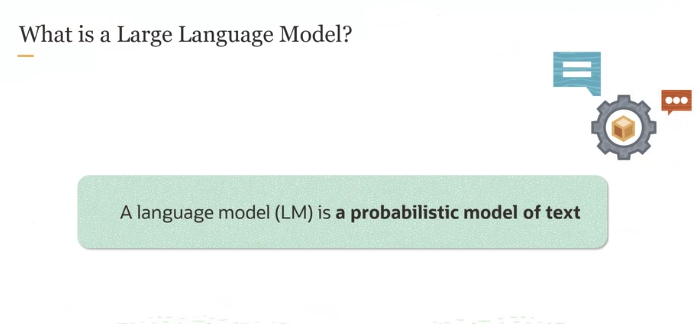
Source for GenAI: <https://www.ibm.com/topics/generative-ai>



Source: <https://www.ibm.com/topics/large-language-models>

**What is a Language Model?**

A **language model** is like a smart guesser for sentences. It helps a computer predict what word will come next in a sentence based on the words that came before.

**How Does It Work?**

* Imagine you’re writing a sentence, and you’ve already typed "I love to eat...".
* A language model will **look at the previous words** ("I love to eat") and try to **predict what word** should come next.
* It might guess something like "pizza", "ice cream", or "pasta", because based on the words already written, those are **likely** to be the next word.

**Why is it Probabilistic?**

The model doesn’t know for sure what word is coming next, so it assigns **probabilities** (likelihoods) to different options.

* For example, the model might think the next word is "pizza" with a **60% chance**, "ice cream" with **30%**, and "pasta" with **10%**.
* Based on this, the model would probably suggest "pizza" because it thinks that’s the **most likely** choice.

**Where is it Used?**

This kind of language model is used in things like:

* **Autocorrect** or **predictive text** on your phone.
* **Search engines** to suggest what you might be looking for.
* **AI chatbots** to help them make sense of what you’re typing and respond appropriately.

A screenshot of a computer

Description automatically generated

**What is Happening in the Example?**

You have a sentence:  
**"I wrote to the zoo to send me a pet. They sent me a \_\_\_\_\_\_."**

The blank at the end needs to be filled with a word that makes sense. **What word would you expect?** Maybe something like "dog" or "cat" since we’re talking about pets and a zoo.

**What Does the Language Model Do?**

A **language model (LM)** helps to figure out which word is most likely to appear in that blank. Here’s how:

* It has a **vocabulary**, which is a list of all possible words it knows.
* The model gives each word in its vocabulary a **probability** — a score that shows how likely it is that word fits in the blank.

**Probability Distribution**

In the picture, you see a list of possible words that could fill the blank:

* **lion**: 0.03 probability
* **elephant**: 0.02 probability
* **dog**: 0.45 probability
* **cat**: 0.4 probability
* **panther**: 0.05 probability
* **alligator**: 0.01 probability

The higher the probability, the **more likely** that word fits in the blank. In this case:

* "Dog" (0.45) is the most likely word.
* "Cat" (0.4) is also quite likely.
* "Lion" and "Elephant" have very low chances (0.03 and 0.02).
* "Alligator" is even less likely (0.01).

**How Does the Model Know Which Words to Include?**

The language model only considers words that **make sense in the context**. Here, the model understands we’re talking about **animals** from a zoo, so it’s giving probabilities to words related to that idea.

Words like "car" or "truck" wouldn’t be considered because they don’t fit the context of the zoo and pets.

**Key Points to Remember:**

1. The **language model** looks at the context (previous words) and predicts which word is most likely to come next.
2. It gives a **probability** to each possible word in its vocabulary.
3. The higher the probability, the more likely that word fits in the sentence.
4. It only considers words that make sense in the **context** (in this case, animals).

**What is a Large Language Model (LLM)?**

When we talk about a **large language model**, we are referring to a model that has a huge number of **parameters**.

**What are Parameters/Weights?**

* **Parameters** are the internal settings in a model that get adjusted during training to help it make better predictions.
* Think of parameters like **dials or knobs** that can be fine-tuned. The more parameters a model has, the more fine-tuning it can do, which helps it **understand complex language patterns** better.

**Why is it Called "Large"?**

* The word **large** doesn’t refer to the physical size of the model or the length of the sentences it can handle.
* Instead, it refers to the **number of parameters**. A **large** language model has **millions or even billions** of parameters, allowing it to capture and process very detailed patterns in language.

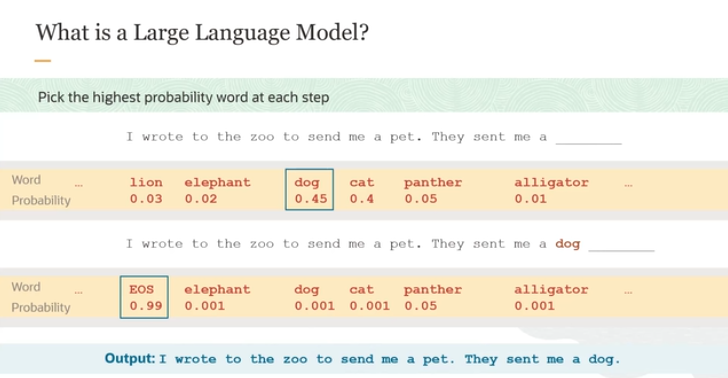
**Is There a Clear Size to Define "Large"?**

No, there’s no exact rule or **threshold/limit** that defines when a model is considered "large" or "small."

* For example, some models with **100 million parameters** might be called large, while others with **billions of parameters** are also considered large.
* It’s more of a **relative term** that indicates the model has a lot of learning capacity, meaning it’s powerful enough to handle complex tasks like writing essays, translating languages, or answering difficult questions.

**Key Point:**

So when we call something a **large language model**, we are mostly highlighting the **scale of its learning power**—thanks to the high number of parameters—rather than giving it an exact size label.



**How Does a Language Model Pick a Word?**

When a large language model (LLM) predicts the next word in a sentence, it doesn’t just guess randomly. It gives **probabilities** to each word in its vocabulary (the list of words it knows), showing how likely each word is to appear next.

**Example From the Slide:**

We have the sentence:  
**"I wrote to the zoo to send me a pet. They sent me a \_\_\_\_\_."**

Since **dog** has the highest probability (45%), the model picks **dog** as the word to fill in the blank.

**What Happens Next?**

Once the model chooses **dog**, it:

1. **Appends "dog"** to the sentence, making the input now:  
   **"I wrote to the zoo to send me a pet. They sent me a dog."**
2. Then, the model **feeds this updated sentence back into itself** to predict what comes next.

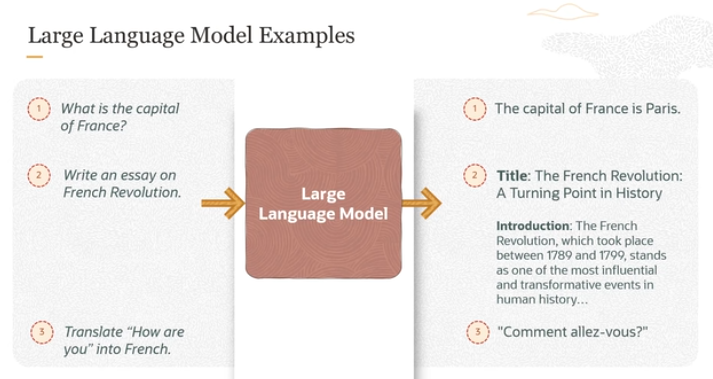
**Changes in Probabilities:**

Now that **"dog"** has been chosen and added to the sentence, the model updates the probabilities of what might come next.

* Words like **elephant** or **lion**, which might have made sense before, now have much lower probabilities because they don’t fit after **"dog."**
* The model now introduces a new token called **EOS (End of Sentence)**, which has a very high probability.

**Final Output:**

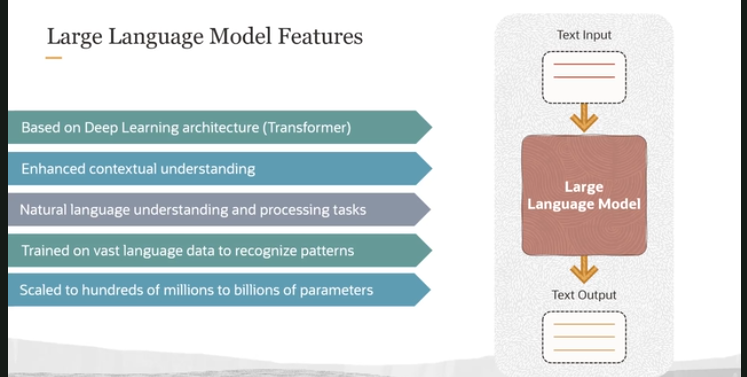
The model produces the full sentence:  
**"I wrote to the zoo to send me a pet. They sent me a dog."**  
Then, the **EOS** token indicates the end of the sentence.



Now what all things can you do with large language models? You can ask questions to a large language model and it can answer your questions. In this case, we are asking it a simple question as to what is the capital of France. And it's returning an answer that capital of France is Paris.

But you could ask it a more complex question. You could give it a puzzle and it would reason over your puzzle and give you all the reasoning and the answer. You can ask it to write an article. Here we are asking it to write an essay on French Revolution and it's returning that essay.

You can ask it to translate text. So you can say something like translate "how are you" in French. And it would return "how are you" in French, [SPEAKING FRENCH]. And that's the formal French translation. So these are some of the examples of where you would use large language models.



**1. LLMs Use Transformer Architecture**

* **Transformers** are a special type of deep learning model that can focus on (or **"pay attention to"**) important parts of the input. This "attention" helps the model better understand what part of a sentence is most relevant when predicting the next word.
* For example, if the input sentence is:  
  "The dog **chased** the ball,"  
  when the model is predicting the next word after "chased," it will focus more on **"dog"** than on other words like "the" or "ball." This helps it understand **context** more effectively.

**2. Enhanced Contextual Understanding**

* Thanks to this **selective attention**, LLMs can understand not just the meaning of each word, but how words relate to one another. This helps the model make more accurate predictions and generate more meaningful text.

Best source for NLP and NLU : <https://www.ibm.com/topics/natural-language-processing>

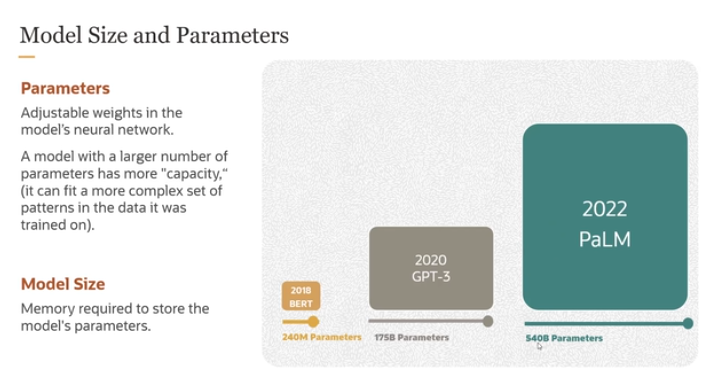
* This ability makes LLMs really good at various tasks, such as:
  + **Question answering** (finding the correct answer to a question)
  + **Language translation** (changing text from one language to another)
  + **Sentiment analysis** (determining the emotion or feeling in text, like whether a review is positive or negative)
  + **Summarization** (shortening long texts while keeping the important information)

**3. LLMs Are Part of Deep Learning**

* **Generative AI** (the AI that creates new content, like text or images) and LLMs belong to **deep learning**, a subset of machine learning.
* In simpler terms, **deep learning** is like teaching a computer to "think" using lots of layers of processing, much like a brain with many neurons.

**4. What Makes Them Large?**

* The **"large"** in **large language model** refers to the number of **parameters** in the model.
* **Parameters** are like adjustable settings or **weights** that the model learns during training to make better predictions.
* LLMs are trained on massive amounts of text data, often from the entire publicly available internet.
* These models can have **hundreds of millions to billions of parameters**, which means they are highly complex and capable of understanding and generating more accurate text.



**1. Parameters in Neural Networks**

* **Parameters** in a model are the **adjustable weights** that the model learns during training. These weights allow the model to make predictions, such as predicting the next word in a sentence.
* As the model sees more data, it adjusts these weights to improve its predictions.

**2. Model Size**

* **Model size** refers to the **memory required** to store all the parameters of the model. The larger the number of parameters, the more memory the model will need.
* For instance, the image shows how models have gotten bigger in recent years:
  + **BERT (2018)** has **240 million parameters**.
  + **GPT-3 (2020)** has **175 billion parameters**.
  + **PaLM (2022)** has a massive **540 billion parameters**.

**3. Growing Number of Parameters**

* Over the past few years, the number of parameters in models has increased **exponentially**, meaning that each newer model has far more parameters than the previous ones.
* While this allows models to learn more complex patterns, having more parameters doesn't always mean the model performs better. Too many parameters can lead to **overfitting**, where the model learns the training data too well and struggles to generalize to new data.

**4. Unknown Parameters**

* Some large language models don’t publicly disclose the number of parameters they use. So, while the image shows the largest model with 540 billion parameters, it’s possible that there are models with even more.

**Summary:**

* **Parameters** are like adjustable settings in a neural network that allow it to make better predictions.
* **Model size** is the memory needed to store all those parameters.
* Models have been growing larger, but bigger doesn’t always mean better. In fact, too many parameters can sometimes cause problems like overfitting.

If we download LLM on our PC so then what will be the pros and cons ?

