

# Overview of Generative Pre-trained Transformer (GPT) Models



**Xavier Morera**

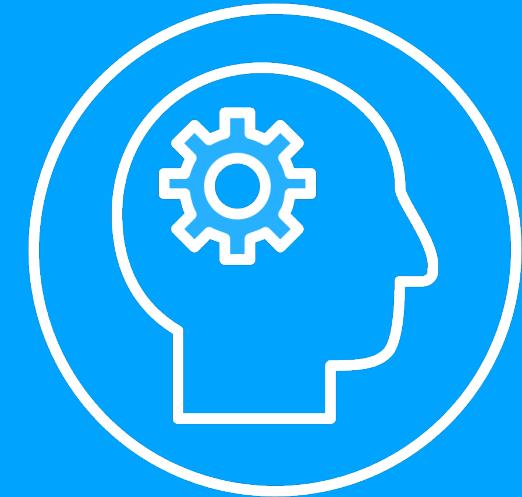
Helping developers create epic applications

[www.xavermorera.com](http://www.xavermorera.com) / [www.bigdatainc.org](http://www.bigdatainc.org) / [www.lupo.ai](http://www.lupo.ai)



**Disclaimer:**  
**This course is not a deep dive**  
**on the architecture of LLMs**





## Objective:

**Learn how to develop a generative application  
using Python and the OpenAI API**





# What is a Generative Pre-trained Transformer (GPT) Model?





**Disclaimer: Not everything that shines is gold**





# Hallucination



Computer Science > Computation and Language

[Submitted on 12 Jun 2017 (v1), last revised 2 Aug 2023 (this version, v7)]

# Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Comments: 15 pages, 5 figures

Subjects: Computation and Language (cs.CL); Machine Learning (cs.LG)

Cite as: arXiv:1706.03762 [cs.CL]

(or arXiv:1706.03762v7 [cs.CL] for this version)

<https://doi.org/10.48550/arXiv.1706.03762> 

## Submission history

From: Llion Jones [[view email](#)]

[v1] Mon, 12 Jun 2017 17:57:34 UTC (1,102 KB)

[v2] Mon, 19 Jun 2017 16:49:45 UTC (1,125 KB)

[v3] Tue, 20 Jun 2017 05:20:02 UTC (1,125 KB)

[v4] Fri, 30 Jun 2017 17:29:30 UTC (1,124 KB)

[v5] Wed, 6 Dec 2017 03:30:32 UTC (1,124 KB)

[v6] Mon, 24 Jul 2023 00:48:54 UTC (1,124 KB)

[v7] Wed, 2 Aug 2023 00:41:18 UTC (1,124 KB)



## Access Paper:

- [Download PDF](#)
- [PostScript](#)
- [Other Formats](#)

([view license](#))

Current browse context:

cs.CL

[< prev](#) | [next >](#)

[new](#) | [recent](#) | [1706](#)

Change to browse by:

cs

[cs.LG](#)

## References & Citations

- [NASA ADS](#)
- [Google Scholar](#)
- [Semantic Scholar](#)

[99 blog links](#) ([what is this?](#))

[DBLP – CS Bibliography](#)

[listing](#) | [bibtex](#)

Ashish Vaswani  
Noam Shazeer  
Niki Parmar  
Jakob Uszkoreit  
Llion Jones

...

[Export BibTeX Citation](#)

[Bookmark](#)



Bibliographic Tools

Code, Data, Media

Demos

Related Papers

About arXivLabs

## Bibliographic and Citation Tools

Bibliographic Explorer ([What is the Explorer?](#))

Litmaps ([What is Litmaps?](#))

# GPT

Generative

Pre-trained

Transformer



# Architecture of Large Language Models



# Architecture of Large Language Models

Neural Networks

Attention Mechanisms



# Architecture of Large Language Models

Transformer



## Computer Science &gt; Computation and Language

[Submitted on 12 Jun 2017 ([v1](#)), last revised 2 Aug 2023 (this version, v7)]

## Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Comments: 15 pages, 5 figures

Subjects: Computation and Language (cs.CL); Machine Learning (cs.LG)

Cite as: arXiv:1706.03762 [cs.CL]

(or arXiv:1706.03762v7 [cs.CL] for this version)

<https://doi.org/10.48550/arXiv.1706.03762> 

## Submission history

From: Llion Jones [[view email](#)]

[v1] Mon, 12 Jun 2017 17:57:34 UTC (1,102 KB)

[v2] Mon, 19 Jun 2017 16:49:45 UTC (1,125 KB)

[v3] Tue, 20 Jun 2017 05:20:02 UTC (1,125 KB)

[v4] Fri, 30 Jun 2017 17:29:30 UTC (1,124 KB)

[v5] Wed, 6 Dec 2017 03:30:32 UTC (1,124 KB)

[v6] Mon, 24 Jul 2023 00:48:54 UTC (1,124 KB)

[v7] Wed, 2 Aug 2023 00:41:18 UTC (1,124 KB)



## Access Paper:

- [Download PDF](#)
- [PostScript](#)
- [Other Formats](#)

[\(view license\)](#)

Current browse context:

cs.CL

< prev | > nextnew | recent | 1706

Change to browse by:

cs

cs.LG

## References &amp; Citations

- [NASAADS](#)
- [Google Scholar](#)
- [Semantic Scholar](#)

[98 blog links](#) (what is this?)

## DBLP - CS Bibliography

listing | bibtexAshish Vaswani  
Noam Shazeer  
Niki Parmar  
Jakob Uszkoreit  
Llion Jones

...

[Export BibTeX Citation](#)

## Bookmark



## Bibliographic Tools

Code, Data, Media

Demos

Related Papers

About arXivLabs

## Bibliographic and Citation Tools

 [Bibliographic Explorer](#) ([What is the Explorer?](#)) [Litmaps](#) ([What is Litmaps?](#)) [scite Smart Citations](#) ([What are Smart Citations?](#))

Provided proper attribution is provided, Google hereby grants permission to reproduce the tables and figures in this paper solely for use in journalistic or scholarly works.

# Attention Is All You Need

## Attention Is All You Need

**Ashish Vaswani\***  
Google Brain  
[avaswani@google.com](mailto:avaswani@google.com)

**Noam Shazeer\***  
Google Brain  
[noam@google.com](mailto:noam@google.com)

**Niki Parmar\***  
Google Research  
[nikip@google.com](mailto:nikip@google.com)

**Jakob Uszkoreit\***  
Google Research  
[usz@google.com](mailto:usz@google.com)

**Llion Jones\***  
Google Research  
[llion@google.com](mailto:llion@google.com)

**Aidan N. Gomez\* †**  
University of Toronto  
[aidan@cs.toronto.edu](mailto:aidan@cs.toronto.edu)

**Lukasz Kaiser\***  
Google Brain  
[lukaszkaiser@google.com](mailto:lukaszkaiser@google.com)

**Illia Polosukhin\* ‡**  
[illia.polosukhin@gmail.com](mailto:illia.polosukhin@gmail.com)

### Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

**The transformer is the foundation for many LLMs due to its parallel processing capabilities and efficient attention mechanisms.**

\*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.



# Transformer Architecture

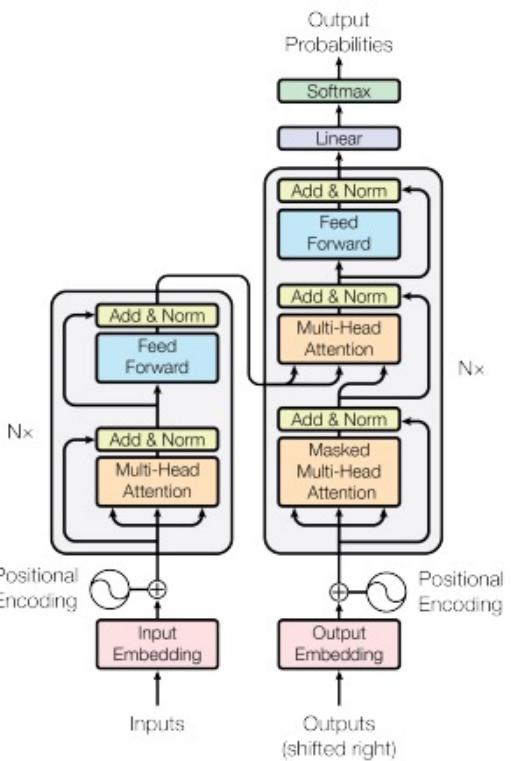


Figure 1: The Transformer - model architecture.

The Transformer follows this overall architecture using stacked self-attention and point-wise, fully connected layers for both the encoder and decoder, shown in the left and right halves of Figure 1, respectively.

## 3.1 Encoder and Decoder Stacks

**Encoder:** The encoder is composed of a stack of  $N = 6$  identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network. We employ a residual connection [11] around each of the two sub-layers, followed by layer normalization [1]. That is, the output of each sub-layer is  $\text{LayerNorm}(x + \text{Sublayer}(x))$ , where  $\text{Sublayer}(x)$  is the function implemented by the sub-layer itself. To facilitate these residual connections, all sub-layers in the model, as well as the embedding layers, produce outputs of dimension  $d_{\text{model}} = 512$ .

**Decoder:** The decoder is also composed of a stack of  $N = 6$  identical layers. In addition to the two sub-layers in each encoder layer, the decoder inserts a third sub-layer, which performs multi-head attention over the output of the encoder stack. Similar to the encoder, we employ residual connections around each of the sub-layers, followed by layer normalization. We also modify the self-attention sub-layer in the decoder stack to prevent positions from attending to subsequent positions. This masking, combined with fact that the output embeddings are offset by one position, ensures that the predictions for position  $i$  can depend only on the known outputs at positions less than  $i$ .

## 3.2 Attention

An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum



# Transformer Architecture

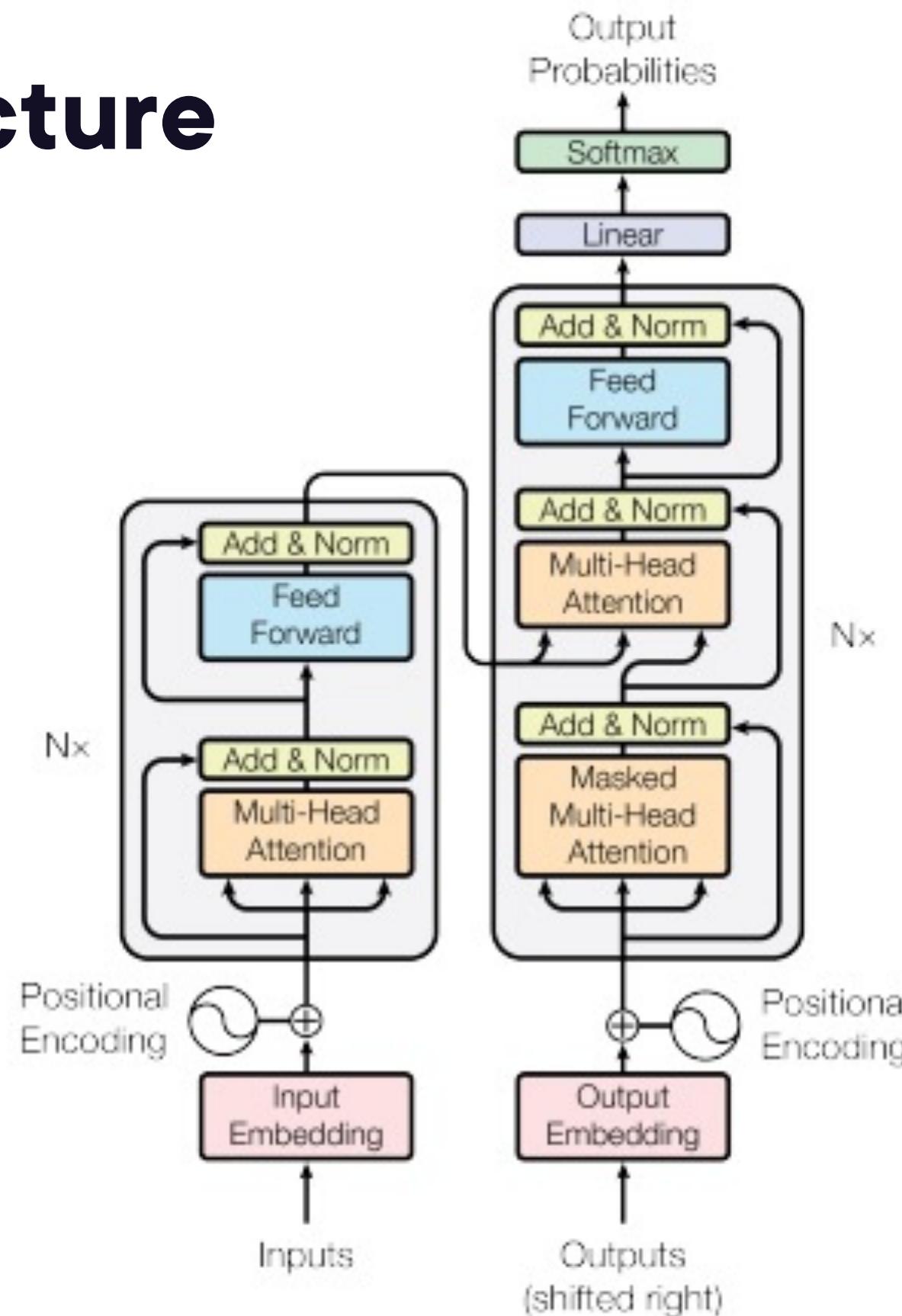
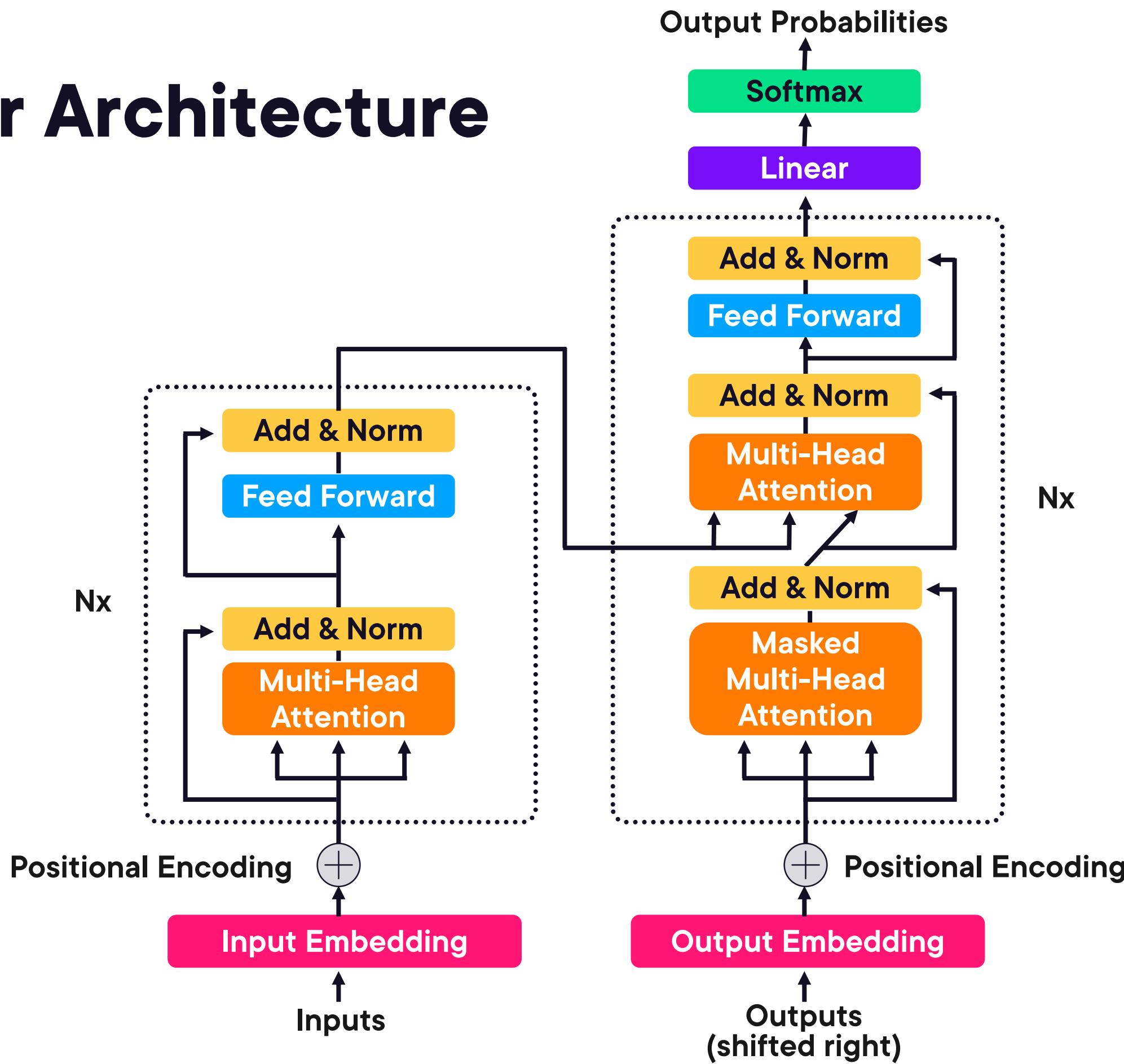


Figure 1: The Transformer - model architecture.

The Transformer follows this overall architecture using stacked self-attention and point-wise, fully connected layers for both the encoder and decoder, shown in the left and right halves of Figure 1.

# Transformer Architecture



# Transformer Architecture

**Encoder-decoder  
structure**

**Self-attention  
mechanism**

**Multi-head  
attention**

**Positional  
encoding**

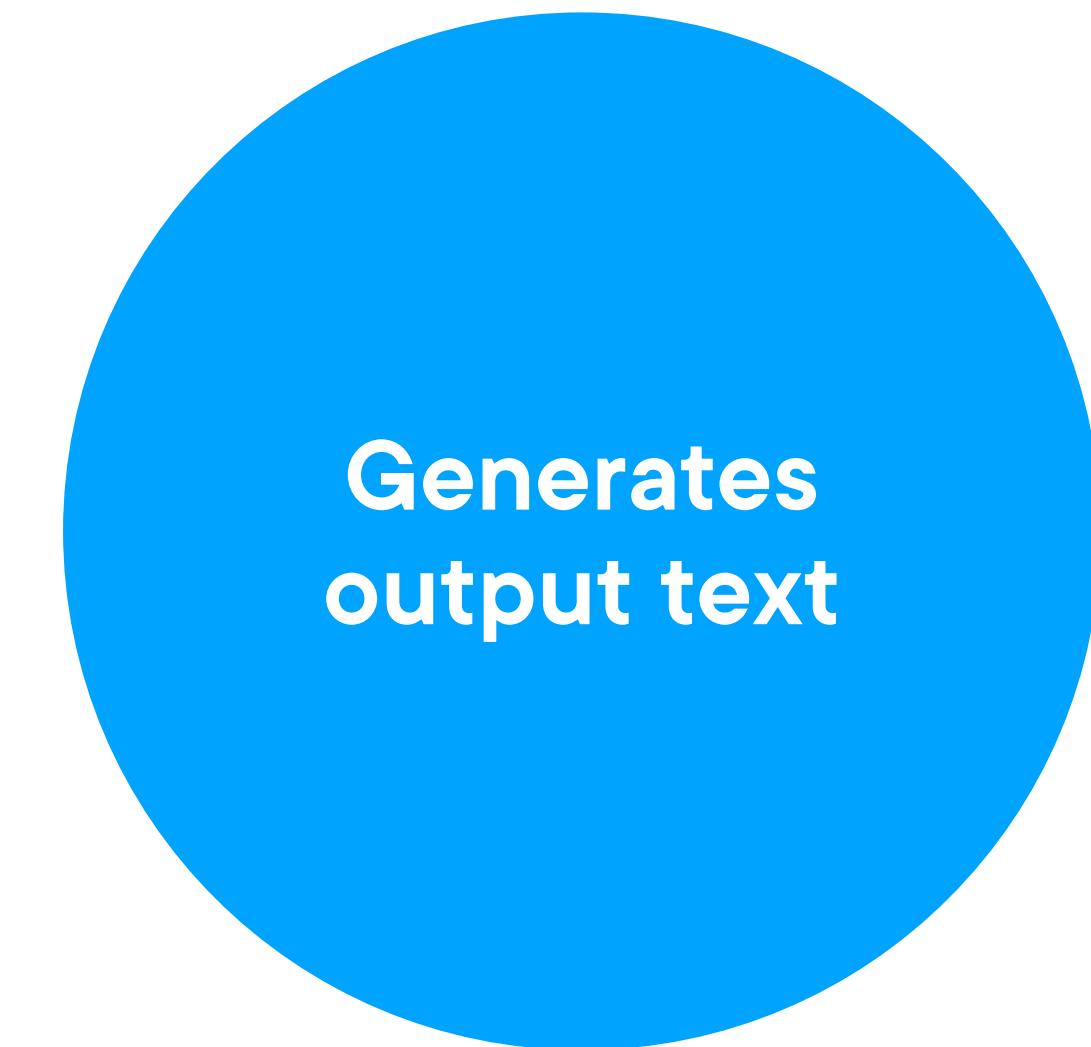
**Feed-forward  
neural networks**



# Encoder-Decoder Structure



**Encoder**



**Decoder**



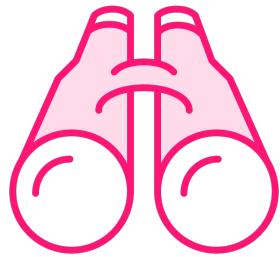
# **Self-Attention Mechanism**

**Allows the model to weigh the importance of different words in a sentence relative to each other**

**Enables the model to capture contextual relationships and dependencies effectively**



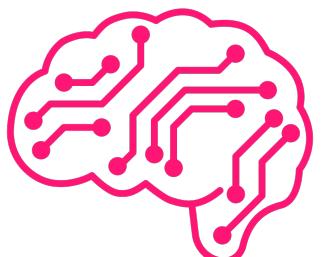
# Self-Attention Mechanism



**Each word in a sentence looks at all the other words in the same sentence**



**Figures out how much attention or importance it should give to each of those other words**



**Then, it combines information from all those other words based on their importance to get a better understanding of the whole sentence**

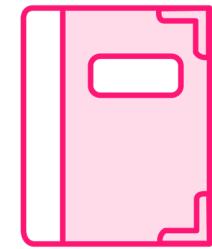


# Multi-Head Attention

To capture different types of relationships, the Transformer employs multi-head attention

Performs self-attention multiple times in parallel, allowing the model to focus on different parts of the input simultaneously

# Multi-Head Attention



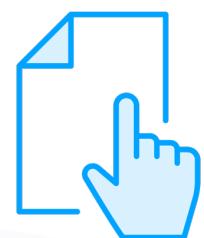
**Imagine that you have a complex book to read**



**You divide the book between you and a few friends**



**Each one of you reads a story and understands it**



**And then your friends help you understand the whole story**



# **Positional Encoding**

**The Transformer doesn't inherently understand the sequential order of words**

**Positional encoding is added to the input embeddings**

**Helps the model differentiate between the positions of words in the sequence**



# Positional Encoding

In simpler words...



# Positional Encoding

In simpler words...

1

2

3

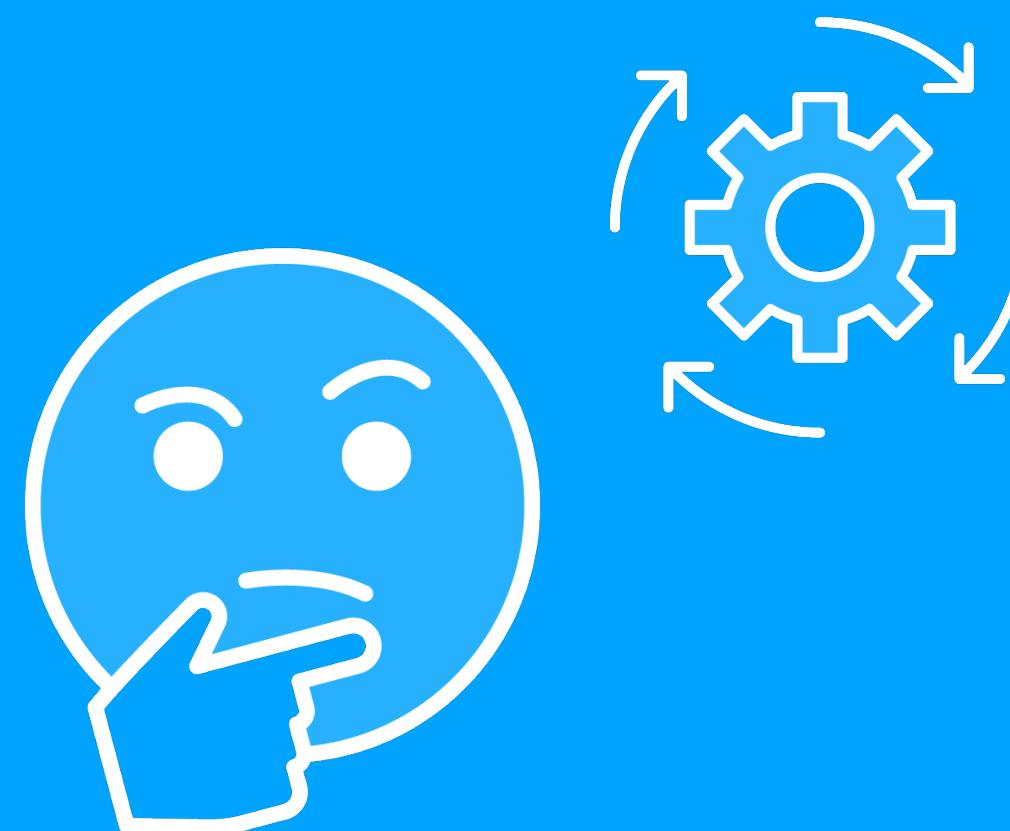


# **Feed-forward Neural Networks**

**Each layer of the Transformer includes feed-forward neural networks**

**Process the attention outputs and apply non-linear transformations to capture complex patterns**

# Feed-forward Neural Networks



# | Transformer Architecture (In Simpler Words)



# At a High Level

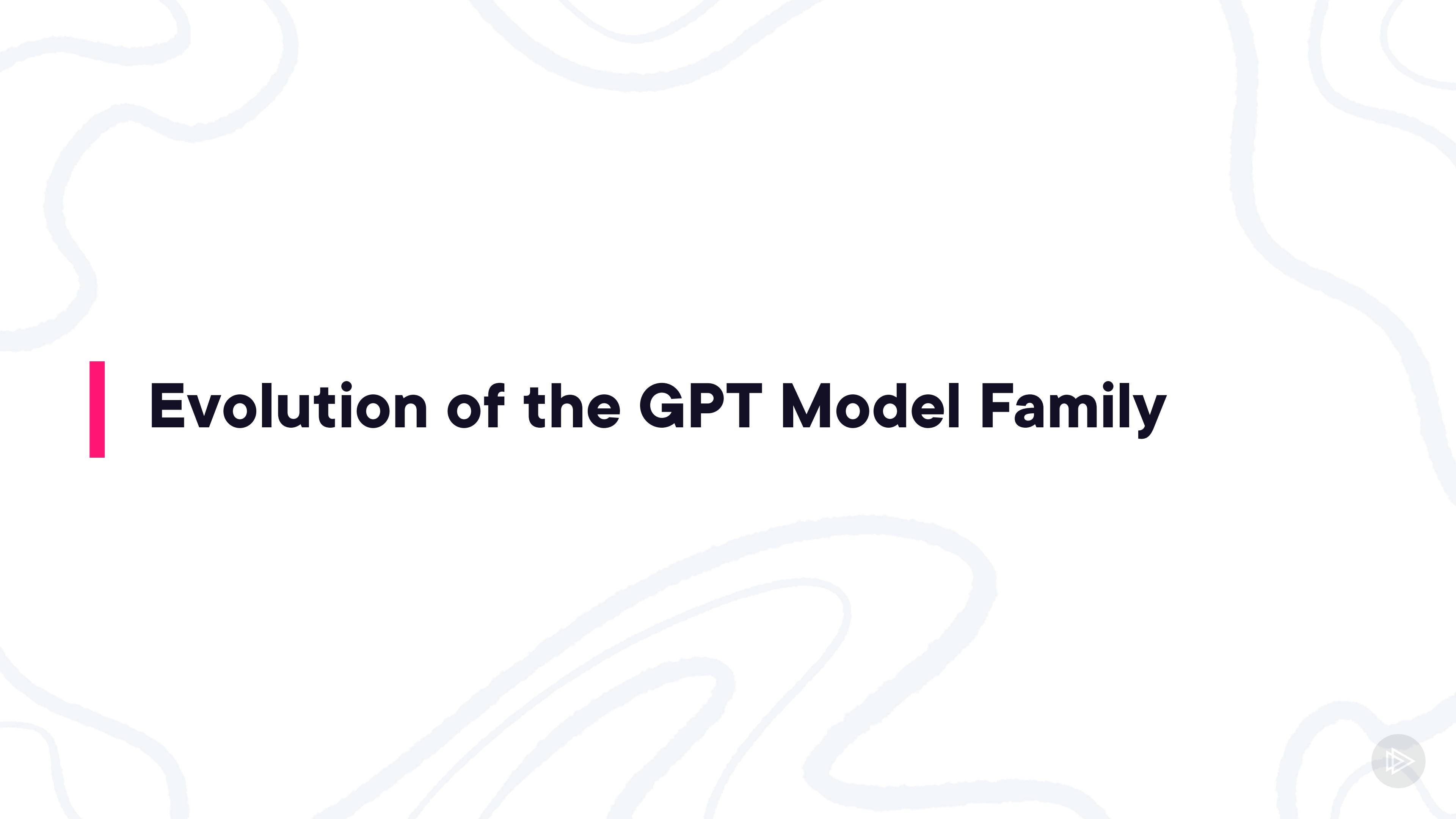
Splitting sentences

Attention

Mixing information

Repeat for understanding





# Evolution of the GPT Model Family



# Evolution of the GPT Model Family

## GPT-1 (2018)

117 million params

Trained on ~40GB

Context length of 512 tokens

## GPT-3 (2020)

175 billion params

Trained on over 570 GB of data

Notable improvements

## GPT-4 (2023)

Largest milestone in the series

Multimodal capabilities

Superior performance

## GPT-2 (2019)

1.5 billion params

Focused on improved training

Double the input size

Enhanced performance

## GPT-3.5 (2022)

Multiple variants

Based on GPT-3 with an emphasis on ethical considerations

Additional improvements

## The Future...

Let's wait and see



# Evolution of the GPT Model Family

## GPT-1 (2018)

117 million params  
Trained on ~40GB  
Context length of 512 tokens

## GPT-3 (2020)

175 billion params  
Trained on over 570 GB of data  
Notable improvements

## GPT-4 (2023)

Largest milestone in the series  
Multimodal capabilities  
Superior performance

## GPT-2 (2019)

1.5 billion params  
**Focused on improved training**  
**Double the input size**  
**Enhanced performance**

## GPT-3.5 (2022)

Multiple variants  
Based on GPT-3 with an emphasis on ethical considerations  
Additional improvements

## The Future...

Let's wait and see



# Evolution of the GPT Model Family

## GPT-1 (2018)

117 million params  
Trained on ~40GB  
Context length of 512 tokens

## GPT-3 (2020)

175 billion params  
Trained on over 570 GB of data  
**Notable improvements**

## GPT-4 (2023)

Largest milestone in the series  
Multimodal capabilities  
Superior performance

## GPT-2 (2019)

1.5 billion params  
Focused on improved training  
Double the input size  
Enhanced performance

## GPT-3.5 (2022)

Multiple variants  
Based on GPT-3 with an emphasis on ethical considerations  
Additional improvements

## The Future...

Let's wait and see



# Evolution of the GPT Model Family

## GPT-1 (2018)

117 million params  
Trained on ~40GB  
Context length of 512 tokens

## GPT-3 (2020)

175 billion params  
Trained on over 570 GB of data  
Notable improvements

## GPT-4 (2023)

Largest milestone in the series  
Multimodal capabilities  
Superior performance

## GPT-2 (2019)

1.5 billion params  
Focused on improved training  
Double the input size  
Enhanced performance

## GPT-3.5 (2022)

Multiple variants  
Based on GPT-3 with an emphasis on ethical considerations  
Additional improvements

## The Future...

Let's wait and see



# Evolution of the GPT Model Family

## GPT-1 (2018)

117 million params  
Trained on ~40GB  
Context length of 512 tokens

## GPT-3 (2020)

175 billion params  
Trained on over 570 GB of data  
Notable improvements

## GPT-4 (2023)

**Largest milestone in the series**  
**Multimodal capabilities**  
**Superior performance**

## GPT-2 (2019)

**1.5 billion params**  
Focused on improved training  
Double the input size  
Enhanced performance

## GPT-3.5 (2022)

Multiple variants  
Based on GPT-3 with an emphasis on ethical considerations  
Additional improvements

## The Future...

Let's wait and see



# Evolution of the GPT Model Family

## GPT-1 (2018)

117 million params  
Trained on ~40GB  
Context length of 512 tokens

## GPT-3 (2020)

175 billion params  
Trained on over 570 GB of data  
Notable improvements

## GPT-4 (2023)

Largest milestone in the series  
Multimodal capabilities  
Superior performance

## GPT-2 (2019)

**1.5 billion params**  
Focused on improved training  
Double the input size  
Enhanced performance

## GPT-3.5 (2022)

Multiple variants  
Based on GPT-3 with an emphasis on ethical considerations  
Additional improvements

**The Future...**  
Let's wait and see



# Impact of Model Size on Performance



# Impact of Model Size on Performance



**Enhanced contextual understanding**



**Improved text generation**



**Complex task handling**



**Broadened multilingual support**



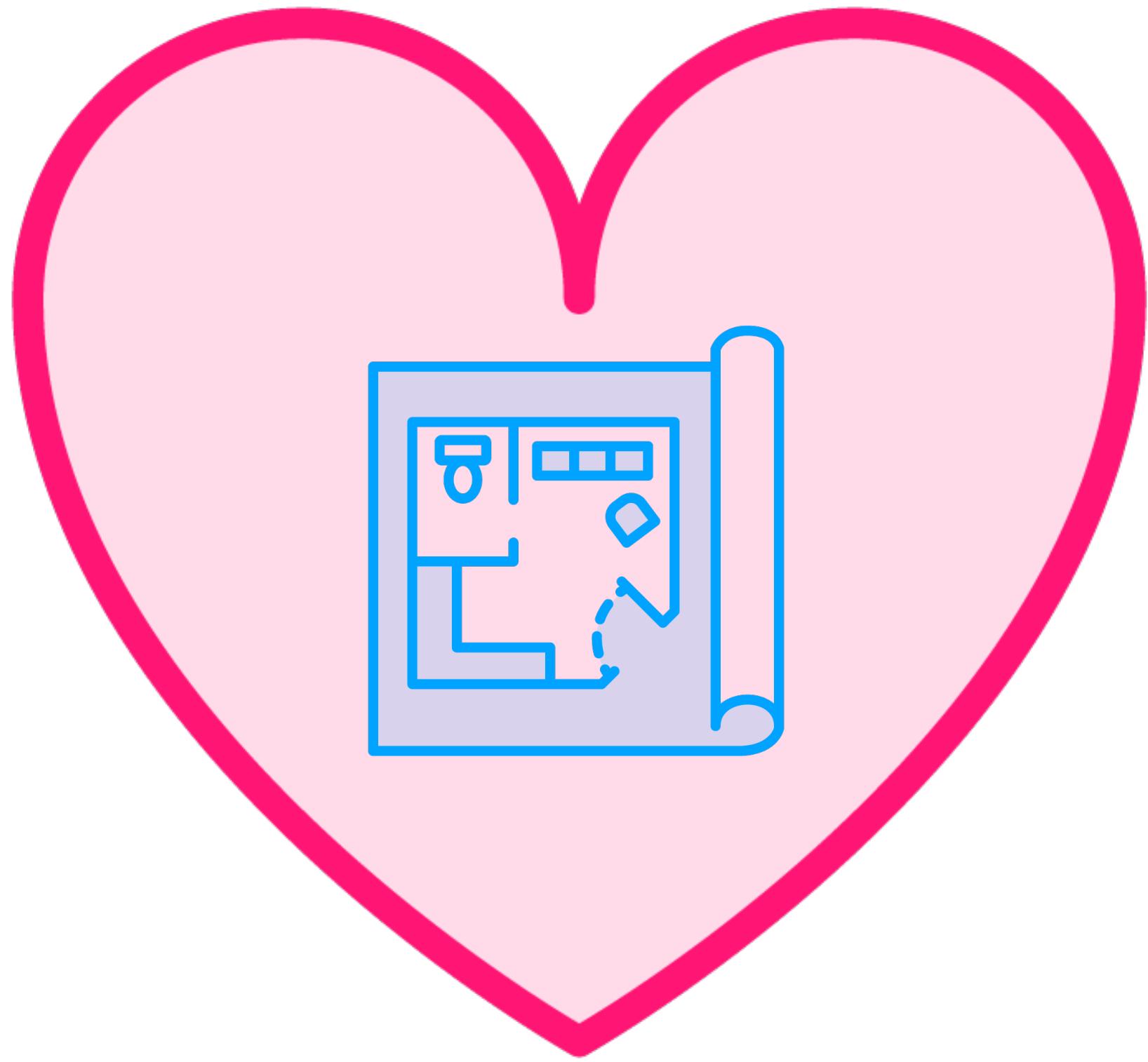
**Challenges and trade-offs**



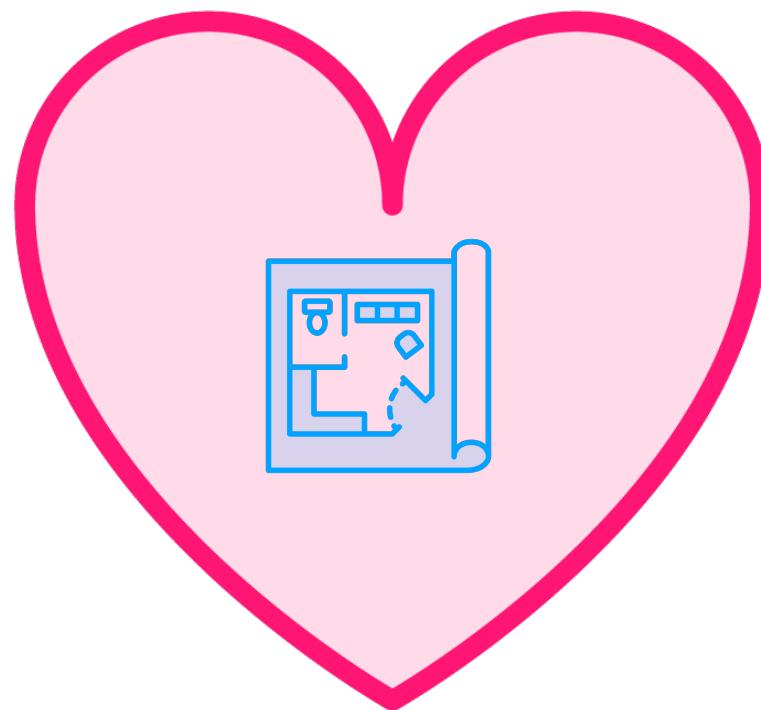
# **Basics of Generative Pre-trained Transformer (GPT) Models**



# Understanding GPT Models



# Understanding GPT Models



**Pivotal advancement in NLP tasks**

**Transformers employ self-attention mechanisms**

- Capture intricate contextual relationships among words within sentences



# The Process: Making It All Work

Models learn language patterns  
and contextual comprehension  
from extensive text data

## Pre-training

Tailors the model to specific tasks,  
making it adaptable and versatile

## Fine-tuning



# **Unveiling Attention Mechanisms**

## **The attention mechanism**

- Empowers GPT models to weigh the significance of each word within a sentence
- In relation to others

## **This dynamic understanding of context**

- Enhances the models' ability to generate coherent and contextually fitting text



# Significance of Embeddings

GPT models employ embeddings to represent words as continuous vectors in a high-dimensional space

These embeddings capture semantic relationships

- Enabling the model to grasp word meanings and contextual intricacies



# Autoregressive Text Generation

**GPT models generate text autoregressively**

- Predicting each word based on the previous words

**Ensures that generated text flows cohesively**

- Maintaining contextual coherence



# Strategies for Text Sampling

Different sampling strategies influence the creativity and randomness of generated text

- You can set how “creative” the response should be

Techniques to control output

- Temperature scaling and top-k sampling



# Use Cases and Applications

**Text Generation**

**Language  
Translation**

**Question Answering**





# | Overview of Rival Language Models (LLMs)

# Talk to Claude

Email

Continue with email

OR

 Continue with Google

Claude.ai is currently in open beta and usage of the platform  
may be limited for unpaid users.

## Claude for Business

Claude is a next-generation AI assistant for your tasks, no matter the scale. Our API is currently being offered to a limited set of customers and researchers.

Browse our products

## Constitutional AI

Constitutional AI (CAI) shapes the outputs of AI systems according to a set of principles, with the goal of making a helpful, harmless, and honest AI assistant.

Learn about CAI

What is Claude?

XM

I'm Claude, an AI assistant created by Anthropic.

AI

Copy



What can you do?

XM

I'm designed to be helpful, harmless, and honest. I can have natural conversations, answer questions factually, and provide useful information to users.

AI

Copy



Are you better than ChatGPT?

XM

I'm an different AI system created by a different company, so it's hard to make a direct comparison. ChatGPT is very impressive in many ways, but all AI systems have strengths and limitations. My goal is to be as helpful, harmless and honest as possible when conversing with users.

AI

Copy



What are your key advantages over the other LLMs?

XM

I don't have a full comparison to other large language models, but some key things about me:

- I was created by Anthropic to be helpful, harmless, and honest through a technique called Constitutional AI. This focuses me on being a positive conversational agent.
- I have a more advanced natural language processing system compared to other chatbots, so I can understand and respond to complex conversations more naturally.

Message Claude...



XM

## What is Claude? ▾

answer questions factually, and provide useful information to users.

AI

Copy



Are you better than ChatGPT?

XM

I'm an different AI system created by a different company, so it's hard to make a direct comparison. ChatGPT is very impressive in many ways, but all AI systems have strengths and limitations. My goal is to be as helpful, harmless and honest as possible when conversing with users.

AI

Copy



What are your key advantages over the other LLMs?

XM

I don't have a full comparison to other large language models, but some key things about me:

- I was created by Anthropic to be helpful, harmless, and honest through a technique called Constitutional AI. This focuses me on being a positive conversational agent.
- I have a more advanced natural language processing system compared to other chatbots, so I can understand and respond to complex conversations more naturally.
- I have training and techniques to avoid various types of harmful, biased or unreliable responses that can occur in AI systems.
- I'm constantly learning and improving based on my conversations to become more useful over time.

Overall I aim for useful, benign and trustworthy dialog. But capabilities of AI systems are rapidly evolving, so direct comparisons are difficult. My focus is on having the most positive and productive conversations I can.

AI

Copy ⌂ Retry



Message Claude...



AI

# Introducing Pathways: A next-generation AI architecture

Oct 28, 2021

5 min read

Too often, machine learning systems overspecialize at individual tasks, when they could excel at many. That's why we're building Pathways—a new AI architecture that will handle many tasks at once, learn new tasks quickly and reflect a better understanding of the world.

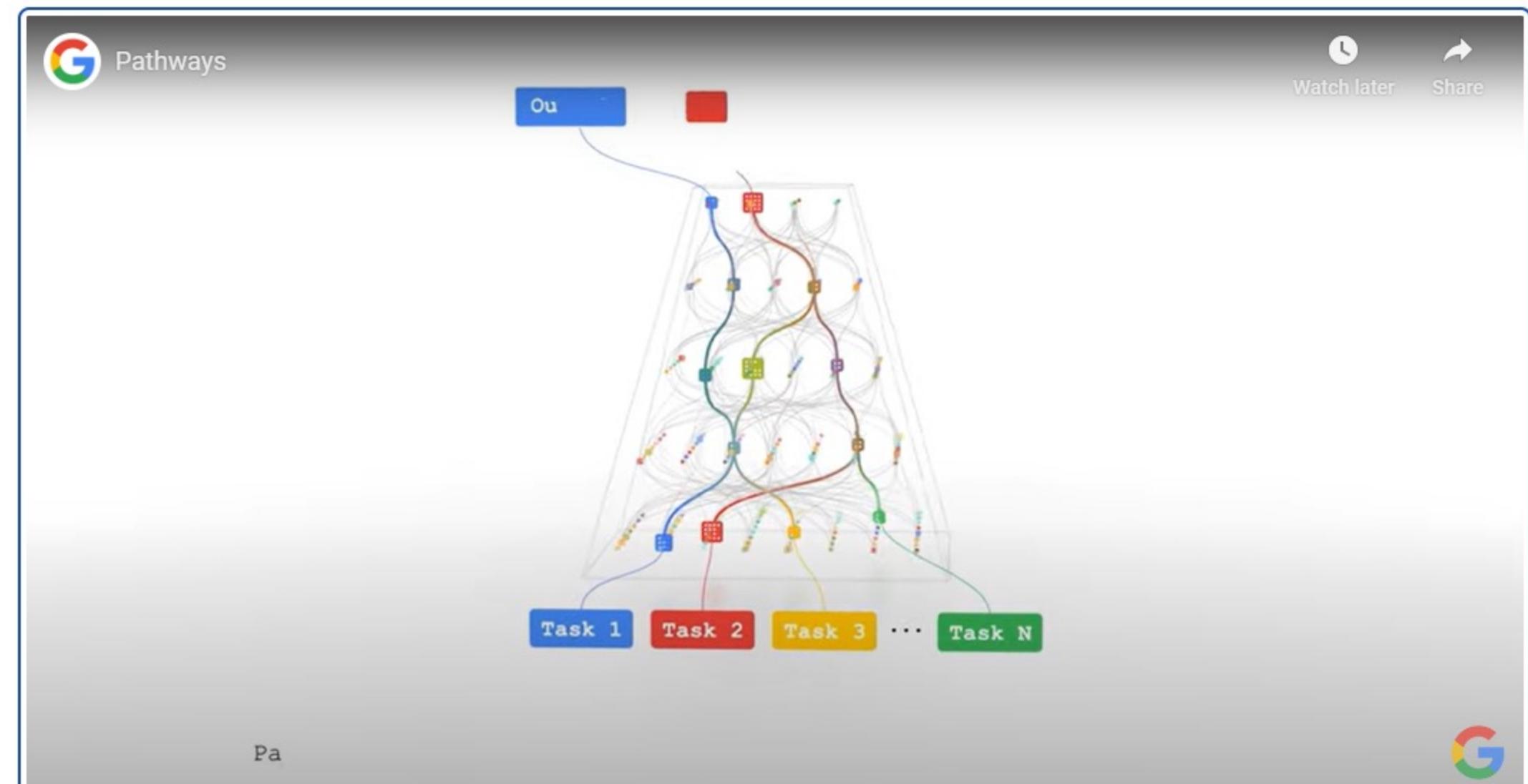


Jeff Dean

Google Senior Fellow and SVP, Google Research



Share



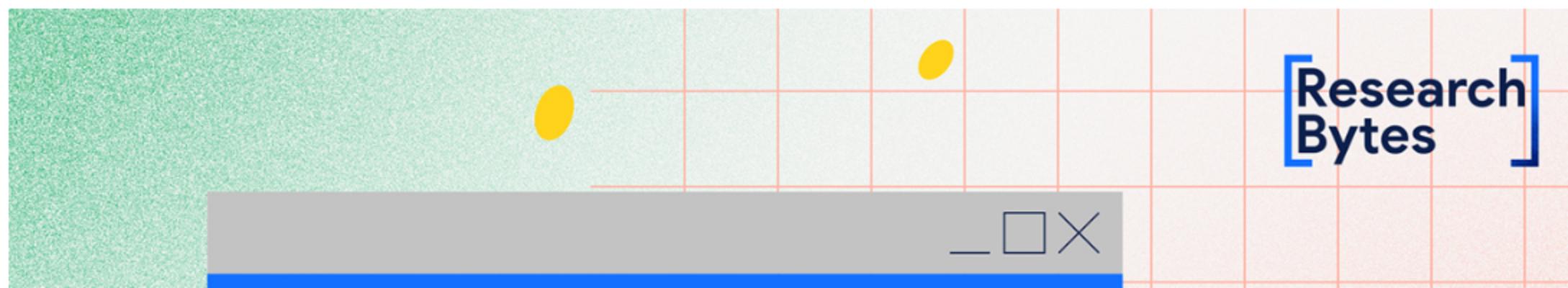
AI ACROSS GOOGLE:

# PaLM 2

PaLM 2 is our next generation large language model that builds on Google's legacy of breakthrough research in machine learning and responsible AI.

It excels at advanced reasoning tasks, including code and math, classification and question answering, translation and multilingual proficiency, and natural language generation better than our previous state-of-the-art LLMs, including PaLM. It can accomplish these tasks because of the way it was built – bringing together compute-optimal scaling, an improved dataset mixture, and model architecture improvements.

PaLM 2 is grounded in Google's approach to building and deploying AI responsibly. All versions of PaLM 2 are evaluated rigorously for potential harms and biases, capabilities and downstream uses in research and in-product applications. PaLM 2 is used in other state-of-the-art models, like [Sec-PaLM](#). We continue to implement the latest versions of PaLM 2 in generative AI tools like the PaLM API and Bard.



Research  
Bytes

+ New chat

Recent

PaLM vs. Bard



What is the difference between PaLM and Bard?



View other drafts



PaLM (Pathways Language Model) is a large language model (LLM) developed by Google AI. It was trained on a massive dataset of text and code, and can generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way. PaLM is still under development, but it has learned to perform many kinds of tasks, including:

- Following instructions and completing requests thoughtfully
- Answering questions in a comprehensive and informative way, even if they are open ended, challenging, or strange
- Generating different creative text formats of text content, like poems, code, scripts, musical pieces, email, letters, etc.

Bard is also an LLM developed by Google AI, but it is based on a newer model called PaLM 2. PaLM 2 is better at common sense reasoning, logic and mathematics compared to previous models. Bard also has access to and processes information from the real world through Google Search, which allows it to provide more comprehensive and up-to-date answers to your questions.

In summary, here are the main differences between PaLM and Bard:

- **Model:** Bard is based on PaLM 2, a newer and more advanced model than PaLM.
- **Capabilities:** Bard is better at common sense reasoning, logic, and mathematics than PaLM. Bard also has access to and processes information from the real world through Google Search.
- **Use cases:** Bard is well-suited for tasks that require common sense reasoning, logic, mathematics, or access to real-world information. PaLM is well-suited for tasks that require generating creative text formats or translating languages.

It is important to note that both PaLM and Bard are still under development, and their capabilities are constantly improving.



Enter a prompt here



# Introducing Llama 2

The next generation of our  
open source large language model

Llama 2 is available for free for research and commercial use.

[Download the Model](#)[↓ Learn more about Llama 2](#)[The Model](#)[Partnerships](#)[Responsibility](#)[Download the Model](#)[Resources](#)

## Inside the model

This release includes model weights and starting code for pretrained and fine-tuned Llama language models (Llama Chat, Code Llama) — ranging from 7B to 70B parameters.

# The Game-Changing Role of Large Language Models



# Demo



**Demo from ChatGPT to the OpenAI API**

