# The AI Era, Part 3: LLMs, SLMs, and Foundation Models

## SMT Perspectives and Prospects

by Dr. Jennie S. Hwang, CEO, H-TECHNOLOGIES GROUP

Since the introduction of ChatGPT on Nov. 30, 2022, and ChatGPT4 on March 14, 2023, large language models (LLMs) have been in everyday news and conversations. LLMs represent a significant advancement in AI, which has the potential to revolutionize multiple fields. This column offers a snapshot of LLMs from the user's perspective.

As a subset of AI models, LLMs are designed to understand, process, and manipulate human language and generate human-like text through learning patterns and relationships. A model is trained on vast datasets, which allow it to recognize, translate, predict, and generate text or other content and perform a wide range of tasks related to natural language processing (NLP).

The recent success of LLMs stems from the following:

- The introduction of transformer architectures
- The capability of increased computational power
- The availability and use of vast training data

LLMs' underlying technology is based on deep learning, particularly neural networks. Deep learning algorithms are capable of a wide range of natural language tasks. The most common architecture for LLMs is the transformer model, introduced in the groundbreaking paper, "Attention Is All You Need" by Vaswani in 2017[1].

## Transformer Architectures

Transformers can derive meanings from long text sequences to understand how different words or semantic components might be related. They can then determine how likely they are to occur in proximity to each other.

The key components include attention mechanisms that focus on different parts of the input sequence when generating output, and self-attention mechanisms to process input data—allowing the model to weigh the importance of different words in a sentence sequence and understand context when making predictions. Its feed-forward neural networks process the attention outputs to produce the final predictions.

**"The architecture comprises an encoder-decoder structure."**

The architecture comprises an encoder-decoder structure. The encoder processes the input sequence and produces a set of continuous representations (embeddings), while the decoder takes the encoder's output and generates the final prediction, e.g., a translated sentence or a continuation of text. Additionally, a multi-head attention mechanism can improve the model's ability to focus simultaneously on different parts of the input sequence. Multiple attention heads enhance the model's capacity to capture diverse linguistic patterns and relationships within the data. Transformer architecture also uses positional encoding to compensate for the lack of sequential processing and maintains information about word order.

Transformer architecture facilitates effective pre-training on large datasets and subsequent fine-tuning for specific tasks. It is a key aspect of LLM development. This pre-training allows the transformer architecture to learn general language patterns while fine-tuning works on specific datasets to improve performance tasks. Many iterations are required for a model to reach the point where it can produce plausible results. The mathematics and coding that go into creating and training generative AI models, particularly LLMs, can be incredibly time-intensive, costly, and complex.

One of the unique advantages of transformer architecture is that it can handle input data in parallel. Parallel processing offers greater efficiency and scalability compared to other architectures, such as a recurrent neural network (RNN) or long short-term memory (LSTM), which process data sequentially.

## LLMs

Based on the concept of transformer architecture, LLMs consist of intricate neural networks trained on large quantities of unlabeled text. An LLM breaks the text into words or phrases and assigns a number to each, using sophisticated computer chips and neural networks to find patterns in the pieces of text through mathematical formulas, and learns to "guess" the next word in a sequence. Then, using NLP, the model can understand what's being asked and reply. Because it uses mathematical formulas rather than text searching to generate responses, it is not ready-made information waiting to be retrieved. Rather, it uses billions or even trillions of numbers to calculate responses from scratch; producing new sequences of words on the fly. However, LLMs are computationally intensive, requiring high computing power and parallel computing, such as graphic processing units (GPUs).

LLMs are characterized by their large param-

Table 1: Key characteristics of ChatGPT models

GPT-1 to GPT –4 (4+)

| Model | Launch Date | Training Data | No. of Parameters | Max. Sequence Length |
|---|---|---|---|---|
| GPT-1 | June 2018 | Common Crawl, BookCorpus | 117 million | 1024 |
| GPT-2 | February 2019 | Common Crawl, BookCorpus, WebText | 1.5 billion | 2048 |
| GPT-3 | June 2020 | Common Crawl, BookCorpus, Wikipedia, Books, Articles, and more | 175 billion | 4096 |
| GPT-4 | March 2023 | (Unknown) | (Estimated to be in the trillions) | (8,000–32,000+) |

GPT 4o
GPT-5> 60 trillion?

eters, which act as the model's knowledge bank. Table 1[2] shows the relative number of parameters and the maximum sequence length of the progressive ChatGPT models: GPT-1, GPT-2, GPT-3, and GPT-4. Models can handle tasks such as generating text, translating, making summaries, answering questions, and analyzing sentiments. They can also be fine-tuned to undertake specific tasks.

How large are LLMs? There is no universally agreed figure. However, they are generally characterized by the number of parameters (billions or even trillions) and the size of the training data they are exposed to. Usually, LLMs have at least 1 petabyte of storage (the human brain stores about 2.5 petabytes of memory data.)

This leads us to another related terminology: foundation models.

## LLMs vs. Foundation Models

Foundation models are base models that provide a versatile "foundation" that can be fine-tuned and adapted for a wide range of applications, from language processing to image recognition. Foundation models are multimodal and can be trained on different data or modalities. In essence, LLMs are foundational models, but not all foundational models are LLMs.

## LLMs vs. SLMs

Recently, "smaller" language models have come into vogue due to practical factors such as cost and readiness. So, what is considered a small language model (SLM)? In terms of size, there are no hard and fast rules. In general, LLMs typically have over 20 billion parameters. For example, GPT-3 has 175 billion as shown in Table 1, while SLMs range from 500 million to 20 billion parameters.

LLMs are broad-spectrum models trained on massive datasets, excelling at deep reasoning, complex context handling, and extensive content generation. SLMs are more specialized, focusing on specific domains or tasks. They may exhibit less bias and are less costly. They are also faster and potentially more accurate (less hallucination) and, accordingly, are more readily able to be put to work.

Artificial intelligence is still nascent but continues to advance. It would not be surprising to see the new frontier offering another level of capabilities and accuracy with a new architecture. **SMT007**

**References**

1. "Attention Is All You Need," by Ashish Vaswani, et al., Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS).

2. Professional Development Course: "Artificial Intelligence: Opportunities, Challenges, and Possibilities," by Jennie S. Hwang.

**Appearances**

Dr. Jennie Hwang will instruct a professional development course on "Artificial Intelligence—Opportunities, Challenges and Possibilities" at the 2024 SMTA International, Oct. 21 in Chicago. She will also deliver the keynote speech titled "Artificial Intelligence Era: Work, Life, Technology, Leadership, and Women!" at the Women's Leadership Program on Oct. 21.

**Dr. Jennie S. Hwang,** an international businesswoman, speaker, and business and technology advisor, is a pioneer and long-standing leader in SMT manufacturing since its inception, and in developing and implementing lead-free electronics technology and manufacturing.

She has served as chair of Artificial Intelligence-Justified Confidence for DoD Command and Control study, chair of AI Committee of the National Academies, and Review Panels of NSF National AI Institutes and Committee of Strategic Thinking for Engineering Research. An International Hall of Famer (Women in Technology), she has been inducted into the National Academy of Engineering, named an R&D-Stars-to-Watch, and received the YWCA Achievement Award. She has held senior executive positions with Lockheed Martin Corp., and was CEO of International Electronic Materials Corp. She is currently CEO of H-Technologies Group, providing business, technology, and manufacturing solutions.

She has served as chair of the Laboratory Assessment Board, the DoD Army Research Laboratory Assessment Board, and the Assessment Board of Army Engineering Centers. She is on the board of Fortune-500 NYSE companies and civic and university boards, Commerce Department's Export Council, National Materials and Manufacturing Board, NIST Assessment Board, various national panels/committees, and international leadership positions.

She is the author of 10 books (four as co-author) and 750+ technical/editorial publications. She is a speaker and author on trade, business, and education issues. Her formal education includes four academic degrees (Ph.D., M.S., M.A., B.S.), as well as Harvard Business School Executive Program and Columbia University Corporate Governance Program. To read previous columns, click here.