**Case Study Comparison between Large Language Models and Small Language Models and Applications of Small Language Models**

**Abstract**

Large Language Models (LLMs) and Small Language Models (SLMs) have recently gained attention for their capability to generate computer programs, particularly through tools such as GitHub Copilot. LLMs rely on free-text input, whereas SLMs may require structured input/output examples, but both aim for the same goal: program synthesis. In this study, we compare how well LLMs and SLMs perform on common program synthesis benchmark problems. Using well-known software metrics, we analyze the structure and diversity of the generated programs. We find that LLMs and SLMs solve a comparable number of benchmark problems (85.2% vs. 77.8%, respectively). LLMs tend to generate smaller and less complex programs than SLMs, while SLMs offer more diverse problem-solving strategies. However, this increase in diversity comes with a cost. When evaluating success rates over 100 runs per problem, LLMs outperform SLMs on over 50% of the tasks.

**Introduction**

In natural language processing (NLP), the development of language models has revolutionized how machines understand and generate human language. With the inception of models in the GPT series, such as GPT-3, the capabilities of these models have reached unprecedented levels. However, there is a growing interest in exploring the effectiveness and efficiency of smaller language models (SLMs) compared to their larger counterparts (LLMs). In this article, we will dive into the distinctions between LLMs and SLMs, their advantages, and the scenarios where each type may be advantageous.

**What are language models?**

Language models are powerful machine learning models used for natural language processing (NLP) tasks, such as text generation and sentiment analysis. These models represent natural language based on the probability of words or sequences of words occurring in a given context.

*Conventional language models* have been used in supervised settings for research purposes where the models are trained on well-labeled text datasets for specific tasks. *Pre-trained language models* offer an accessible way to get started with AI and have become more widely used in recent years. These models are trained on large-scale text corpora from the internet using deep neural networks and can be fine-tuned on smaller datasets for specific tasks.

The size of a language model is determined by its number of parameters, or *weights*, that determine how the model processes input data and generates output. Parameters are learned during the training process by adjusting the weights within layers of the model to minimize the difference between the model's predictions and the actual data. The more parameters a model has, the more complex and expressive it is, but also the more computationally expensive it is to train and use.

In general, **small language models** have *fewer than 10 billion parameters*, and **large language models** have *more than 10 billion parameters*. For example, the new Microsoft Phi-3 model family has three versions with different sizes: mini (3.8 billion parameters), small (7 billion parameters), and medium (14 billion parameters).

**Definitions**

**Large Language Models (LLMs)**

Large Language Models (LLMs) are advanced AI systems with billions of parameters, which are essentially adjustable values that shape the model’s predictions and behaviors. These parameters enable the model to learn intricate patterns from massive amounts of data, allowing it to generate coherent and contextually accurate text based on the input it receives. LLMs are highly effective at tasks like text generation, language translation, summarization, and question answering.

Some LLMs are multimodal, meaning they can process not just text but also other types of data like images and audio. This versatility makes them valuable across a wide range of applications.

The main advantage of LLMs is their ability to understand and generate human-like text, making them highly accurate and natural in their responses. However, their power also comes with challenges. Ethical concerns like the potential to spread misinformation, perpetuate bias, and violate privacy have become prominent, underscoring the importance of responsible AI usage. Researchers are actively working to address these issues while continuing to explore new applications for these powerful models.

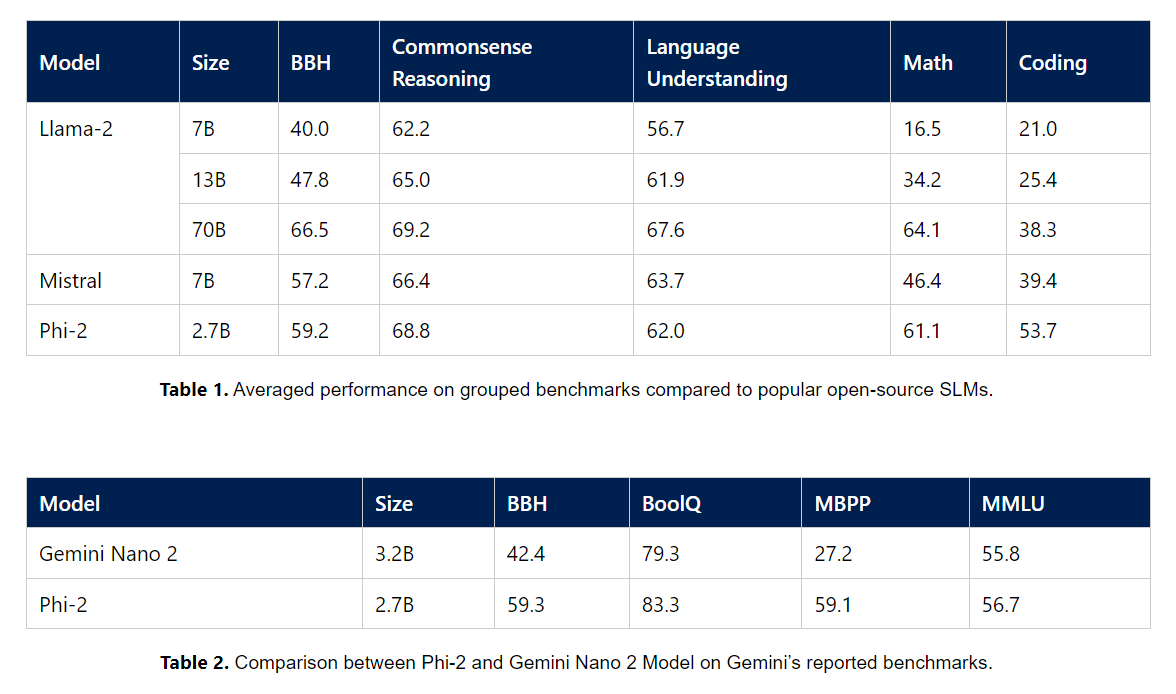
**Small Language Models (SLMs)**

Small Language Models (SLMs) are characterized by their reduced scale and simplified architecture compared to larger models. They are trained on less data with fewer parameters, so they may struggle to capture intricate language patterns effectively.

These models are crafted to be more lightweight and resource-efficient, making them suitable for deployment in environments with limited computational resources.

While SLMs may not match the scale and performance of LLMs, they offer distinct advantages in terms of speed, memory footprint and energy consumption. This makes them particularly appealing for real-time applications where low latency and efficient resource utilization are crucial factors.

As demonstrated in the table below, an example of an SLM keeping pace with LLMs is Microsoft’s Phi-2, which is an open-source model intended for research purposes. Phi-2 significantly outperforms all 4 models in mathematics (except Llama-2–70B) and coding tasks.



Source: Microsoft Phi 2: The surprising power of small language models” (12/12/23)

While these models excel in task-specific adaptability and cost-effective training, they encounter scalability issues when confronted with larger datasets or more intricate tasks. As a result, determining the suitability of small language models hinges on evaluating the specific needs of the application and striking a balance between resource efficiency and performance.

**Comparisons**

**Size:**

* LLMs, like Claude 3 and Olympus, have up to 2 trillion parameters.
* SLMs, like Phi-2, have around 2.7 billion parameters.

**Training Data:**

* LLMs require extensive, diverse datasets for broad general learning.
* SLMs use more specialized, focused, and smaller datasets.

**Training Time:**

* LLMs take months to train due to their size and complexity.
* SLMs can be trained within weeks.

**Computing Power and Resources:**

* LLMs consume a lot of computing power and resources due to large datasets and parameter sizes.
* SLMs require far less computing power, making them more sustainable.

**Proficiency:**

* LLMs excel at handling complex, sophisticated, and general tasks.
* SLMs are more suited to simpler, adequate tasks.

**Adaptation:**

* LLMs are harder to fine-tune and adapt to customized tasks.
* SLMs are easier to fine-tune and customize for specific needs.

**Inference:**

* LLMs require specialized hardware like GPUs and cloud services, needing an internet connection.
* SLMs are small enough to run locally on devices like a Raspberry Pi or a phone, without requiring an internet connection.

**Latency:**

* LLMs generally have higher latency, taking seconds to respond.
* SLMs, due to their smaller size, are much quicker.

**Cost:**

* LLMs have higher token costs because they consume more resources for inference.
* SLMs are cheaper to run due to lower computing and resource needs.

**Control:**

* LLMs are dependent on model builders, and changes in the model may cause drift or forgetting.
* SLMs can be run on private servers, allowing full control, tuning, and freezing to prevent changes.

**Use Cases and Applications**

**Small Language Models (SLMs)**

1. **Specialized Knowledge Tasks:**
   * SLMs are ideal for applications requiring domain-specific knowledge, such as medical terminology processing, legal document understanding, or scientific data interpretation.
2. **Real-Time Language Processing:**
   * Due to their smaller size and faster processing times, SLMs are well-suited for real-time applications such as voice assistants, which require quick and efficient responses.
3. **Mobile Applications and IoT Devices:**
   * SLMs can be deployed in resource-constrained environments, such as mobile devices and IoT platforms, offering low-latency language processing without relying on cloud computing or internet access.
4. **Voice Assistants:**
   * SLMs are efficient in processing voice commands with low latency, making them suitable for embedded systems like smart home devices or in-car assistants.
5. **Educational Purposes:**
   * SLMs are widely used in teaching NLP concepts and experimenting with machine learning models due to their simplicity, smaller datasets, and lower computing power requirements**.**
6. **Small-Scale NLP Projects:**
   * SLMs are beneficial for smaller projects or startups needing natural language understanding without the overhead of massive computing resources, enabling quick deployment and cost savings.

**Contrasting characteristics**

• **Model size:** LLMs are characterized by their large size, typically containing hundreds of billions of parameters, whereas SLMs have a smaller parameter count to optimize resource consumption. LLMs such as ChatGPT can contain up to 1.7 trillion parameters, while an open SLM such as Phi-2 contains 2.7 billion parameters.

• **Performance:** LLMs tend to outperform SLMs in terms of accuracy and fluency, thanks to their extensive training on vast datasets. However, SLMs can still achieve satisfactory performance for many NLP tasks while offering faster inference times.

• **Resource requirements:** LLMs demand substantial computational resources, including high-end GPUs or TPUs for training and inference. In contrast, SLMs can run efficiently on devices with limited resources, making them more accessible for deployment in various settings.

• **Deployment flexibility:** SLMs are more flexible for deployment in resource-constrained environments such as mobile devices, IoT devices, or edge computing platforms, whereas LLMs are better suited for cloud-based applications where ample computational resources are available.

• **Understanding and field specialization:** SLMs are trained on data from specialized domains, potentially limiting their grasp of broader information across multiple fields. However, they often demonstrate exceptional proficiency within their designated domain. In contrast, LLMs aim to replicate human intelligence on a much larger scale. Leveraging large datasets of training data, they are expected to demonstrate capable performance across various domains compared to domain specific SLMs. Additionally, LLMs possess greater adaptability and can be refined for enhanced performance in downstream tasks like programming.

• **Bias:** LLMs can often exhibit biases due to insufficient fine-tuning and training on raw data from the internet. This source of training data may lead to a misrepresentation of certain groups or ideas, as well as erroneous labeling. Inherent biases in language, influenced by factors like dialect, geographic location and grammar rules contribute to further complexity. On the other hand, small language models (SLMs), which train on smaller, domain-specific datasets, inherently carry a lower risk of bias compared to LLMs.

**Limitations of Small Language Models**

**Niche Focus and Limited Generalization**

While the specialized focus of SLMs is a significant advantage, it also poses limitations. These models may not perform well outside their specific domain of training, lacking the broad knowledge base that allows LLMs to generate relevant content across a wide range of topics. This limitation requires organizations to potentially deploy multiple SLMs to cover different areas of need, which could complicate the AI infrastructure.

**Rapid Evolution and Technical Challenges**

The field of Language Models is rapidly evolving, with new models and approaches being developed at a fast pace. This constant innovation, while exciting, presents challenges in keeping up with the latest advancements and ensuring that deployed models remain state-of-the-art. Additionally, customizing and fine-tuning SLMs to specific enterprise needs can require specialized knowledge and expertise in data science and machine learning, resources that not all organizations may have readily available.

**Evaluation and Selection Difficulties**

With the burgeoning interest in SLMs, the market has seen an influx of various models, each claiming superiority in certain aspects. However, LLM evaluation and selecting the appropriate Small Language Model for a specific application can be daunting. Performance metrics can be misleading, and without a deep understanding of the model size underlying technology, businesses may struggle to choose the most effective model for their needs.

**Conclusion**

In conclusion, both LLMs and SLMs have their unique strengths and applications in natural language processing. While LLMs offer unparalleled performance and accuracy, SLMs provide efficiency and flexibility, particularly in resource-constrained environments. Your choice between LLMs and SLMs depends on the specific requirements of the application, balancing performance with resource constraints to achieve optimal results in various NLP tasks. The suitability of language models depends entirely on the specific use case and the resources at one's disposal. For some businesses, using an LLM as a chat agent for support teams may be to their advantage as it can handle large volumes of inquiries.

For function-specific tasks, SLMs typically excel where specialized and proprietary knowledge is of the utmost importance. In such cases, training an SLM in-house, and leveraging domain-specific expertise can yield sophistication in these specialized sectors.

As the field continues to evolve, advancements in both types of models will further broaden the landscape of possibilities in natural language processing and AI-driven applications.

**References**

**IEEE Transactions and Evolutionary Computation (Abstract)**

**Rackspace. (2024, September 13). *Large language models (LLMs) vs. small language models (SLMs)*. Rackspace. Retrieved September 17, 2024**

**Simms, K. (n.d.). 10 differences between small language models (SLM) and large language models (LLM). LinkedIn. Retrieved September 17, 2024**

**Aisera. (n.d.). *Small language models: Limitations and use cases***

**Microsoft Phi 2: The surprising power of small language models” (12/12/23)**

**Microsoft. (n.d.). *Concepts of AI and machine learning: Language models*. Microsoft Learn. Retrieved September 17, 2024,**