SMALL LANGUAGE MODELS

Introduction:

In today's fast-growing world, we face many challenges and problems that often seem impossible to solve. At the same time, the technology world is rapidly evolving. One of the most powerful and fast-growing technologies today is **Artificial Intelligence (AI)**.

Unlike many technologies that fade away over time, AI has grown rapidly and gained global attention. Its progress has been as steep as a mountain peak. As a result, **Large Language Models (LLMs)** came into the picture. The foundation of LLMs was introduced in 2017 through the research paper **"Attention Is All You Need"**, published by Google.

As the use of LLMs increased, new challenges such as high cost, large memory usage, and heavy computation also emerged. To address these issues, **Small Language Models (SLMs)** were introduced. This blog explores what SLMs are, why they matter, and how they differ from LLMs.

Small Language Models (SLMs)

A **Small Language Model (SLM)** is an artificial intelligence model designed to understand text, recognize patterns, and generate responses based on user queries. As the name suggests, SLMs are smaller in scope and scale when compared to Large Language Models (LLMs).

In terms of size, SLMs are typically trained on **a few million to a few billion parameters**. Parameters are internal variables of the model, such as **weights and biases**, which help the model learn patterns from data.

Due to their compact size, Small Language Models are more efficient than large models. They require **less memory and computational power**, which makes them suitable for **edge devices such as mobile phones and small servers**. In some cases, SLMs can even work **offline**, enabling interaction without a constant internet connection.

How Small Language Models Work

Similar to Large Language Models (LLMs), **Small Language Models (SLMs)** are built using a neural network architecture called the **Transformer**. Transformers have become the backbone of modern **Natural Language Processing (NLP)** because they can efficiently understand relationships between words in a sentence.

The transformer architecture mainly consists of two parts: the **Encoder** and the **Decoder**. The encoder converts input text into a machine-readable format called **embeddings**. These embeddings capture both the meaning of words and their **positions** in a sentence, which helps the model understand word order and avoid confusion.

A key component of transformers is the **self-attention mechanism**. Self-attention allows the model to focus on the most relevant words when processing each word in a sentence. It calculates attention scores using mathematical and probabilistic methods to determine how strongly words are related to each other.

The decoder then uses the encoder's embeddings and the self-attention mechanism to generate the output text. It predicts the **most statistically probable next word** at each step, eventually producing a complete and meaningful response.

Example Of SLM Models:

- DistilBERT
- Gemma
- GPT-4o mini
- Granite
- Llama
- Ministral
- Phi

Benefits of Small Language Models

Accessibility: Small Language Models make AI more accessible to researchers and enterprises because they do not require complex infrastructure such as multiple GPU configurations. This lowers the barrier to entry and allows more teams to experiment with AI.

Lower Latency: Since SLMs have fewer parameters, they process data faster. This results in quicker response times, making them ideal for real-time applications such as chatbots and on-device assistants.

Reduced Cost: Organizations using SLMs do not need to invest heavily in expensive hardware or deployment infrastructure. As a result, both training and inference costs are significantly reduced.

Reference:IBM