targeted Applications**: SLMs excel in scenarios where the language processing needs are specific and well-defined, such as parsing legal documents, analyzing customer feedback, or handling domain-specific queries.
* **Customization**: Due to their size, SLMs can be more easily customized and fine-tuned on niche datasets, allowing businesses to achieve high accuracy in specialized tasks without the overhead of larger models.

**Fast and efficient language processing**

* **Real-time Interactions**: Businesses requiring real-time language processing—such as interactive chatbots, instant customer support, or live translation services—benefit from the faster response times of SLMs.
* **Lower Latency**: Since SLMs can be deployed on local servers or even on edge devices, they reduce the latency associated with cloud processing, making them ideal for applications where speed is critical.

**Sensitivity to operational costs and resource usage**

* **Reduced Infrastructure Requirements**: Companies mindful of the costs associated with powerful computing resources find SLMs attractive because they require significantly less computational power and can run on less specialized hardware.
* **Economic Efficiency**: SLMs can drastically reduce the costs related to data processing and storage, making them a cost-effective option for startups and medium-sized businesses or those with limited IT budgets.

**High value on data privacy**

* **Local Data Processing**: For industries such as healthcare and finance where data privacy is paramount, SLMs offer an advantage as they can be fully operated in-house. Being able to process data locally ensures that sensitive data does not leave the organizational boundary.
* **Compliance with Regulations**: Organizations that are obligated to comply with strict data protection regulations (like GDPR or HIPAA) might prefer SLMs as they can help maintain control over data handling and storage practices.

Potential use cases

A small language model may excel at specific tasks such as multi-step reasoning, commonsense reasoning, or language understanding. Depending on the task the language model excels at, businesses can look into the following use cases:

1. **Customer Service Automation**: SLMs can [manage complex customer queries](https://kili-technology.com/data-labeling/customer-story-how-covea-leveraged-kili-technology-to-transform-customer-satisfaction-analysis) that require understanding context and multiple interaction steps. For example, in troubleshooting tech products or services, an SLM can guide a customer through a series of diagnostic steps based on the problem described. Additionally, because small language models have less computational requirements, they can process with lower latency which means a smoother customer experience.
2. **Home Automation Systems**: Small language models can power voice-activated assistants that control smart home devices. Basic commonsense reasoning and human language understanding are essential for executing user commands in context and reacting appropriately, which SLMs are just as capable of doing as LLMs.
3. **Virtual assistants:**Small language models can further automate more menial tasks in several industries. In industries that require a lot of filing, such as the finance, legal, and medical domains, an SLM can perform multiple simple tasks, such as classifying documents and extracting entities to create a summary or to fill in a database.
4. **Content Creation**: SLMs can assist in content creation for blogs, reports, and marketing materials by understanding the thematic elements of the requested content and generating contextually relevant text.
5. **Code assistance**: Some SLMs in the market have also been trained to generate code. For example, Microsoft's Phi-2 has been fed some programming languages. Though writing software code effectively is still out of reach for even the most popular large language models, small language models can still perform well enough to act as assistant coders for developers.

Examples of small language models

Smol LM

[Smol LM](https://kili-technology.com/large-language-models-llms/what-is-smollm-a-guide-to-hugging-face-s-small-language-model) is a highly efficient open-sourced language model developed by Hugging Face, tailored for scenarios that require lightweight models with fast response times. It's designed to be easily integrated into various applications where space and speed are critical, such as mobile apps or real-time systems.

It offers a streamlined version of larger models, providing sufficient linguistic capabilities for a range of tasks like summarization, sentiment analysis, or simple conversational interfaces.

Its compact architecture makes it ideal for businesses that need quick deployment and scalability while keeping resource consumption to a minimum.

GPT-4o Mini

[GPT-4o Mini](https://kili-technology.com/large-language-models-llms/a-guide-to-gpt4o-mini-openai-s-smaller-more-efficient-language-model) is a compact yet powerful version of the GPT-4 series, designed to bring the benefits of advanced language modeling to environments where computational resources are limited.

Despite its smaller size, GPT-4o Mini retains a significant portion of the capabilities of its larger counterparts, making it an excellent choice for businesses looking to implement natural language processing tasks like content generation, customer support automation, or data analysis without the heavy infrastructure typically required for larger models.

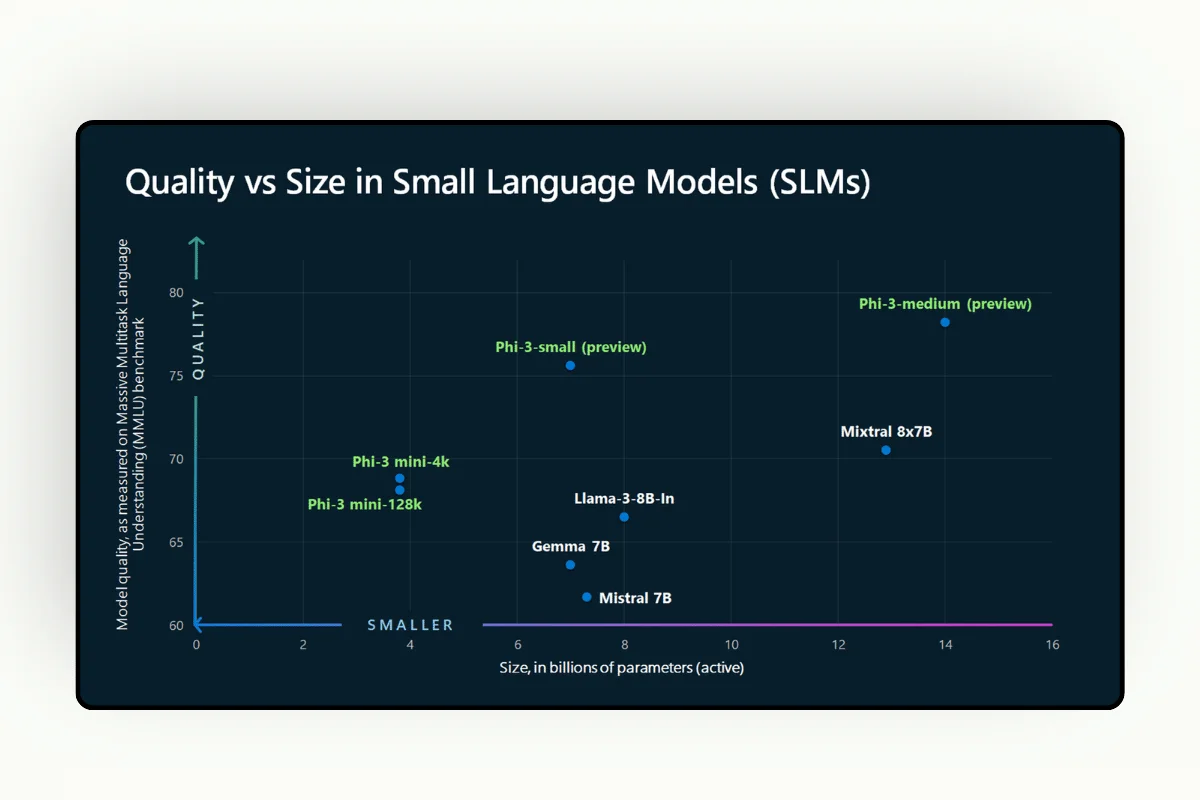
It strikes a balance between efficiency and performance, offering a cost-effective solution for organizations looking to leverage AI without the need for extensive hardware investments.

OpenELM Model by Apple

The [OpenELM model by Apple](https://machinelearning.apple.com/research/openelm" \t "_self), hosted on Hugging Face, is an efficient language model family designed to be openly available for training and inference. The models vary from 270 million to 3 billion parameters, providing different performance levels for various tasks.

They are pretrained on a dataset that includes multiple sources, amounting to about 1.8 trillion tokens. OpenELM is designed for efficiency and scalability, utilizing a layer-wise scaling strategy in transformer models. It's available for different applications, including those that require instruction tuning.

Microsoft's Phi-3

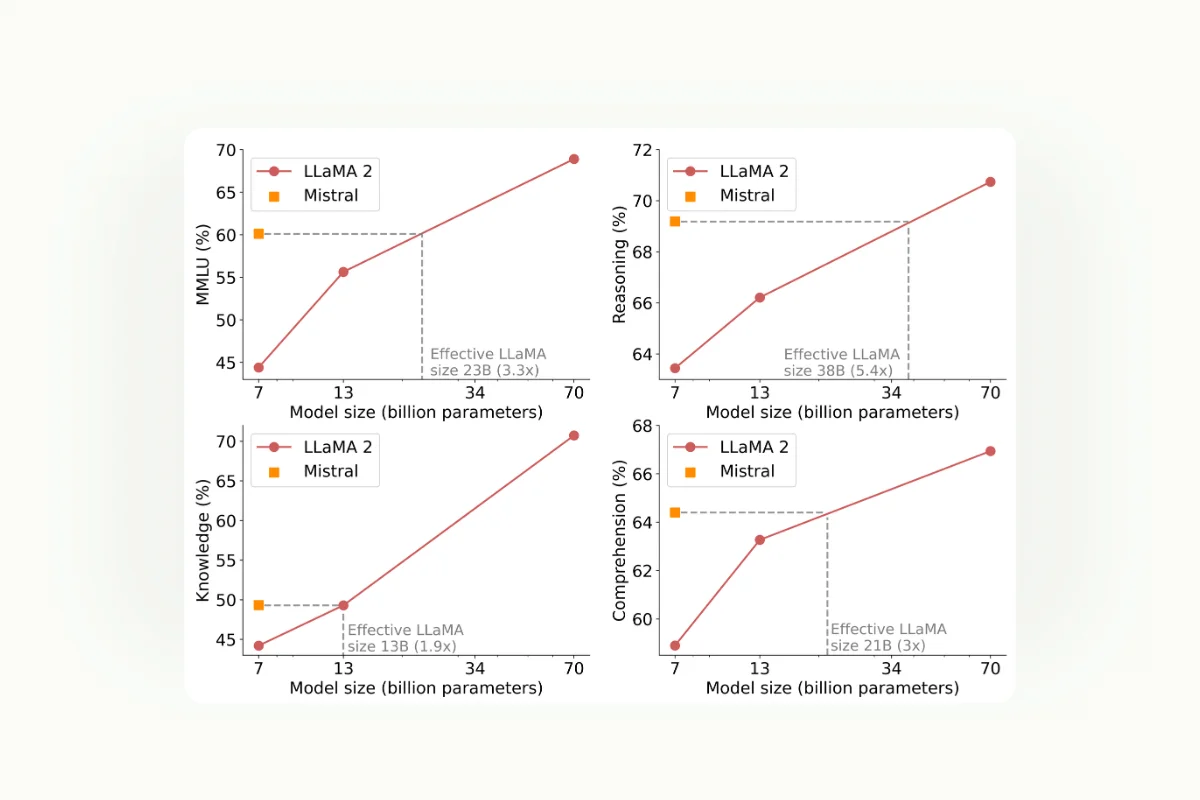


[Microsoft's Phi-3 model](https://news.microsoft.com/source/features/ai/the-phi-3-small-language-models-with-big-potential/)represents a significant advancement in the field of small language models (SLMs). The smallest model in the family is Phi-3-mini. Phi-3-mini has 3.8 billion parameters and was trained on 3.3 trillion tokens.

Despite its size, Phi-3-mini's performance is comparable to larger models like Mixtral 8x7B and GPT-3.5. It achieves a 69% score on the MMLU benchmark, which tests a model's understanding across a broad range of language tasks, and an 8.38 score on the MT-bench, which focuses on machine translation. These scores indicate that Phi-3-mini performs well in both general language understanding and specific tasks like translation.

One of the key innovations of the Phi-3 model is its ability to operate efficiently on smaller devices, like mobile devices, without the need to connect to the cloud. This capability minimizes latency and maximizes privacy, making it suitable for applications in remote or mobile settings where internet connectivity is limited or non-existent.​

Mistral 7B

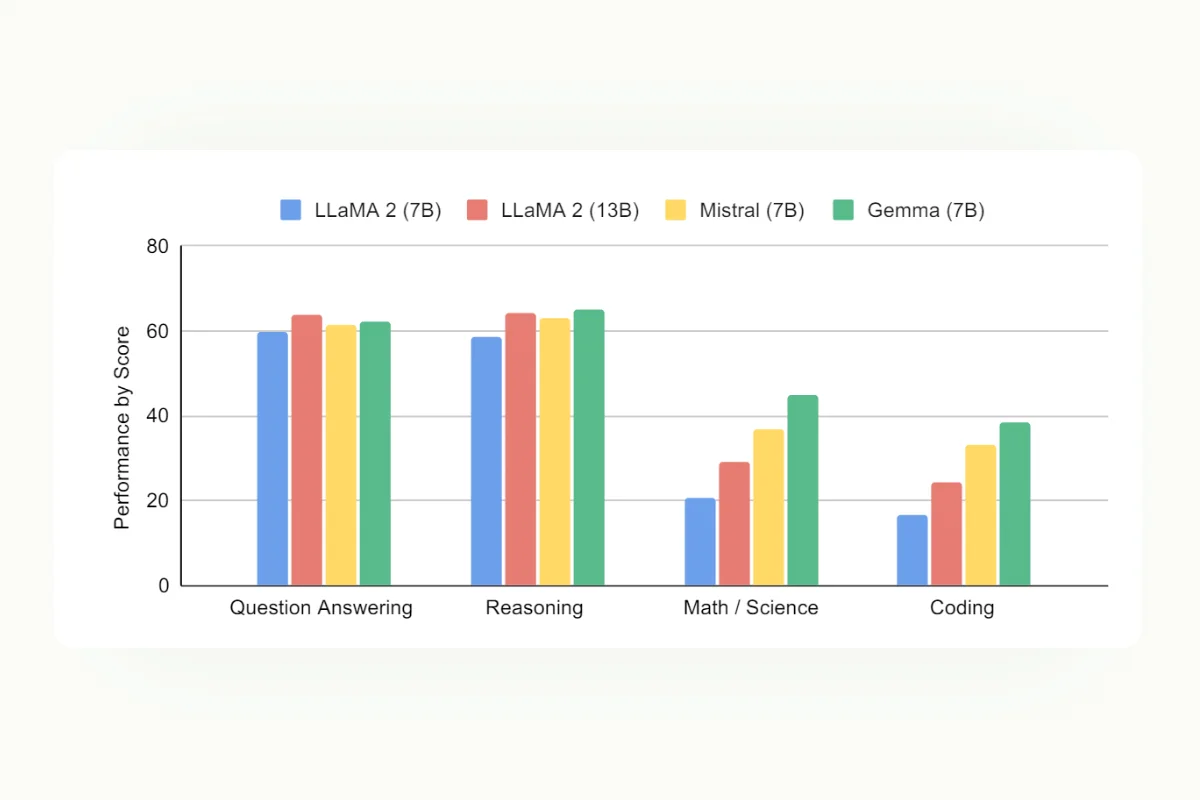


[Mistral 7B](https://mistral.ai/news/announcing-mistral-7b/) is a high-performance language model developed by Mistral AI, notable for its efficiency and the powerful results it delivers compared to its size. It is a 7-billion-parameter model optimized for a range of tasks including reasoning, math, and code generation.

It uses a decoder-only Transformer architecture, equipped with innovative features like Sliding Window Attention and Grouped Query Attention, which enhance both its speed and effectiveness.

In benchmarks, Mistral 7B has shown superior performance across various domains, including common sense reasoning, world knowledge, and specialized tasks like code generation and math, outperforming comparable models like Llama 2 13B and approaching the performance of some larger models.

Google's Gemma 7B



[Gemma 7B](https://huggingface.co/google/gemma-7b) is part of a new family of Google models designed to be lightweight yet deliver state-of-the-art performance. This model has achieved notable success in benchmarks such as MMLU, GSM8K, and HumanEval, excelling in language understanding, math problem-solving, and code generation tasks.

Gemma 7B uses advanced training techniques and is optimized for performance with features like parameter-efficient tuning and compatibility with multiple backends, enabling fast execution on a wide range of hardware, including consumer-grade GPUs.

Leveraging small language models

As we mentioned earlier, SLMs provide the most value after being fine-tuned to perform well on more specific, narrow tasks such as classification, summarization, content creation, and code assistance. Here are two ways to leverage small language models:

Fine-tuning small language models

Fine-tuning SLMs to turn them into [domain-specific language models](https://kili-technology.com/large-language-models-llms/building-domain-specific-llms-examples-and-techniques) or capable small language models for narrow and specific tasks can be much easier than doing so with large language models.

Given these models' reduced complexity and smaller size, aligning their capabilities with specific business needs or niche tasks can be accomplished more efficiently. This is particularly true when the training involves proprietary datasets designed to enhance the model's performance in targeted areas.

Fine-tuning smaller models can be considerably less expensive regarding computational resources and considering the volume of data required.

For instance, [in our approach to fine-tuning](https://kili-technology.com/large-language-models-llms/webinar-recap-fine-tuning-llm-deep-dive-and-demos), we recommend starting with a minimal amount of high-quality data to establish a baseline. This method mirrors successful strategies employed in the industry, such as Meta's [targeted data curation,](https://kili-technology.com/large-language-models-llms/llama-3-guide-everything-you-need-to-know-about-meta-s-new-model-and-its-data) which emphasizes the importance of an optimal data mix over sheer volume.

An illustrative case is when we fine-tuned the Llama 3 8B model with just 800 data points for a financial analysis task. This small dataset was specifically chosen to train the AI to deliver precise, clear, consistent responses in a defined format and tone.

Watch video

[Watch our LLM Fine-Tuning Webinar](https://resources.kili-technology.com/webinar/using-llms-to-build-custom-conversational-engines-0-0)

The initial results were promising, showing that even a limited amount of carefully selected data could significantly enhance the model’s output for specialized tasks. Expanding the dataset could be considered as needed based on the model's performance and the complexity of the task at hand.

This fine-tuning strategy reduces initial costs and allows for agile adjustments and enhancements, proving particularly effective for small language models tasked with specific functions.

RAG for SLMs

Just like large language models (LLMs), small language models can effectively leverage retrieval-augmented generation (RAG) to enhance their specialized capabilities. RAG is a powerful technique that bolsters the model’s responses by integrating external information. It operates by retrieving pertinent documents or data from an extensive database and then utilizing this retrieved content to enrich the generation process of the model. This integration allows the model to produce responses that are not only contextually relevant but also factually accurate and up-to-date.

In the case of small language models, pairing them with RAG can imbue them with additional domain-specific expertise, making them particularly useful for targeted applications.

However, the effectiveness of this combination heavily relies on the quality and relevance of the information sources RAG accesses. It is crucial for organizations to use credible, well-maintained data sources that align with their specific needs.

It is also important that SLMs be aligned and combined with RAG for a specific narrow task; otherwise, they can suffer from what is called an "alignment tax."

In [a study by Anthropic](https://arxiv.org/abs/2204.05862), it was found that large language models can receive an alignment bonus, wherein model performance improves after RLHF. In the meantime, small language models suffered from an alignment tax, which means the model's performance was reduced after RLHF.

Organizations should implement a[robust evaluation and monitoring strategy](https://kili-technology.com/large-language-models-llms/a-guide-to-rag-evaluation-and-monitoring-2024)to maximize the effectiveness of small language models paired with RAG. This strategy is essential to ensure that the SLM and RAG stack produces the desired outputs and meets specific operational requirements.

Pro Tip 💡

Check out our RAG Evaluation and Monitoring guide to learn more about how you can ensure your language model and RAG implementations are performing well.

Read the blog

Regular assessments allow organizations to gauge the performance of these models in real-world applications, identifying any discrepancies or areas where the outputs may not align perfectly with expected outcomes.

By continuously monitoring these systems, organizations can quickly identify when the models need tuning or realignment. This approach not only helps maintain the accuracy and relevance of the outputs but also ensures that the models remain effective as the nature of the tasks or the surrounding data environment evolves.

Such proactive evaluation and adjustment are crucial for sustaining the utility and efficiency of small language models enhanced by RAG, particularly in specialized or dynamic application areas.

What's next? Using small language models for strategic business applications

SLMs represent a significant technological advancement by making AI accessible to a broader range of businesses, especially those with limited computational resources or specific, narrow operational needs. They provide a cost-effective, agile, and simpler alternative to large language models (LLMs), enabling businesses to leverage AI for targeted tasks without the hefty investment in infrastructure typically associated with LLMs.

The efficiency and customizability of SLMs allow for rapid deployment and fine-tuning, making them ideal for businesses that prioritize quick adaptation to market changes and data security. Whether it's enhancing customer service, streamlining operations, or generating content, these tailored language models can significantly boost productivity and competitiveness.

However, the successful implementation of SLMs requires careful planning, from dataset preparation to continuous evaluation and monitoring. Organizations must adopt a proactive approach to manage these models, ensuring they continually meet the evolving demands and maintain alignment with business goals.

Addressing the challenge of constructing and managing a high-quality training data—whether for training, fine-tuning, evaluating, or monitoring—is the initial and a critical step in creating a truly effective AI application.

Building a proprietary dataset tailored to the specific use cases we've explored can be challenging, even for seasoned data science teams. Kili Technology provides a comprehensive solution for assembling high-quality datasets that can accelerate your development efforts. This includes a state-of-the-art tool renowned for its speed, flexibility, and quality, along with dedicated project management to ensure seamless execution, and a broad network of experts and data labelers who uphold the highest quality standards.

Our expert team is eager to discuss your specific needs and help identify the most effective strategies for your AI projects. We recognize the pivotal role of a dependable and meticulously curated dataset in the success of AI initiatives, and our services are designed to meet these essential needs efficiently. Whether you are just starting to explore AI possibilities or seeking to enhance an existing system, our tailored solutions are here to provide the support you need to achieve your objectives.