# Large language models.pdf

The matter of LLM's exhibiting intelligence or understanding has two main aspects - the first is how to model thought and language in a computer system, and the second is how to enable the computer system to generate human like language. These aspects of language as a model of cognition have been developed in the field of cognitive linguistics . American linguist George Lakoff presented Neural Theory of Language (NTL) as a computational basis for using language as a model of learning tasks and understanding. The NTL Model outlines how specific neural structures of the human brain shape the nature of thought and language and in turn what are the computational properties of such neural systems that can be applied to model thought and language in a computer system. After a framework for modeling language in a computer systems was established, the focus shifted to establishing frameworks for computer systems to generate language with acceptable grammar. In his 2014 book titled The Language Myth: Why Language Is Not An Instinct , British cognitive linguist and digital communication technologist Vyvyan Evans mapped out the role of probabilistic context-free grammar (PCFG) in enabling NLP to model cognitive patterns and generate human like language.

The first AI language models trace their roots to the earliest days of AI. The Eliza language model debuted in 1966 at MIT and is one of the earliest examples of an AI language model. All language models are first trained on a set of data, and then they make use of various techniques to infer relationships and then generate new content based on the trained data. Language models are commonly used in natural language processing ( NLP ) applications where a user inputs a query in natural language to generate a result.

In addition to teaching human languages to artificial intelligence (AI) applications, large language models can also be trained to perform a variety of tasks like understanding protein structures, writing software code, and more. Like the human brain, large language models must be pre-trained and then fine-tuned so that they can solve text classification, question answering, document summarization, and text generation problems. Their problem-solving capabilities can be applied to fields like healthcare, finance, and entertainment where large language models serve a variety of NLP applications , such as translation, chatbots, AI assistants, and so on.

A large language model (LLM) is a deep learning algorithm that can perform a variety of natural language processing (NLP) tasks. Large language models use transformer models and are trained using massive datasets — hence, large. This enables them to recognize, translate, predict, or generate text or other content.

Because language models may overfit to their training data, models are usually evaluated by their perplexity on a test set of unseen data. This presents particular challenges for the evaluation of large language models. As they are trained on increasingly large corpora of text largely scraped from the web, it becomes increasingly likely that models' training data inadvertently includes portions of any given test set.

Despite the tremendous capabilities of zero-shot learning with large language models, developers and enterprises have an innate desire to tame these systems to behave in their desired manner. To deploy these large language models for specific use cases, the models can be customized using several techniques to achieve higher accuracy. Some techniques include prompt tuning , fine-tuning, and adapters .

In 2023, Nature Biomedical Engineering wrote that "it is no longer possible to accurately distinguish" human-written text from text created by large language models, and that "It is all but certain that general-purpose large language models will rapidly proliferate... It is a rather safe bet that they will change many industries over time." Goldman Sachs suggested in 2023 that generative language AI could increase global GDP by 7% in the next ten years, and could expose to automation 300 million jobs globally.

Large language models are trained using unsupervised learning . With unsupervised learning, models can find previously unknown patterns in data using unlabelled datasets. This also eliminates the need for extensive data labeling, which is one of the biggest challenges in building AI models.

Large language models (LLM) are very large deep learning models that are pre-trained on vast amounts of data. The underlying transformer is a set of neural networks that consist of an encoder and a decoder with self-attention capabilities. The encoder and decoder extract meanings from a sequence of text and understand the relationships between words and phrases in it.

Notably, gender bias refers to the tendency of these models to produce outputs that are unfairly prejudiced towards one gender over another. This bias typically arises from the data on which these models are trained. Large language models often assign roles and characteristics based on traditional gender norms. For example, it might associate nurses or secretaries predominantly with women and engineers or CEOs with men.

In addition to accelerating natural language processing applications — like translation, chatbots and AI assistants — large language models are used in healthcare , software development and use cases in many other fields .

Large language models are among the most successful applications of transformer models . They aren’t just for teaching AIs human languages, but for understanding proteins, writing software code, and much, much more.

A large language model ( LLM ) is a language model notable for its ability to achieve general-purpose language generation and understanding. LLMs acquire these abilities by learning statistical relationships from text documents during a computationally intensive self-supervised and semi-supervised training process. LLMs are artificial neural networks , the largest and most capable of which are built with a decoder-only transformer -based architecture. Some recent implementations are based on other architectures, such as recurrent neural network variants and Mamba (a state space model).

Enabling more accurate information through domain-specific LLMs developed for individual industries or functions is another possible direction for the future of large language models. Expanded use of techniques such as reinforcement learning from human feedback , which OpenAI uses to train ChatGPT, could help improve the accuracy of LLMs, too. There's also a class of LLMs based on the concept known as retrieval-augmented generation -- including Google's Realm (short for Retrieval-Augmented Language Model) -- that will enable training and inference on a very specific corpus of data, much like how a user today can specifically search content on a single site.

Some datasets have been constructed adversarially, focusing on particular problems on which extant language models seem to have unusually poor performance compared to humans. One example is the TruthfulQA dataset, a question answering dataset consisting of 817 questions which language models are susceptible to answering incorrectly by mimicking falsehoods to which they were repeatedly exposed during training. For example, an LLM may answer "No" to the question "Can you teach an old dog new tricks?" because of its exposure to the English idiom you can't teach an old dog new tricks , even though this is not literally true.

While LLMs have shown remarkable capabilities in generating human-like text, they are susceptible to inheriting and amplifying biases present in their training data. This can manifest in skewed representations or unfair treatment of different demographics, such as those based on race, gender, language, and cultural groups. Since English data is overrepresented in current large language models' training data, it may also downplay non-English views.

Transformer neural network architecture allows the use of very large models, often with hundreds of billions of parameters. Such large-scale models can ingest massive amounts of data, often from the internet, but also from sources such as the Common Crawl , which comprises more than 50 billion web pages, and Wikipedia, which has approximately 57 million pages.

When one subtracts out from the y-axis the best performance that can be achieved even with infinite scaling of the x-axis quantity, large models' performance, measured on various tasks, seems to be a linear extrapolation of other (smaller-sized and medium-sized) models' performance on a log-log plot. However, sometimes the line's slope transitions from one slope to another at point(s) referred to as break(s) in downstream scaling laws, appearing as a series of linear segments connected by arcs; it seems that larger models acquire "emergent abilities" at this point(s). These abilities are discovered rather than programmed-in or designed, in some cases only after the LLM has been publicly deployed.

Amazon SageMaker JumpStart is a machine learning hub with foundation models, built-in algorithms, and prebuilt ML solutions that you can deploy with just a few clicks With SageMaker JumpStart, you can access pretrained models, including foundation models, to perform tasks like article summarization and image generation. Pretrained models are fully customizable for your use case with your data, and you can easily deploy them into production with the user interface or SDK.

Because of the rapid pace of improvement of large language models, evaluation benchmarks have suffered from short lifespans, with state of the art models quickly "saturating" existing benchmarks, exceeding the performance of human annotators, leading to efforts to replace or augment the benchmark with more challenging tasks. In addition, there are cases of "shortcut learning" wherein AIs sometimes "cheat" on multiple-choice tests by using statistical correlations in superficial test question wording in order to guess the correct responses, without necessarily understanding the actual question being asked.

Learn more about large language models .

Large language models by themselves are " black boxes ", and it is not clear how they can perform linguistic tasks. There are several methods for understanding how LLM work.

All large language models are generative AI 1 .

Large language models largely represent a class of deep learning architectures called transformer networks . A transformer model is a neural network that learns context and meaning by tracking relationships in sequential data, like the words in this sentence.

A language model is a machine learning model that aims to predict and generate plausible language. Autocomplete is a language model, for example.

A large language model (LLM) is a type of artificial intelligence ( AI ) algorithm that uses deep learning techniques and massively large data sets to understand, summarize, generate and predict new content. The term generative AI also is closely connected with LLMs, which are, in fact, a type of generative AI that has been specifically architected to help generate text-based content.

Large language models are also referred to as neural networks (NNs) , which are computing systems inspired by the human brain. These neural networks work using a network of nodes that are layered, much like neurons.

Popular large language models have taken the world by storm. Many have been adopted by people across industries. You've no doubt heard of ChatGPT, a form of generative AI chatbot.

In 2021, NVIDIA and Microsoft developed Megatron-Turing Natural Language Generation 530B , one of the world’s largest models for reading comprehension and natural language inference, which eases tasks like summarization and content generation.

Advancements across the entire compute stack have allowed for the development of increasingly sophisticated LLMs. In June 2020, OpenAI released GPT-3 , a 175 billion-parameter model that generated text and code with short written prompts. In 2021, NVIDIA and Microsoft developed Megatron-Turing Natural Language Generation 530B , one of the world’s largest models for reading comprehension and natural language inference, with 530 billion parameters.

The popular ChatGPT AI chatbot is one application of a large language model. It can be used for a myriad of natural language processing tasks.

The largest models, such as Google's Gemini 1.5 , presented in February 2024, can have a context window sized up to 1 million (context window of 10 million was also "successfully tested"). Other models with large context windows includes Anthropic's Claude 2.1, with a context window of up to 200k tokens. Note that this maximum refers to the number of input tokens and that the maximum number of output tokens differs from the input and is often smaller. For example, the GPT-4 Turbo model has a maximum output of 4096 tokens.

On the other hand, the use of large language models could drive new instances of shadow IT in organizations. CIOs will need to implement usage guardrails and provide training to avoid data privacy problems and other issues. LLMs could also create new cybersecurity challenges by enabling attackers to write more persuasive and realistic phishing emails or other malicious communications.

Flamingo demonstrated the effectiveness of the tokenization method, finetuning a pair of pretrained language model and image encoder to perform better on visual question answering than models

trained from scratch. Google PaLM model was finetuned into a multimodal model PaLM-E using the tokenization method, and applied to robotic control. LLaMA models have also been turned multimodal using the tokenization method, to allow image inputs, and video inputs.

Advances in software and hardware have reduced the cost substantially since 2020, such that in 2023 training of a 12-billion-parameter LLM computational cost is 72,300 A100-GPU -hours, while in 2020 the cost of training a 1.5-billion-parameter LLM (which was two orders of magnitude smaller than the state of the art in 2020) was between $80 thousand and $1.6 million. Since 2020, large sums were invested in increasingly large models. For example, training of the GPT-2 (i.e. a 1.5-billion-parameters model) in 2019 cost $50,000, while training of the PaLM (i.e. a 540-billion-parameters model) in 2022 cost $8 million.

While quantized models are typically frozen, and only pre-quantized models are finetuned, quantized models can still be finetuned.

Political bias refers to the tendency of algorithms to systematically favor certain political viewpoints, ideologies, or outcomes over others. Language models may also exhibit political biases. Since the training data includes a wide range of political opinions and coverage, the models might generate responses that lean towards particular political ideologies or viewpoints, depending on the prevalence of those views in the data.

Thanks to the extensive training process that LLMs undergo, the models don’t need to be trained for any specific task and can instead serve multiple use cases. These types of models are known as foundation models.

Some notable LLMs are OpenAI 's GPT series of models (e.g., GPT-3.5 and GPT-4 , used in ChatGPT and Microsoft Copilot ), Google 's PaLM and Gemini (the latter of which is currently used in the chatbot of the same name ), Meta 's LLaMA family of open-source models, and Anthropic 's Claude models.

Some commenters expressed concern over accidental or deliberate creation of misinformation, or other forms of misuse. For example, the availability of large language models could reduce the skill-level required to commit bioterrorism; biosecurity researcher Kevin Esvelt has suggested that LLM creators should exclude from their training data papers on creating or enhancing pathogens.