# From Google

**What is a language model?**

A language model is a machine learning [model](https://developers.google.com/machine-learning/glossary#model) that aims to predict and generate plausible language. Autocomplete is a language model, for example.

These models work by estimating the probability of a [token](https://developers.google.com/machine-learning/glossary#token) or sequence of tokens occurring within a longer sequence of tokens. Consider the following sentence:

When I hear rain on my roof, I \_\_\_\_\_\_\_ in my kitchen.

If you assume that a token is a word, then a language model determines the probabilities of different words or sequences of words to replace that underscore. For example, a language model might determine the following probabilities:

cook soup 9.4%

warm up a kettle 5.2%

cower 3.6%

nap 2.5%

relax 2.2%

...

A "sequence of tokens" could be an entire sentence or a series of sentences. That is, a language model could calculate the likelihood of different entire sentences or blocks of text.

Estimating the probability of what comes next in a sequence is useful for all kinds of things: generating text, translating languages, and answering questions, to name a few.

**What is a large language model?**

Modeling human language at scale is a highly complex and resource-intensive endeavor. The path to reaching the current capabilities of language models and large language models has spanned several decades.

As models are built bigger and bigger, their complexity and efficacy increases. Early language models could predict the probability of a single word; modern large language models can predict the probability of sentences, paragraphs, or even entire documents.

The size and capability of language models has exploded over the last few years as computer memory, dataset size, and processing power increases, and more effective techniques for modeling longer text sequences are developed.

**How large is large?**

The definition is fuzzy, but "large" has been used to describe BERT (110M parameters) as well as PaLM 2 (up to 340B parameters).

[Parameters](https://developers.google.com/machine-learning/glossary#parameter) are the [weights](https://developers.google.com/machine-learning/glossary#weight) the model learned during training, used to predict the next token in the sequence. "Large" can refer either to the number of parameters in the model, or sometimes the number of words in the dataset.

**Transformers**

A key development in language modeling was the introduction in 2017 of Transformers, an architecture designed around the idea of [attention](https://developers.google.com/machine-learning/glossary#attention). This made it possible to process longer sequences by focusing on the most important part of the input, solving memory issues encountered in earlier models.

Transformers are the state-of-the-art architecture for a wide variety of language model applications, such as translators.

If the input is **"I am a good dog."**, a Transformer-based translator transforms that input into the output **"Je suis un bon chien."**, which is the same sentence translated into French.

Full Transformers consist of an [encoder](https://developers.google.com/machine-learning/glossary#encoder) and a [decoder](https://developers.google.com/machine-learning/glossary#decoder). An encoder converts input text into an intermediate representation, and a decoder converts that intermediate representation into useful text.

**Self-attention**

Transformers rely heavily on a concept called self-attention. The self part of self-attention refers to the "egocentric" focus of each token in a corpus. Effectively, on behalf of each token of input, self-attention asks, "How much does every other token of input matter to **me**?" To simplify matters, let's assume that each token is a word and the complete context is a single sentence. Consider the following sentence:

The animal didn't cross the street because it was too tired.

There are 11 words in the preceding sentence, so each of the 11 words is paying attention to the other ten, wondering how much each of those ten words matters to them. For example, notice that the sentence contains the pronoun **it**. Pronouns are often ambiguous. The pronoun **it** always refers to a recent noun, but in the example sentence, which recent noun does **it** refer to: the animal or the street?

The self-attention mechanism determines the relevance of each nearby word to the pronoun **it**.

**What are some use cases for LLMs?**

LLMs are highly effective at the task they were built for, which is generating the most plausible text in response to an input. They are even beginning to show strong performance on other tasks; for example, summarization, question answering, and text classification. These are called [emergent abilities](https://research.google/pubs/pub52065/). LLMs can even solve some math problems and write code (though it's advisable to check their work).

LLMs are excellent at mimicking human speech patterns. Among other things, they're great at combining information with different styles and tones.

However, LLMs can be components of models that do more than just generate text. Recent LLMs have been used to build sentiment detectors, toxicity classifiers, and generate image captions.

**LLM Considerations**

Models this large are not without their drawbacks.

The largest LLMs are expensive. They can take months to train, and as a result consume lots of resources.

They can also usually be repurposed for other tasks, a valuable silver lining.

Training models with upwards of [a trillion parameters](https://cloud.google.com/blog/products/ai-machine-learning/training-a-recommender-model-of-100-trillions-parameters-on-google-cloud) creates engineering challenges. Special infrastructure and programming techniques are required to coordinate the flow to the chips and back again.

There are ways to mitigate the costs of these large models. Two approaches are [offline inference](https://developers.google.com/machine-learning/glossary#offline-inference) and [distillation](https://ai.googleblog.com/2021/12/training-machine-learning-models-more.html).

Bias can be a problem in very large models and should be considered in training and deployment.

As these models are trained on human language, this can introduce numerous potential ethical issues, including the misuse of language, and bias in race, gender, religion, and more.

It should be clear that as these models continue to get bigger and perform better, there is continuing need to be diligent about understanding and mitigating their drawbacks. Learn more about Google's approach to [responsible AI](https://ai.google/responsibility/responsible-ai-practices/).

------ <https://developers.google.com/machine-learning/resources/intro-llms>