Large language models (LLMs) are a category of deep learning models trained on immense amounts of data, making them capable of understanding and generating natural language and other types of content to perform a wide range of tasks. LLMs are built on a type of neural network architecture called a transformer which excels at handling sequences of words and capturing patterns in text.

LLMs work as giant statistical prediction machines that repeatedly predict the next word in a sequence. They learn patterns in their text and generate language that follows those patterns.

LLMs represent a major leap in how humans interact with technology because they are the first AI system that can handle unstructured human language at scale, allowing for natural communication with machines. Where traditional search engines and and other programmed systems used algorithms to match keywords, LLMs capture deeper context, nuance and reasoning. LLMs, once trained, can adapt to many applications that involve interpreting text, like summarizing an article, debugging code or drafting a legal clause. When given agentic capabilities, LLMs can perform, with varying degrees of autonomy, various tasks that would otherwise be performed by humans.

LLMs are the culmination of decades of progress in natural language processing (NLP) and machine learning research, and their development is largely responsible for the explosion of artificial intelligence advancements across the late 2010s and 2020s. Popular LLMs have become household names, bringing generative AI to the forefront of the public interest. LLMs are also used widely in enterprises, with organizations investing heavily across numerous business functions and use cases.

LLMs are easily accessible to the public through interfaces like Anthropic's Claude, Open AI's ChatGPT, Microsoft's Copilot, Meta's Llama models, and Google's Gemini assistant, along with its BERT and PaLM models. IBM maintains a Granite model series on watsonx.ai, which has become the generative AI backbone for other IBM products like watsonx Assistant and watsonx Orchestrate.