- LLMs
- How to use LLMs
- Some Interesting Tips
- Q&A

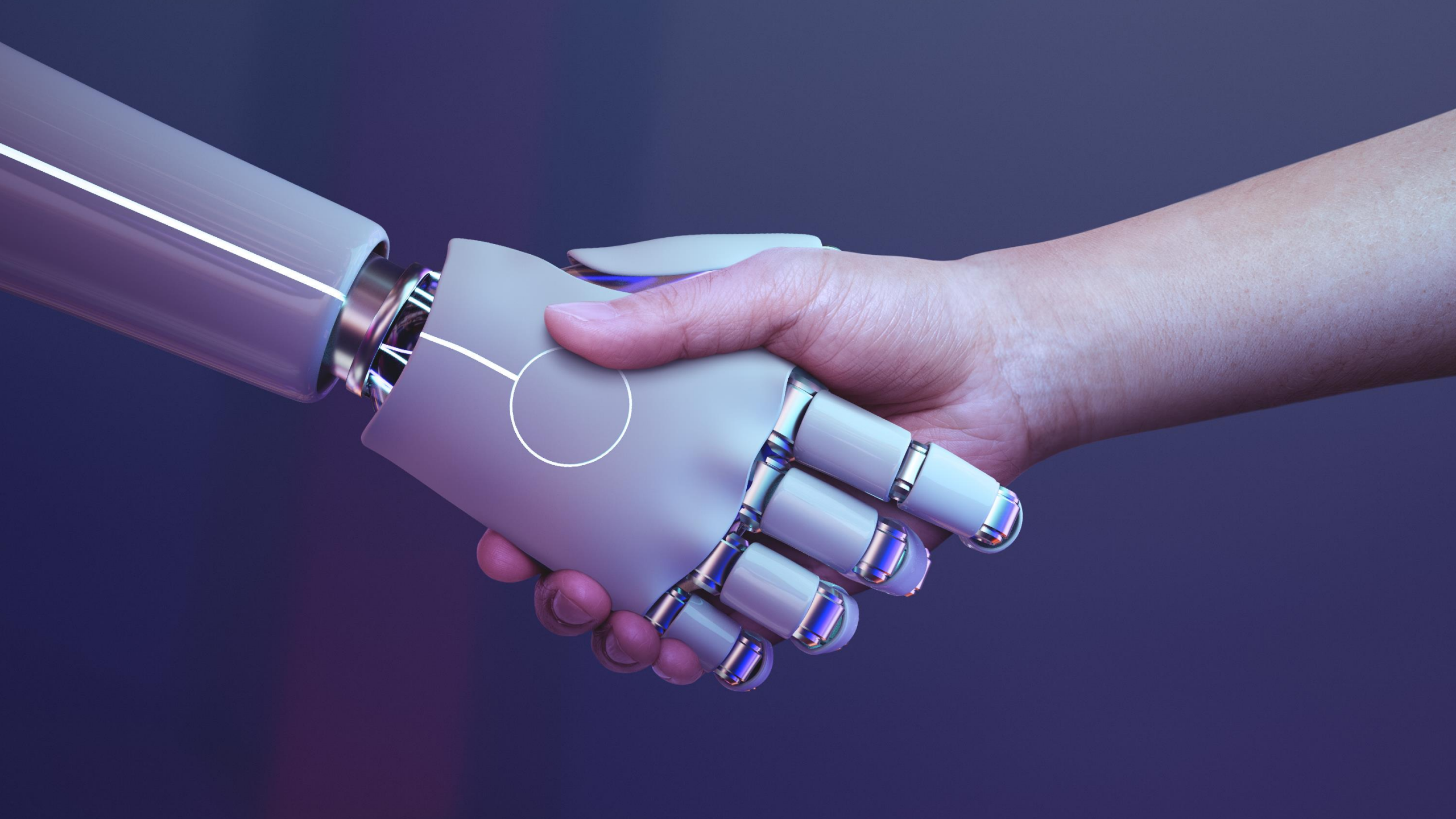
LLMs

Large Language Models

Ryan Heida

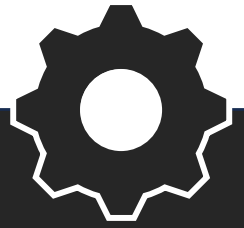
AI.Mastery.Student


LLMs

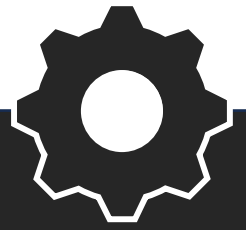


ChatGPT
was trained
on over 300 billion
words!

LLMs Timeline



LLMs Timeline

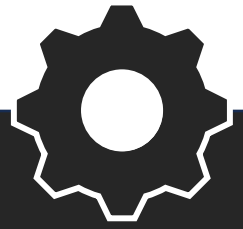


Pre-Transformer

1990 - 2017

RNN, LSTM, NMT, ELMo model

LLMs Timeline



Pre-Transformer

1990 - 2017

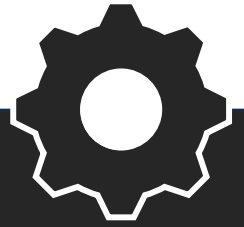
RNN, LSTM, NMT, ELMo model

Transformer

2017 - 2018

The rise of Transformers - No specific LLMs Yet…

LLMs Timeline



Pre-Transformer

1990 - 2017

RNN, LSTM, NMT, ELMo model

Transformer

2017 - 2018

The rise of Transformers - No specific LLMs Yet…

Post-Transformer

2018 - Present

GPT-1, BERT – GPT-2, XLNet – GPT3, T5 - BLOOM, Codex, GPT-4

Attention Is All You Need

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ 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 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

Attention Is All You Need

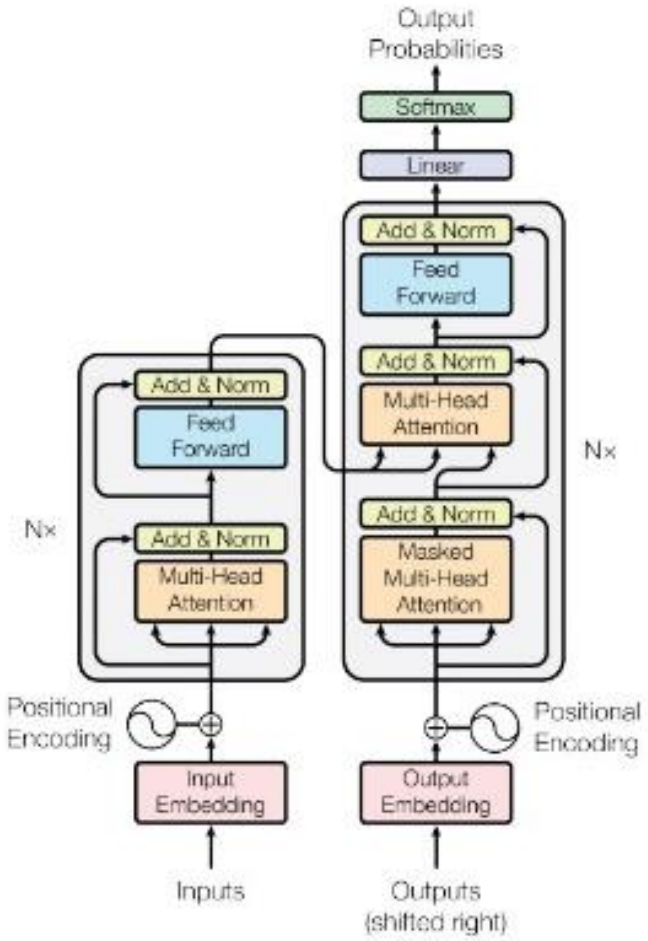


Figure 1: The Transformer - model architecture.

wise fully connected feed-forward network. We employ a residual connection [10] around each of the two sub-layers, followed by layer normalization [1]. That is, the output of each sub-layer is

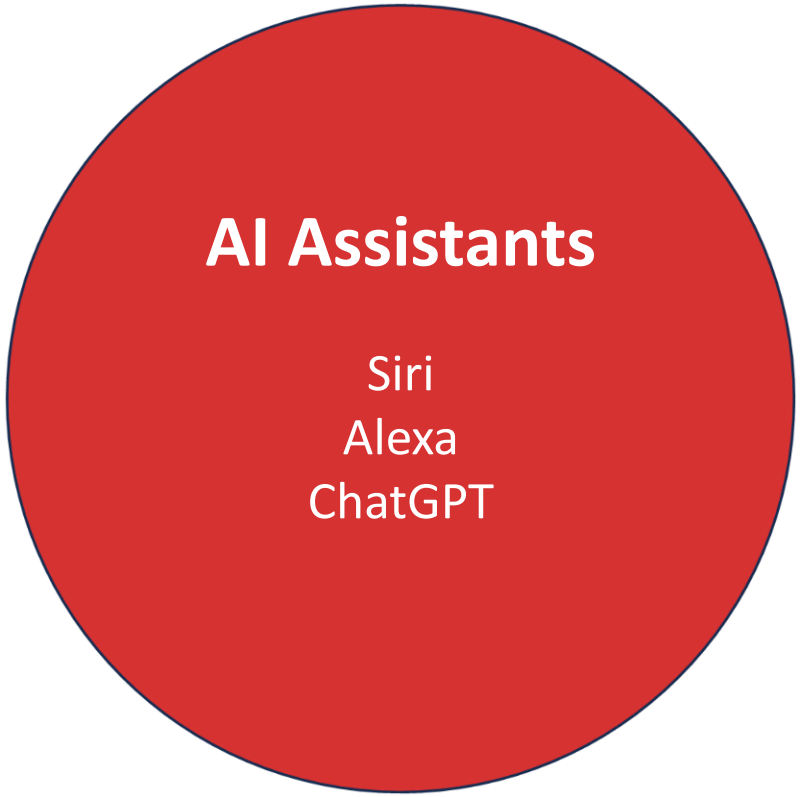
Why Are LLMs Transformative?

Why Are LLMs Transformative?

Why Are LLMs Transformative?

Why Are LLMs Transformative?

LLMs in Everyday Life



LLMs in Everyday Life

LLMs in Everyday Life

LLMs in Everyday Life

How LLMs Work

Next Word Prediction

INPUT

OUTPUT

How LLMs Work

Next Word Prediction

The Car is ----

How LLMs Work

Next Word Prediction

The Car is ----

Red

How LLMs Work

Next Word Prediction

The Car is red

How LLMs Work

Next Word Prediction

Tell a story.

How LLMs Work

Next Word Prediction

Tell a story.

Once

How LLMs Work

Next Word Prediction

Tell a story. Once

How LLMs Work

Next Word Prediction

Tell a story. Once

upon

How LLMs Work

Next Word Prediction

Tell a story. Once upon

How LLMs Work

Next Word Prediction

Tell a story. Once upon

a time

How LLMs Work

Next Word Prediction

Tell a story. Once upon a time

.....

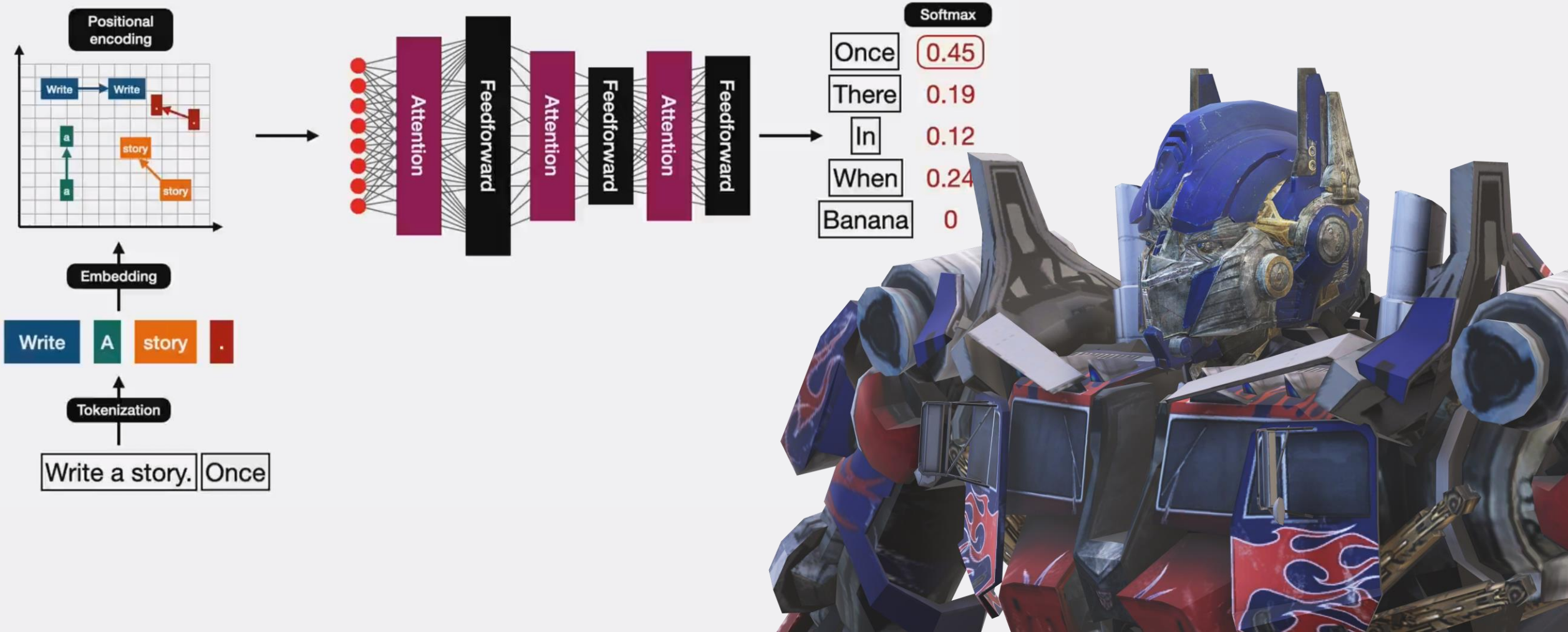
A little deeper

How LLMs Work

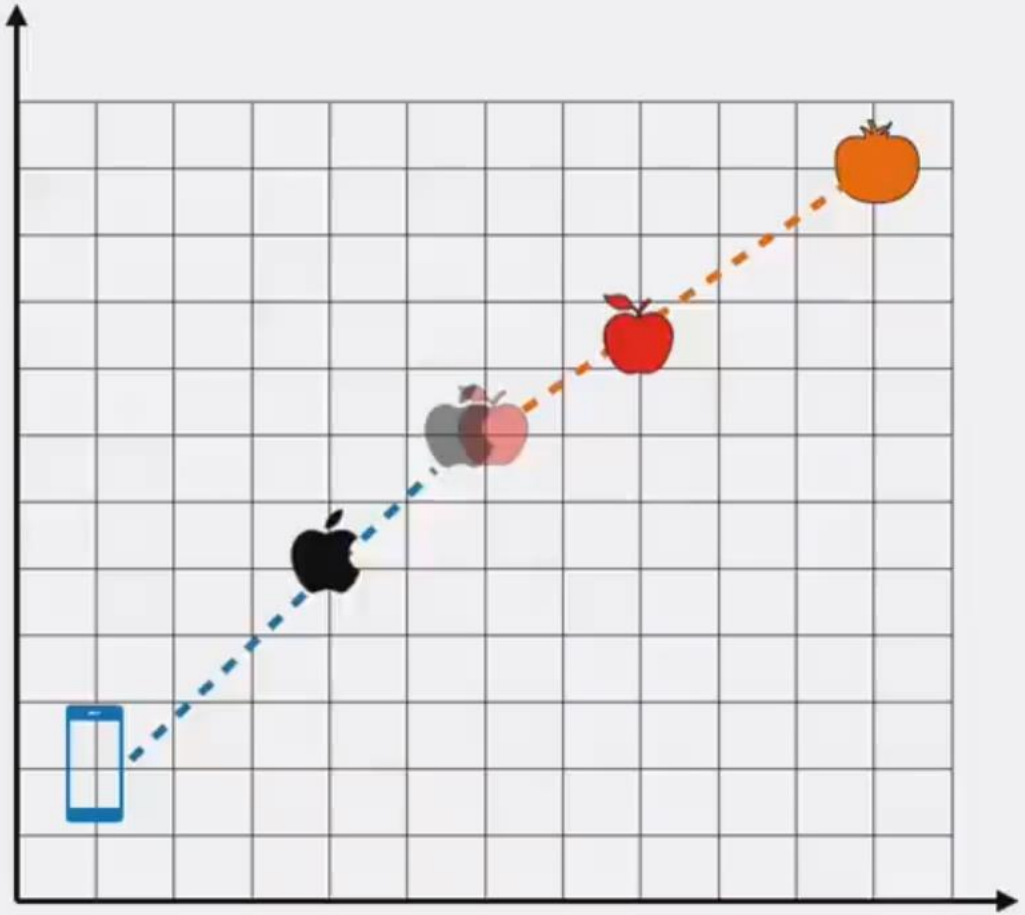
Write a story.

A little deeper

How LLMs Work



Words pulling words



please buy an apple and an orange

apple unveiled the new phone

1-gram

Hello, how are **happy**

1 word

Data

… are **you** …

… are **sad** …

… are**?**…

… are **happy** …

… are **ready** …

… are **happy** …

… are **free** …

3-gram

Hello, how are

3 words

Data

Hello, how are **you**?

Hello, how are **things** going?

Hello, how are **things** today?

Hello, how are **the** kids?

Hello, how are **the** others?

Hello, how are **they** doing?

Hello, how are **things** happening?

10-gram

Hello, today I would like to inform you that ____

10 words

Data

Hello, today I would like to inform you that **the**

Hello, today I would like to inform you that **the**

Transformers

Key Components

Transformers

Key Components

Input Embeddings: Convert text to numerical vectors.

Transformers

Key Components

Input Embeddings: Convert text to numerical vectors.

Positional Encoding: Capture the order of words.

Transformers

Key Components

Input Embeddings: Convert text to numerical vectors.

Positional Encoding: Capture the order of words.

Self-Attention Layers: Highlight relationships between words.

Transformers

Key Components

Input Embeddings: Convert text to numerical vectors.

Positional Encoding: Capture the order of words.

Self-Attention Layers: Highlight relationships between words.

Feedforward Neural Network: Process attention outputs into meaningful patterns.

Transformers

Key Components

Input Embeddings: Convert text to numerical vectors.

Positional Encoding: Capture the order of words.

Self-Attention Layers: Highlight relationships between words.

Feedforward Neural Network: Process attention outputs into meaningful patterns.

Result: Complex understanding of long-range dependencies.

What Can’t LLMs Do?

Lack of True Understanding

LLMs don’t "understand" text the way humans do.
They rely on **statistical patterns** in data rather than reasoning or comprehension.

What Can't LLMs Do?

Dependence on Input Phrasing

LLMs are sensitive to the wording of **prompts**, which can drastically affect their output.

What Can't LLMs Do?

Hallucinations and Fabrication

LLMs can "**hallucinate**" facts,
generating entirely false or misleading information **with confidence**.

What Can't LLMs Do?

Resource Intensive

LLMs require **vast computational resources** for training and inference.

What Can't LLMs Do?

Bias in Outputs

LLMs **inherit biases** present in their **training data**,
which can lead to unfair or **inappropriate outputs**.

What Can't LLMs Do?

Limited Domain-Specific Expertise

While LLMs are **generalists**,
they may **struggle** with **highly specialized** tasks without fine-tuning.

What Can't LLMs Do?

Ethical and Security Concerns

LLMs can be **misused** to **generate harmful content** like misinformation, phishing attacks, or deepfakes.

Ethical and Security Concerns

Ethical and Security Concerns

A summery of “An Overview of Artificial Intelligence Ethics” article

Ethical and Security Concerns

A summery of “An Overview of Artificial Intelligence Ethics” article

1

A categorization of AI ethical issues at individual, societal, and environmental levels.

Ethical and Security Concerns

A summery of “An Overview of Artificial Intelligence Ethics” article

1

A categorization of AI ethical issues at individual, societal, and environmental levels.

2

A global review of 146 AI ethics guidelines, highlighting key principles like transparency, fairness, and accountability.

Ethical and Security Concerns

A summery of “An Overview of Artificial Intelligence Ethics” article

1

A categorization of AI ethical issues at individual, societal, and environmental levels.

2

A global review of 146 AI ethics guidelines, highlighting key principles like transparency, fairness, and accountability.

3

A review of approaches to address ethical issues, including ethical, technological, and legal strategies.

Ethical and Security Concerns

A summery of “An Overview of Artificial Intelligence Ethics” article

1

A categorization of AI ethical issues at individual, societal, and environmental levels.

2

A global review of 146 AI ethics guidelines, highlighting key principles like transparency, fairness, and accountability.

3

A review of approaches to address ethical issues, including ethical, technological, and legal strategies.

4

A discussion on the challenges in implementing ethical AI and future directions for research.

Ethical and Security Concerns

A summery of “An Overview of Artificial Intelligence Ethics” article

https://github.com/Ryan-Heida-Translations/fa-An-Overview-of-Artificial-Intelligence-Ethics


ChatGPT o1-preview
OU
What can I help with?
Message ChatGPT
Have you ever tried GPT o1-preview in voice chat?
ChatGPT can make mistakes. Check important info.
?

What it the differences between different versions of chatGPT

ChatGPT has evolved through several versions, each introducing enhancements in capabilities, performance, and user experience. Here's an overview of the key versions and their differences:

1. GPT-3.5 (ChatGPT Free Version):

- **Release:** 2022
- **Capabilities:** Primarily text-based interactions with a maximum input limit of 4,096 tokens (approximately 3,072 words). TALKATIVE
- **Limitations:** Lacks advanced features like image recognition and voice interaction.

2. GPT-4 (ChatGPT Plus):

- **Release:** March 2023
- **Capabilities:**
  - Enhanced reasoning and understanding, offering more nuanced responses.
  - Supports multimodal inputs, including text and images, allowing it to describe visuals and interpret diagrams.
  - Extended context window up to 8,192 tokens, enabling better retention of conversation context. WIKIPEDIA
- **Access:** Available through a subscription to ChatGPT Plus.

What it the differences between different versions of chatGPT

ChatGPT has evolved through several versions, each introducing enhancements in capabilities, performance, and user experience. Here's an overview of the key versions and their differences:

1. GPT-3.5 (ChatGPT Free Version):

- **Release:** 2022
- **Capabilities:** Primarily text-based interactions with a maximum input limit of 4,096 tokens (approximately 3,072 words). TALKATIVE
- **Limitations:** Lacks advanced features like image recognition and voice interaction.

2. GPT-4 (ChatGPT Plus):

- **Release:** March 2023
- **Capabilities:**
  - Enhanced reasoning and understanding, offering more nuanced responses.
  - Supports multimodal inputs, including text and images, allowing it to describe visuals and interpret diagrams.
  - Extended context window up to 8,192 tokens, enabling better retention of conversation context. WIKIPEDIA
- **Access:** Available through a subscription to ChatGPT Plus.

3. GPT-4o:

- **Release:** May 2024
- **Capabilities:**
  - Processes and generates text, images, and audio, facilitating real-time multimodal interactions.
  - Faster response times and improved performance in non-English languages.
  - Achieved state-of-the-art results in voice, multilingual, and vision benchmarks. WIKIPEDIA
- **Access:** Offered for free to all users, democratizing access to advanced AI capabilities.

4. GPT-4o Mini:

- **Release:** July 2024
- **Capabilities:**
  - A more compact version of GPT-4o, designed for efficiency.
  - Supports text and image inputs, with plans to include audio and video capabilities.
  - Cost-effective, being 60% cheaper than the previous GPT-3.5 Turbo model. THE WALL STREET JOURNAL
- **Access:** Integrated into the ChatGPT interface, replacing GPT-3.5 Turbo.

5. o1 and o1-Mini:

- **Release:** September 2024
- **Capabilities:**

3. **GPT-4o:**

- **Release:** May 2024
- **Capabilities:**
  - Processes and generates text, images, and audio, facilitating real-time multimodal interactions.
  - Faster response times and improved performance in non-English languages.
  - Achieved state-of-the-art results in voice, multilingual, and vision benchmarks. WIKIPEDIA
- **Access:** Offered for free to all users, democratizing access to advanced AI capabilities.

4. **GPT-4o Mini:**

- **Release:** July 2024
- **Capabilities:**
  - A more compact version of GPT-4o, designed for efficiency.
  - Supports text and image inputs, with plans to include audio and video capabilities.
  - Cost-effective, being 60% cheaper than the previous GPT-3.5 Turbo model.
    THE WALL STREET JOURNAL
- **Access:** Integrated into the ChatGPT interface, replacing GPT-3.5 Turbo.

5. **o1 and o1-Mini:**

- **Release:** September 2024
- **Capabilities:**

THE WALL STREET JOURNAL

- **Access:** Integrated into the ChatGPT interface, replacing GPT-3.5 Turbo.

5. o1 and o1-Mini:

- **Release:** September 2024
- **Capabilities:**
  - Designed to tackle complex problems by allocating more time to formulate responses, enhancing accuracy.
  - Excels in scientific reasoning, coding, and mathematical tasks.
  - o1-Mini offers a faster and more cost-effective alternative with similar capabilities. WIKIPEDIA
- **Access:** Available to ChatGPT Plus and Team subscribers.

Choosing the Right Version:

- **Casual Users:** The free version with GPT-4o provides robust capabilities suitable for general use.
- **Professional or Intensive Users:** Subscribing to ChatGPT Plus grants access to advanced features and models like GPT-4 and o1, beneficial for tasks requiring higher accuracy and multimodal interactions.

Each iteration of ChatGPT has expanded its functionalities, offering users a range of options tailored to diverse needs and applications.

Sources

LLMs we can use...

Open-Source

Hugging Face

Llama

APIs

GPT-3

GPT-4

Codex

Claude

Questions

PyTorch Offline Installation

https://gist.github.com/Ryan-PG/9b5198e099bf77f33e94e13a434acf89

How to Store PyTorch for Installation on Environments

If you have PyTorch installed and want to store it for future use in other environments without re-downloading, follow these steps:

Step 1: Download PyTorch Locally

Use `pip` to download the PyTorch package and any additional dependencies to a local folder:

```
pip download torch torchvision torchaudio -d /path/to/your/cache_directory
```

This command downloads the specified PyTorch packages into the `cache_directory`.

Step 2: Install PyTorch from the Local Folder

To install PyTorch from the local folder in a different environment, use the following command:

```
pip install --find-links=/path/to/your/cache_directory torch torchvision torchaudio
```

This tells `pip` to look for the packages in the specified directory before attempting to download them from the internet.

Benefits

- **Offline Installation**: Once stored locally, you can install PyTorch on any environment without internet access.
- **Faster Setup**: Saves time by avoiding repeated downloads.

By following these steps, you can efficiently store and reuse PyTorch packages for different Python environments.

Questions

Differences between python packages

- PyTorch و TensorFlow/Keras بیشتر برای **مدل‌های یادگیری عمیق** مناسب هستند. PyTorch برای تحقیق و توسعه، و TensorFlow برای تولید.
- Scikit-Learn برای **الگوریتم‌های کلاسیک یادگیری ماشین** و کارهای سبک‌تر استفاده می‌شود.
- XGBoost برای **مدل‌های درختی و تحلیل آماری** محبوب است.
- MXNet به‌دلیل **قابلیت توزیع‌پذیری و پشتیبانی از گراف‌های پویا**، برای مدل‌هایی که به اجرا در سطح وسیع نیاز دارند، مناسب است.

Hands on code...

Transformers, StreamLit, Ollama, G4F

RAGs and Crew AI

Large language models are not just tools of innovation.

They are a **reflection** of our collective knowledge and creativity, guiding us toward a future **limited** only by **our imagination**.

مدل‌های زبانی بزرگ فقط ابزار نوآوری نیستند.

آنها **انعکاسی** از دانش و خلاقیت جمعی ما هستند و ما را به سوی آینده ای هدایت می کنند که فقط توسط **تخیل ما محدود** می‌شود.

References & Continue

Hugging Face Tutorial

What Are Transformers

Transformers

What are Transformer Models and how do they work?

An Overview of Artificial Intelligence Ethics - Persian Translation

My Recommendations...

links.ryanheida.com

@RYANHEIDA

Thanks For Watching