**5 Small Language Models Examples Boosting Business Efficiency**

<https://www.netguru.com/blog/small-language-models-examples>

**Summary**

The article explores the rising significance of Small Language Models (SLMs) in enhancing business efficiency. While Large Language Models (LLMs) like GPT-3 and GPT-4 have garnered significant attention for their expansive capabilities, SLMs offer a domain-specific alternative that is more efficient and cost-effective. The author presents five promising examples of SLMs, detailing their unique features and potential business applications.

**Introduction**

* **Shift from LLMs to SLMs:** Businesses are recognizing that while LLMs are powerful, they may not always be the most practical solution for domain-specific tasks due to their size, cost, and resource requirements.
* **Advantages of SLMs:** SLMs are gaining traction because they are more efficient, less costly to implement, and better suited for specialized tasks.

**What Are Small Language Models (SLMs)?**

* **Definition:** SLMs are artificial intelligence models based on natural language processing (NLP) with fewer parameters than LLMs.
* **Capabilities:** They can comprehend and mimic human language patterns but are less capable of processing and generating complex text compared to larger models.
* **Ideal Tasks:** SLMs excel in specific, less complex tasks such as text classification, sentiment analysis, clustering, tagging, and information extraction.
* **Business Use Cases:** Suitable for applications that don't require complex analysis but need efficient processing, like automating customer feedback analysis or generating basic content.

**SLMs vs. LLMs**

* **Domain-Specific vs. Generalist:**
  + **SLMs:** Great for niche, domain-specific tasks; can be trained with specialist terminology to provide expert, granular information.
  + **LLMs:** Operate on vast datasets across various disciplines, making them generalists.
* **Deployment and Accessibility:**
  + **SLMs:** Compact and can be deployed on smartphones, tablets, and edge devices; require less computational power.
  + **LLMs:** Require significant computational resources and specialized hardware; not easily deployable on smaller devices.
* **Security and Privacy:**
  + **SLMs:** Operate locally, reducing the risk of data breaches and enhancing privacy.
  + **LLMs:** Often process data on the cloud, making them more susceptible to security risks.

**Comparison Table**

| **Aspect** | **Small Language Models (SLMs)** | **Large Language Models (LLMs)** |
| --- | --- | --- |
| **Definition** | Millions to few billions of parameters | Dozens or hundreds of billions of parameters |
| **Performance** | Good for simpler tasks | Excels in complex language understanding |
| **Training Resources** | Less computational power required | Significant computational resources needed |
| **Training Time** | Shorter and less costly | Time-consuming and expensive |
| **Use Cases** | Embedded systems, mobile apps, local tasks | Advanced conversational agents, complex systems |
| **Deployment** | Easier on various platforms | Requires specialized hardware or cloud servers |
| **Adaptability** | Quick fine-tuning for specific domains | Flexible but requires significant effort to tune |
| **Environmental Impact** | Lower energy consumption and carbon footprint | Higher energy consumption and carbon footprint |
| **Security** | Lower risk due to local deployment | Higher risk due to cloud dependency |

**Examples of Small Language Models**

**1. PHI-3 – Tiny but Mighty**

* **Developer:** Microsoft (announced in April 2024).
* **Model Details:**
  + **Phi-3-mini:** The first in the PHI-3 family with 3.8 billion parameters.
  + Claimed to outperform models twice its size.
* **Capabilities:**
  + Strong logic and reasoning abilities.
  + Potential to summarize long, domain-specific documents.
  + Can be used to set up chatbots that access CRM records.
  + Generates marketing collateral like social media posts.
* **Safety and Ethics:**
  + Built following Microsoft's Responsible AI standards.
  + Emphasizes privacy, security, reliability, and inclusiveness.

**2. Mixtral of Experts – Advanced Reasoning**

* **Architecture:**
  + Decoder-only model using a Mixture-of-Experts (MoE) approach.
  + Utilizes a router neural network to select the best "experts" for processing each token.
* **Model Details:**
  + Total of 46.7 billion parameters but only uses 12.9 billion per token.
* **Advantages:**
  + Handles complex tasks efficiently and cost-effectively.
  + Comparable performance to ChatGPT 3.5.
  + Can run on local machines, reducing dependency on powerful hardware.
* **Capabilities:**
  + Leverages diverse knowledge across various domains.
  + Unique architecture allows it to act like an LLM with fewer resources.

**3. Llama 3 – Powerful and Accessible**

* **Developer:** Meta (Facebook).
* **Model Details:**
  + Trained on a dataset seven times larger than Llama 2.
  + Contains four times more code than its predecessor.
  + Can process text up to 8,000 tokens long.
* **Capabilities:**
  + Enhanced reasoning abilities.
  + Top-tier performance on industry benchmarks.
* **Integration:**
  + Used as Meta AI across Facebook, Instagram, WhatsApp, and Messenger.
  + Allows users to retrieve real-time information without switching apps.
* **Objective:**
  + To promote the next wave of AI innovation in applications, tools, and optimizations.

**4. DeepSeek-Coder-V2 – Your AI Developer**

* **Purpose:** Advanced AI tool for code generation and mathematical reasoning.
* **Model Details:**
  + Built using the Mixture-of-Experts (MoE) technique.
  + Pre-trained with 6 trillion tokens.
  + Supports 338 languages.
  + Context length of 128k tokens.
* **Performance:**
  + Achieves performance rates similar to GPT-4 Turbo.
  + Highest accuracy on HumanEval with a 90.2% success rate.
* **Ideal Use Cases:**
  + Businesses needing top-level analytical capabilities.
  + Situations where code sharing via cloud services is a concern.
* **Advantages:**
  + Can replace tools like Gemini Code or Copilot.
  + Enhances development productivity while maintaining data security.

**5. MiniCPM-Llama3-V 2.5 – Multimodal Excellence**

* **Model Details:**
  + Latest in the MiniCPM-V series with 8 billion parameters.
  + Built on SigLip-400M and Llama3-8B-Instruct.
* **Performance:**
  + Scored 65.1 points on OpenCompass, outperforming larger models.
  + Low hallucination rate of 10.3% on Object HalBench.
* **Capabilities:**
  + Excels in Optical Character Recognition (OCR).
  + Processes images up to 1.8 million pixels with any aspect ratio.
  + Supports over 30 languages, including English, German, French, Korean, and Japanese.
* **Use Cases:**
  + Digitization of images and handwritten text.
  + Data extraction from structured and unstructured data.
  + Mobile deployment ensures data stays on the device, enhancing security.

**General Business Use Cases for SLMs**

1. **Retrieving Information:**
   * Efficient searching and summarizing to find relevant data within large text volumes.
2. **Tagging and Clustering:**
   * Converting text into metadata.
   * Tagging comments or product opinions.
   * Clustering data to gain actionable insights.
3. **Creating JSON Data:**
   * Transforming text into structured formats compatible with API systems.

**Implementing SLMs in Your Business**

* **Media Note Conversion:**
  + Transform media notes into API calls with tags.
  + Automate information gathering and analysis.
  + Useful for staying updated with industry news and financial analysis.
* **Customer Feedback Analysis:**
  + Tag and cluster product opinions and comments.
  + Leverage insights to improve products based on user feedback.
* **Competitive Analysis:**
  + Examine competitors' offerings by extracting features from their web pages.
  + Cluster and analyze data to enhance your own products.

**Conclusion**

* **Cost and Efficiency Benefits:**
  + SLMs are less expensive to train and deploy compared to LLMs.
  + Require fewer computational resources, making them accessible on various devices.
* **Domain Specificity:**
  + Can be trained on specialized data, making them highly effective for niche tasks.
* **Performance and Accuracy:**
  + Despite having fewer parameters, SLMs maintain high performance and output accuracy in domain-specific applications.
* **Practical Solution for Businesses:**
  + Offer a cost-effective alternative without sacrificing quality.
  + Enhance operational efficiency across various industries.

**Key Takeaways for Businesses:**

* **Adopt SLMs for Specialized Tasks:**
  + Ideal for tasks that require domain expertise and precision.
* **Leverage SLMs for Enhanced Security:**
  + Local deployment reduces the risk of data breaches.
* **Improve Efficiency and Reduce Costs:**
  + Lower resource requirements translate to cost savings.
* **Stay Competitive:**
  + Utilize SLMs to gain insights from data, improve products, and stay ahead in the market.

**Final Thoughts**

As an experienced AI and ML engineer, I recognize that Small Language Models represent a significant advancement in making AI more accessible and practical for businesses. By focusing on domain-specific tasks, SLMs enable organizations to harness the power of AI without the substantial investment typically associated with Large Language Models. They offer a balanced approach, providing high performance and accuracy while ensuring efficiency and cost-effectiveness.

SLMs are particularly advantageous for:

* **Data Privacy:** Local processing enhances security, which is crucial for industries handling sensitive information.
* **Resource-Constrained Environments:** Their ability to run on devices with limited computational power expands the possibilities for AI deployment.
* **Customized Solutions:** Tailoring models to specific business needs leads to more relevant and actionable outputs.

In conclusion, Small Language Models are poised to play a critical role in the next wave of AI-driven business innovation, offering practical solutions that align with organizational goals and resource capacities.

**Full Article**

**For a long time, everyone talked about the capabilities of large language models. And while they’re truly powerful, some use cases call for a more domain-specific alternative.**

That’s where businesses can turn to small language models. They’re currently gaining traction, and for good reason. Not only are they more efficient but also less costly to implement. In this article, I share some of the most promising examples of small language models on the market. I also explain what makes them unique, and what scenarios you could use them for.

**What are Small Language Models (SLMs)?**

Language models are [tools based on artificial intelligence](https://www.netguru.com/blog/process-optimization-ai-assistants) and natural language processing.

They can comprehend and mimic human language based on the patterns and training data they’re given.

[Small language models](https://www.netguru.com/blog/small-language-models) are less capable of processing and generating text as they have fewer parameters as opposed to larger models. This means they’re better at handling less complex tasks, which are more specific, like text classification, sentiment analysis, and basic text generation. These models are ideal for business use cases that don't require complex analysis. They are perfect for clustering, tagging, or extracting necessary information.

**Small language models vs large language models**

Apart from pure size, how do large language models vs small language models differ?

Small models are great for niche, domain-specific tasks, and can provide more expert, granular information. For example, if you’re in an industry like banking, you could feed it with specialist terminology and turn it into a financial model.

LLMs, on the other hand, are like generalists; they have a wider dataset. They’ll operate on data from various disciplines. The more detailed or industry-specific your need, the harder it may be to get a precise output. Being the domain expert, an small language model would likely outperform a large language model.

Also, due to their compact nature, it’s easy and fast to set up an SLM not only on smartphones and tablets but also on edge computing devices. This can’t be said about LLMs, which require large computational resources to be deployed.

Finally, in terms of security and privacy, small language models tend to be safer, too. How so?

Since they operate locally, you don’t exchange data with external servers, reducing the risk of sensitive data breach. As I’ve covered in a post on [local language data security](https://www.netguru.com/blog/lower-data-breaches-and-security-risks-with-local-language-models), large language models are more susceptible to hacks, as they often process data on the cloud.

|  |  |  |
| --- | --- | --- |
| **Aspect** | **Small Language Models (SLMs)** | **Large Language Models (LLMs)** |
| Definition | Fewer parameters (millions to few billions) | Vast number of parameters (dozens or hundreds of billions) |
| Performance | Good for simpler tasks | Excels in complex language understanding and generation |
| Training and resources | Less computational power, memory, and storage required | Significant computational resources, memory, and storage required |
| Training time | Shorter and less costly | Time-consuming and expensive |
| Use cases | Embedded systems, mobile applications, local generation | Advanced conversational agents, content creation, complex translation systems |
| Deployment | Easier to deploy on various platforms | Often requires specialized hardware and powerful cloud servers |
| Adaptability | Quicker to be fine tuned or adjusted for specific tasks or domains | Greater flexibility and adaptability but requires significant effort to fine-tune |
| Environmental impact | Lower energy consumption and carbon footprint | Higher energy consumption and larger carbon footprint |
| Security | Lower risk of exposure due to smaller deployment footprint | Higher risk due to larger attack surface and dependency on cloud infrastructure |

**Small Language Models Examples**

**1. PHI-3 – tiny but mighty**

In April 2024, Microsoft announced they’re creating a family of [PHI-3 open models](https://news.microsoft.com/source/features/ai/the-phi-3-small-language-models-with-big-potential/), which they called “the most capable and cost-effective small language models available.”

The first to come from this Microsoft small language models’ family is Phi-3-mini, which boasts 3.8 billion parameters. What piqued my interest is that the company said it can perform better than models twice its size.

This tiny language model is said to have great logic and reasoning abilities. This means that it could work well for scenarios like:

* Summarizing long, domain specific documents like new regulations, and extracting the key points from it.
* Setting up a chatbot, which could accurately answer customer questions and dive into CRM records to suggest relevant upgrades.
* Generate marketing collateral like social media posts or product/service descriptions.

Phi-3 models are built in a safety-first approach, following [Microsoft’s Responsible AI](https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE5cmFl?culture=en-us&country=us) standards. These cover areas like privacy, security, reliability, and inclusiveness (thanks to training on high-quality, inclusive data).

If the company lives up to their promise, we can expect the phi-3 family to be among the best small language models on the market.

**2. Mixtral of experts – advanced mix of experts for better reasoning**

[Mixtral](https://mistral.ai/news/mixtral-of-experts/) is among the best small language models out there. It operates as a decoder-only model, selecting parameters from 8 different sets to process each text part or token. Designed with efficiency and capability in mind, it utilizes a specialized type of neural network, called a router, to pick the best 'experts' for processing each text segment.

In total, Mixtral has around 46.7 billion parameters but uses only 12.9 billion to analyze any given token. The beauty of it is that while it can handle complicated tasks, just like LLMs do, it’s much more efficient and cheaper. It’s trained on open web data and learns from experts and the router – all at once.

You can compare its performance to that of ChatGPT 3.5. Here is a summary:

* It’s able to leverage a wide spectrum of knowledge through a blend of various domains.
* It features unique architecture, and it can act as an LLM, but only uses a fraction of its potential.
* Mixtral creates new models capable of running on local machines while still achieving comparable power to full-scale LLMs.

**3. Llama 3 – one of the most capable small language models on your computer**

[Llama 3](https://llama.meta.com/llama3/) is an advanced language model from Meta, which is much more powerful than its predecessor. The dataset it’s been trained on is seven times as big as that of [Llama 2 and features](https://www.netguru.com/blog/comparison-llama-gpt-open-source-closed-ai-development)four times more code.

This AI model can understand text as long as 8,000 tokens, which is twice the capacity of its older brother, making it capable of comprehending and generating longer and more complex text pieces.

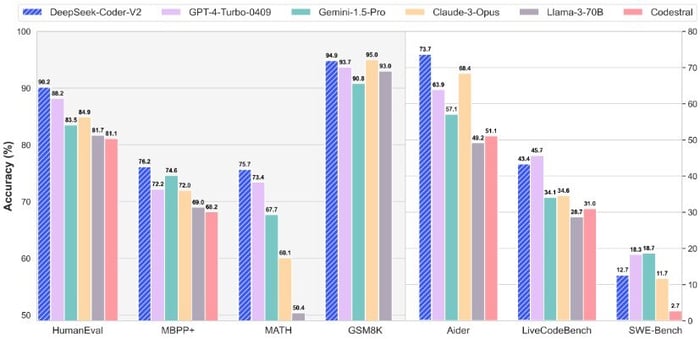
Llama 3 has enhanced reasoning capabilities and displays top-tier performance on various industry benchmarks. No wonder, they’re viewed as the best open-source models in their category. Meta made it available to all their users, [intending to promote](https://about.fb.com/news/2024/04/meta-ai-assistant-built-with-llama-3/) “the next wave of AI innovation impacting everything from applications and developer tools to evaluation methods and inference optimizations”.

Llama 3 is used in Meta products: It’s called Meta AI. Meta AI is accessible through search, feeds, and chats on Meta, Instagram, WhatsApp, and Messenger. Users can retrieve real-time information from the Internet without switching between apps.

**4. DeepSeek-Coder-V2 – additional developer on your machine**

Calling it an “additional developer” isn’t an understatement. It’s a resourceful [AI development tool](https://www.netguru.com/blog/ai-development-tools), and is among the best small language models for code generation. Tests prove that it has amazing coding and mathematical reasoning capabilities. So much so that it could replace Gemini Code or Copilot, when used on your machine.

DeepSeek-Coder-V2 is an open source model built through the Mixture-of-Experts (MoE) machine learning technique. As we can find out from its [‘Read me’ documents on GitHub](https://github.com/deepseek-ai/DeepSeek-Coder-V2?tab=readme-ov-file), it comes pre-trained with 6 trillion tokens, supports 338 languages, and has a context length of 128k tokens. Comparisons show that, when handling coding tasks, it can reach performance rates similar to GPT4-Turbo.



[Image Source](https://github.com/deepseek-ai/DeepSeek-Coder-V2?tab=readme-ov-file)

*Large vs small language models; DeepSeek-Coder-V2 as a powerful small language model example*

DeepSeek-Coder-V2 has the highest accuracy on HumanEval, with an impressive [90.2% rate](https://github.com/deepseek-ai/DeepSeek-Coder-V2?tab=readme-ov-file).

Which businesses could most benefit from this model? In my opinion, it’s best suited not only for those who need their SLM to have top-level analytical capabilities. It’s also perfect when you can’t share code through your critical systems, if those operate on the cloud.

**5. MiniCPM-Llama3-V 2.5: A GPT-4V Level Multimodal LLM on Your Phone**

MiniCPM-Llama3-V 2.5 is the newest model in the MiniCPM-V series and includes 8 billion parameters, built on SigLip-400M and Llama3-8B-Instruct. It’s a lot more capable than its predecessor, MiniCPM-V 2.0. It received 65.1 points on [OpenCompass](https://huggingface.co/openbmb/MiniCPM-Llama3-V-2_5" \t "_blank), scoring better than GPT-4V-1106, Gemini Pro, Claude 3, and Qwen-VL-Max, despite smaller number of parameters.

This small language model has a very low hallucination rate of just [10.3% on Object HalBench](https://github.com/OpenBMB/MiniCPM-V), making it more trustworthy than GPT-4V-1106 (with 13.6%).

It’s also worth mentioning that you can use it in over 30 languages, such as English, German, French, Korean, and Japanese. This relates to what I believe is the single-most powerful capability of this model, i.e., that it excels in optical character recognition (OCR).

It can process images with up to 1.8 million (!) pixels, with any aspect ratio. An OCR-specific performance test, OCRBench, gave it an impressive score of 700, outranking GPT-4o and Gemini Pro. You can ask questions about any uploaded image and receive specific, accurate answers.

This makes it perfectly suitable for:

* Digitization of images, as well as digital and handwritten text
* Data extraction for both structured and unstructured data. For example, converting tables to markdown copy.

You can use it on your mobile, making sure the images stay on your phone, which improves data security and privacy.

**General business use cases for small language models**

Small Language Models (SLMs) like PHI-3, Mixtral, Llama 3, DeepSeek-Coder-V2, and MiniCPM-Llama3-V 2.5 enhance various operations with their advanced capabilities.

**Use cases:**

1. **Retrieving Information:** Leverage searching and summarizing capabilities to find relevant information within large volumes of text.
2. **Tagging and Clustering:** Quickly convert any text into the required metadata. For example, tag comments on your webpage or product opinions and cluster them to gain insights.
3. **Creating JSON Data:** Convert any text into relevant information for API systems.

**How to use small language models in your business?**

* Convert media notes into proper API calls with tags, allowing your systems to automatically gather and analyze information. This is especially useful for staying up to date with industry news and analyzing financial information.
* Tag and cluster all opinions and comments about your products to improve based on user feedback.
* Analyze the competition by examining their pages, extracting all the features they offer, clustering this data, and using the insights to enhance your products.

**Small language models have fewer parameters but are great for domain-specific tasks**

Small language models offer significant benefits in terms of cost savings, efficiency, and versatility. They are less expensive to train and deploy than large language models, making them accessible for a wider range of applications.

Since they use computational resources efficiently, they can offer good performance and run on various devices, including smartphones and edge devices. Additionally, since you can train them on specialized data, they can be extremely helpful when handling niche tasks.

Overall, domain specific language models provide a practical, [cost-effective solution](https://www.netguru.com/blog/8-affordable-ways-to-implement-an-ai-strategy) for businesses, without sacrificing performance and output accuracy.