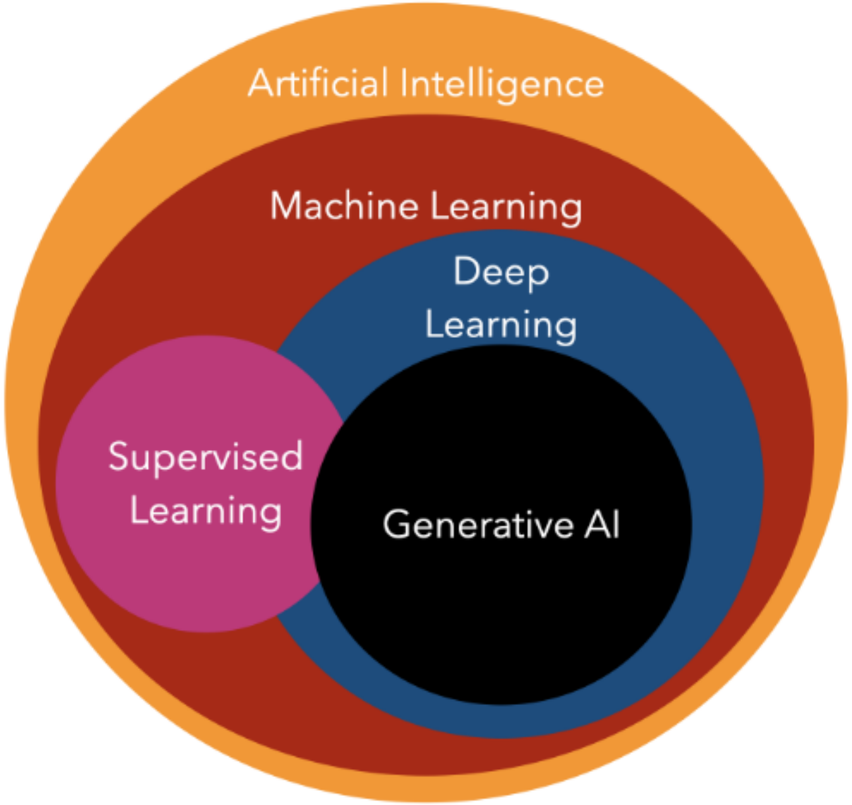
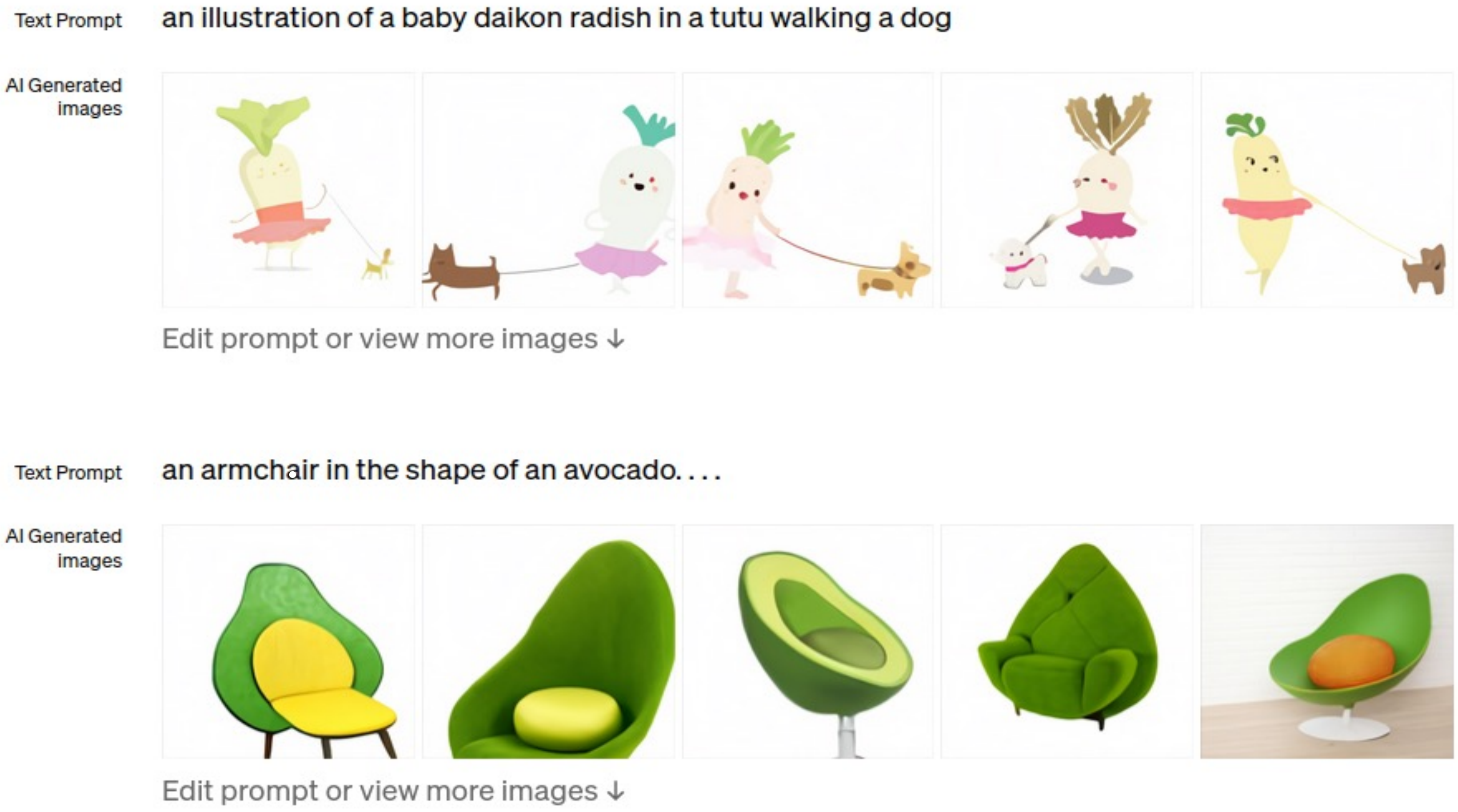
# What is Generative AI?

- Generative AI refers to a deep learning model which enables users to quickly generate new content based on a variety of inputs. Inputs and outputs to these models can include text, images, sounds, animation, 3D models, or other types of data.
- Unlike traditional AI focused on analyzing and interpreting data, generative AI takes inspiration from human creativity and produces unique outputs.
- Deep learning models in general can be divided into discriminative model and generative models.



Sources:- https://www.cmu.edu/intelligentbusiness/expertise/genai-principles.pdf , https://www.nvidia.com/en-us/glossary/generative-ai/

# Examples of Generative AI



Text Prompt an illustration of a baby daikon radish in a tutu walking a dog

AI Generated images

Edit prompt or view more images ↓

Text Prompt an armchair in the shape of an avocado. . . .

AI Generated images

Edit prompt or view more images ↓

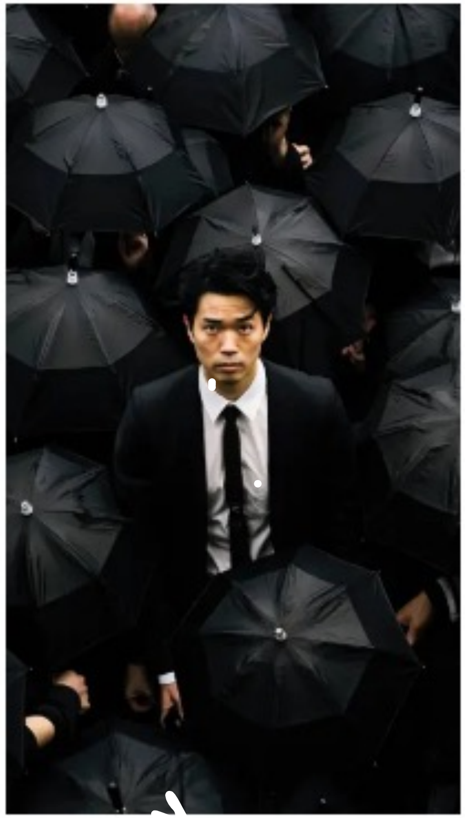
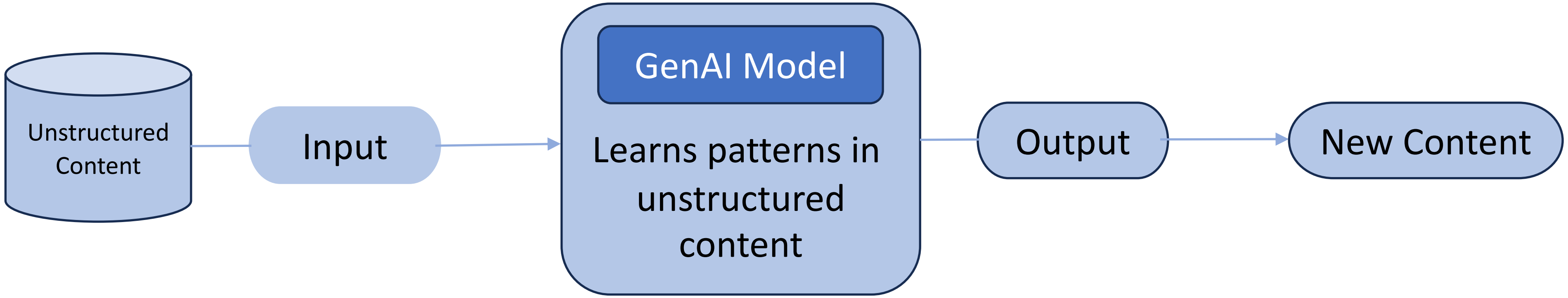


Figure : An image-based GenAI model, Midjourney's response to the prompt — *"Businessman in Tokyo amidst rush hour, his gaze fixed ahead, surrounded by a sea of black umbrellas."*

Source: DALL-E, Midjourney

# How does Generative AI work?

- Generative AI models use neural networks to identify patterns in existing data to generate new content. Trained on unsupervised and semi-supervised learning approaches, organizations can create foundation models from large, unlabeled data sets, essentially forming a base for AI systems to perform tasks.
- Some examples of foundation models include LLMs, GANs,(Generative adversarial network) VAEs,(Variational Autoencoder) and Multimodal, which power tools like ChatGPT, DALL-E, and more. ChatGPT draws data from GPT-3 and enables users to generate a story based on a prompt. Another foundation model Stable Diffusion enables users to generate realistic images based on text input.



Source: https://www.nvidia.com/en-us/glossary/generative-ai/

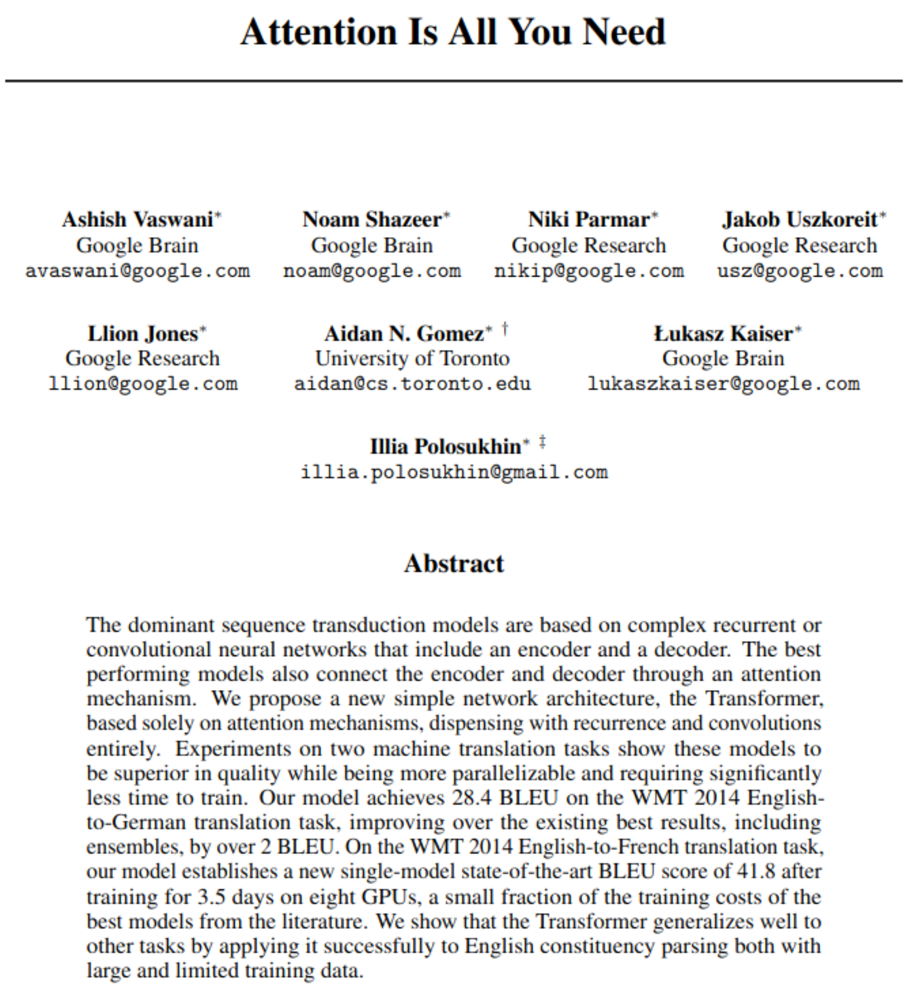
# Evolution of Generative AI

| Year | Developed Technology | Impact |
|---|---|---|
| 1950s-1960s | Rule-based models | Limited capabilities, relied on hand-crafted linguistic rules and features |
| 1980s-1990s | Statistical Language Models | Handle larger datasets, more accurate than rule-based models |
| Mid-2010s | Recurrent Neural Network Language Model (RNNLM), Google Neural Machine Translation (GNMT) system | Model context of words, introduction of deep learning techniques |
| 2017 | Transformer Model | Learn longer-term dependencies in language, parallel training on multiple GPUs |
| 2018 | OpenAI's GPT-1 | Revolutionize NLP tasks, contextually relevant sentences |
| 2020 | OpenAI's GPT-3 | Highly coherent and natural-sounding text, potential for various NLP tasks |
| 2021 | OpenAI's DALL-E | Generative AI model that can create images from text |
| Present | Subsequent models (GPT-4, Llama, Bard, Pangu) | Continuation of advancements in NLP tasks |

In 2017, the transformer model,—a groundbreaking method in the field of natural language processing was proposed. Large language models (LLMs) such as GPT3, RoBERT, Gopher, and BERT started to gain widespread popularity and adoption.

# Break Through

- The breakthrough in LLMs came with the introduction of the Transformer architecture.
- Transformers are a type of deep learning model introduced in the paper “Attention Is All You Need” by Vaswani et al.
- They have since become the foundation for state-of-the-art models in natural language processing tasks. The key innovation in transformers is the “attention mechanism” that allows the model to focus on different parts of the input data differently, much like how humans pay attention to specific parts of a sentence when understanding it.



**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

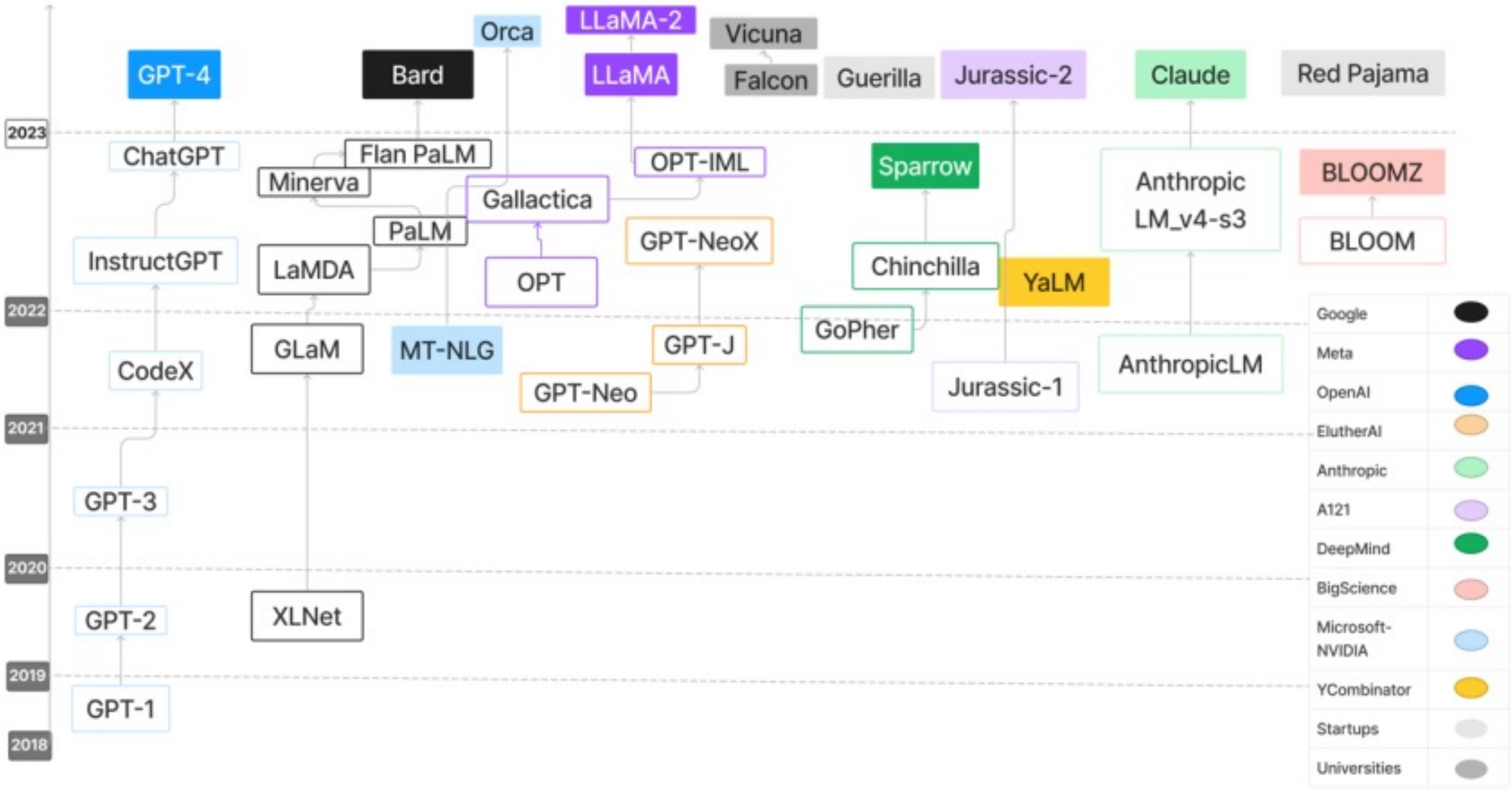
**Łukasz Kaiser**[*]
Google Brain
lukaszkaiser@google.com

**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 41.8 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.

https://arxiv.org/abs/1706.03762



Orca

LLaMA-2

Vicuna

GPT-4

Bard

LLaMA

Falcon

Guerilla

Jurassic-2

Claude

Red Pajama

2023

ChatGPT

Flan PaLM

Minerva

OPT-IML

Sparrow

Anthropic LM_v4-s3

BLOOMZ

Gallactica

PaLM

GPT-NeoX

InstructGPT

LaMDA

OPT

Chinchilla

YaLM

BLOOM

2022

GLaM

MT-NLG

GPT-J

GoPher

CodeX

GPT-Neo

Jurassic-1

AnthropicLM

2021

GPT-3

2020

GPT-2

XLNet

2019

GPT-1

2018

| | |
|---|---|
| Google | ● |
| Meta | ● |
| OpenAI | ● |
| ElutherAI | ● |
| Anthropic | ● |
| A121 | ● |
| DeepMind | ● |
| BigScience | ● |
| Microsoft-NVIDIA | ● |
| YCombinator | ● |
| Startups | ● |
| Universities | ● |

https://www.techrxiv.org/users/618307/articles/682263-large-language-models-a-comprehensive-survey-of-its-applications-challenges-limitations-and-future-prospects

# Why Now?

| Factors Making Generative artificial intelligence possible now | | |
|---|---|---|
| **Large Data sets** | **Computational Power** | **Innovative DL Models** |
| • Availability of large and diverse datasets.<br>• Al models learn patterns, correlations, and characteristics of large datasets.<br>• Pre-trained state-of-the-art models. | • Advancements in hardware; GPUs.<br>• Access to cloud computing.<br>• Open-source software, Hugging Face. | • Generative Adversarial Networks (GANs).<br>• Transformers Architecture.<br>• Reinforcement learning from human feedback (RLHF). |

# Use Cases OF Large Language Models (LLMs)

| Industry | Use Cases |
|---|---|
| Entertainment | • Creative Content Generation<br>• Virtual Avatars<br>• Art and Design |
| Healthcare | • Medical Imaging<br>• Drug Discovery<br>• Personalized Treatment Plans |
| Finance | • Risk Assessment<br>• Portfolio Management<br>• Algorithmic Trading |
| Marketing & Advertising | • Content Creation<br>• Customer Engagement<br>• Targeted Advertising |
| Manufacturing | • Product Design<br>• Process Optimization<br>• Predictive Maintenance |
| Education | • Personalized Learning<br>• Language Learning<br>• Virtual Learning Environments |
| Retail | • Personalized Recommendations<br>• Visual Merchandising<br>• Inventory Management |

Source: https://openknowledge.worldbank.org/server/api/core/bitstreams/e8f93001-27f9-4a9b-b320-4ff3a4102b76/content

# Types of Generative AI Models

- Generative AI or foundation models are designed to generate different types of content, such as text and chat, images, code, video, and embeddings. Researchers can modify these models to fit specific domains and tackle tasks by adjusting the generative AI's learning algorithms or model structures.

| Generative Language Model | Generative Image Models |
|---|---|
| Generative language models learn about patterns in language through training data.<br>Then, Given some text, they predict what comes next. | Generative image models produce new images using techniques like diffusion.<br>Then, given a prompt or related imagery, they transform random noise into images or generate images from prompts. |

# Types of Generative AI Based on data

| Input : Image | | |
|---|---|---|
| **Output: Text** | **Output: Image** | **Output: Video** |
| Image Captioning | Super Resolution | Animation |
| Visual Question Answering | Image Completion | |
| Image Search | | |

# Types of Generative AI Based on data

| Input : Text | | | |
|---|---|---|---|
| **Output: Text** | **Output: Image** | **Output: Audio** | **Output: Decisions** |
| Translation | Image Generation | Text to speech | Play Games |
| Summarization | Video Generation | | |
| Question answering | | | |
| Grammer Correction | | | |

# Transformer based Large Language models

- GPT 3
- GPT 3.5
- GPT 4
- LLaMa-1
- LLaMa-2
- Claude
- Guanaco
- Falcon
- PaLM2 etc.

# Large Language Models

- LLMs are a type of neural network model that are called LLMs because of their size.
- A language model usually consists of hundreds of billions of parameters. Because of the model's size, it can learn about complex relationships between words and phrases in the input text.
- For example, BERT had about 340 million parameters. OpenAI's GPT-2 (introduced in 2019) has 1.5 billion parameters, and GPT-3 (introduced in 2020) has 175 billion.
- The size of these models determines their quality. A model with many parameters allows things to be done that could not be done before.
- These models differ in their training strategies, model architectures, and use cases.
- To provide a clearer understanding of the LLM landscape we categorize them into two types: encoder-decoder or encoder-only language models and decoder-only language models.

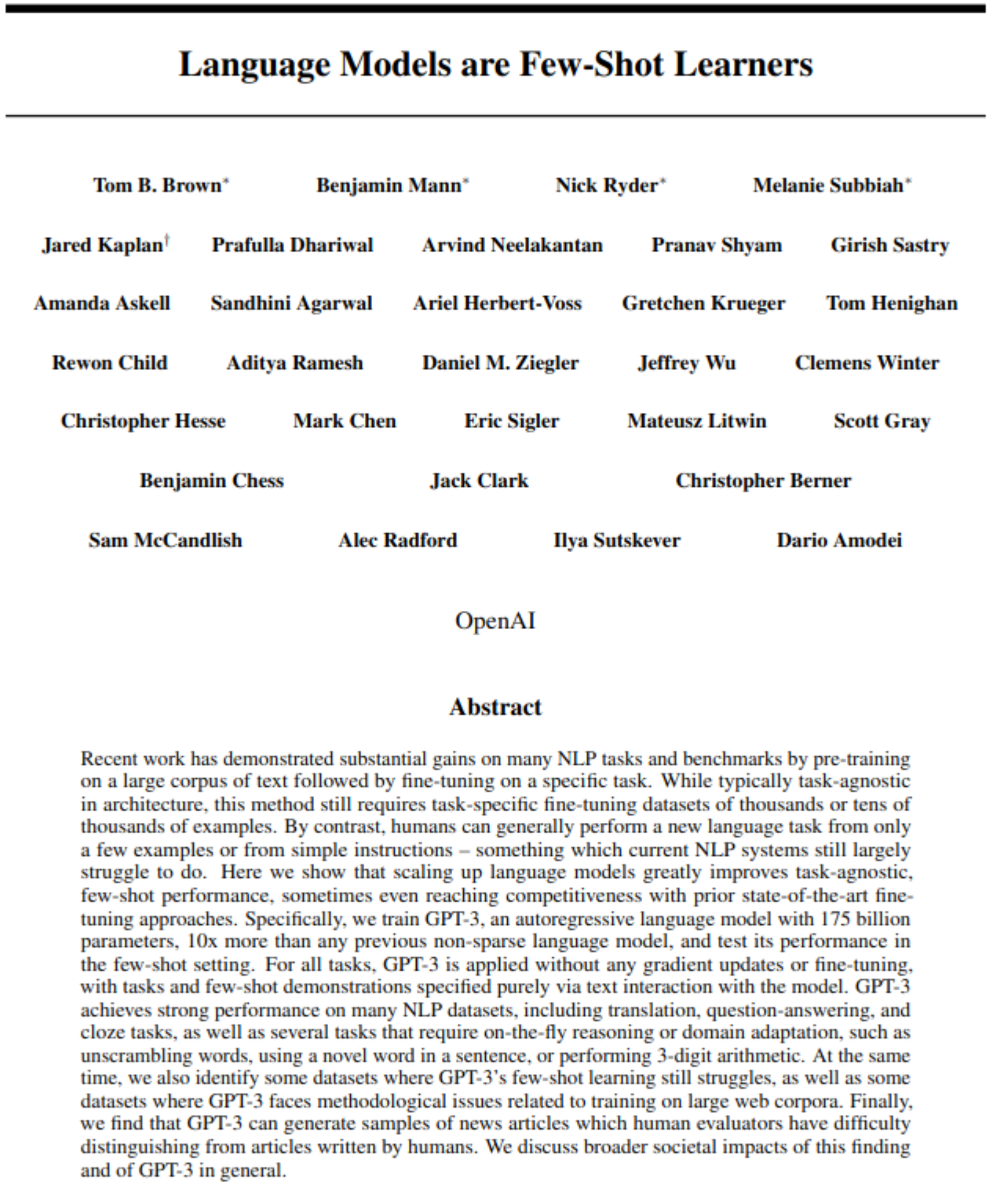
- **Encoder Models:** These models map input sequences to a vector representation. Useful for extracting features (BERT)
- **Decoder Models:** These models generate an output sequence from a fixed length input vector. Useful for generation text, images etc. (GPT-3)
- **Encoder-Decoder Models:** These models are a combination of both encoder and decoder. Encoder is responsible for mapping input into vector and decoder generates output sequence from that vector. (BART/ T5/ FLAN UL2)

| **Categories** | **Characteristics** | **LLM** |
|---|---|---|
| Encoder-Decoder or Encoder-only (BERT-style) | Training: Masked Language Models<br>Model type: Discriminative<br>Pretrain task: Predict masked words | ELMo, BERT, RoBERTa, DistilBERT, BioBERT, XLM, Xlnet, ALBERT, ELECTRA, T5 , GLM, XLM-E, ST-MoE, AlexaTM |
| Decoder-only (GPT-style) | Training Autoregressive Language Models<br>Model type: Generative<br>Pretrain task: Predict next word | GPT-3 , OPT, PaLM, BLOOM, MT-NLG, GLaM, Gopher, chinchilla, LaMDA, GPT-J, LLaMA, GPT-4, BloombergGPT |

# GPT 3

- On June 11, 2020, GPT-3 was launched as a beta version. The full version of GPT-3 has a capacity of 175 billion ML parameters. GPT-2 has 1.5 billion parameters that show how massively powerful GPT-3 is. The original paper introducing GPT-3 was presented by OpenAI engineers.

- GPT-3 can generate various text content depending on the context and predict statements according to available sentences. Due to its context-based nature, it has incredible creative capabilities. The text predictor of GPT-3 processes all the text existing on the internet and can calculate the most statistically expected output. It can write poetry, blogs, PR content, resumes, and technical documentation.



## Language Models are Few-Shot Learners

Tom B. Brown[*] Benjamin Mann[*] Nick Ryder[*] Melanie Subbiah[*]
Jared Kaplan[†] Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

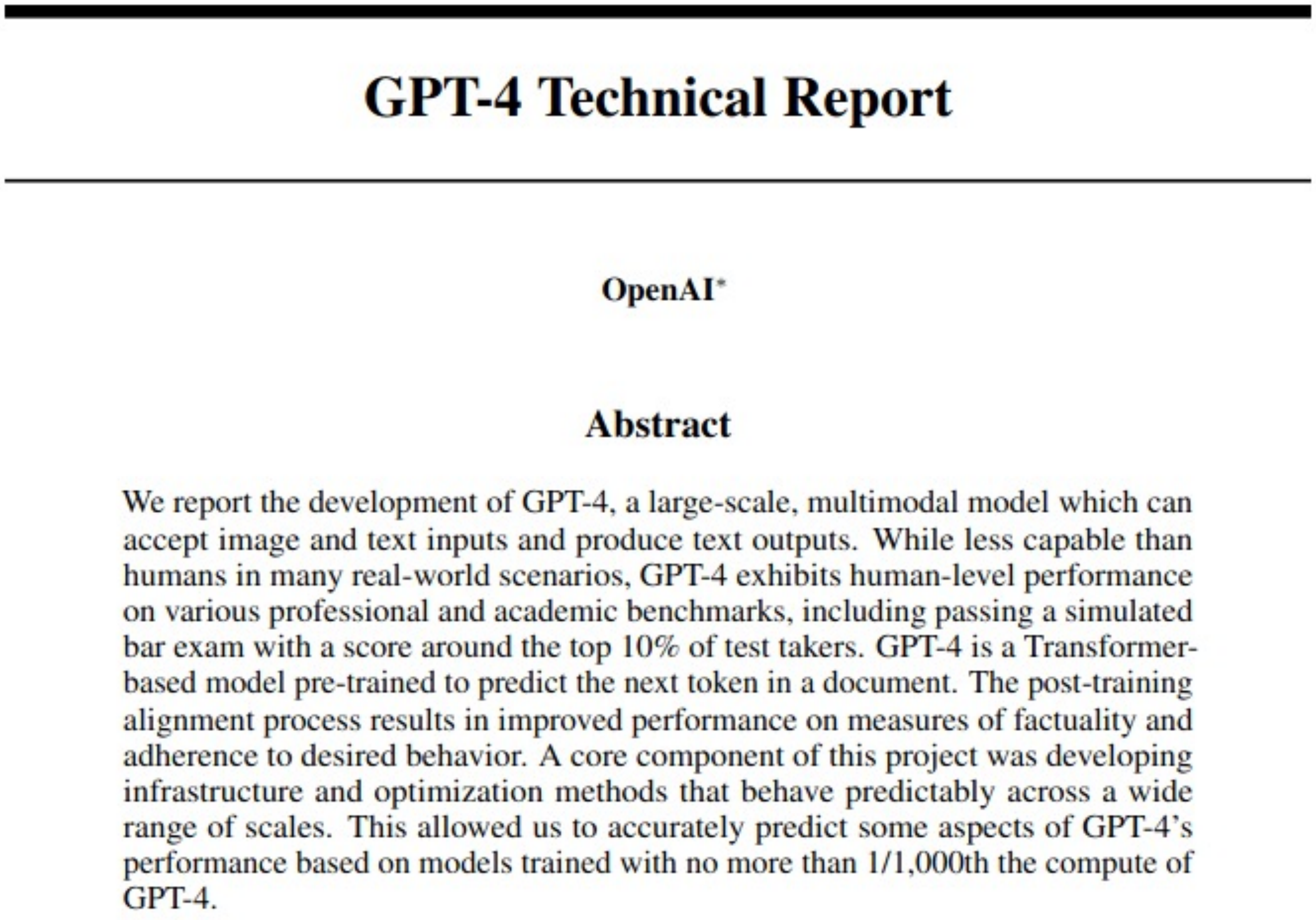
OpenAI

### Abstract

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

https://arxiv.org/pdf/2005.14165.pdf

# GPT 4



## GPT-4 Technical Report

OpenAI*

### Abstract

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4.

https://arxiv.org/pdf/2303.08774v4.pdf

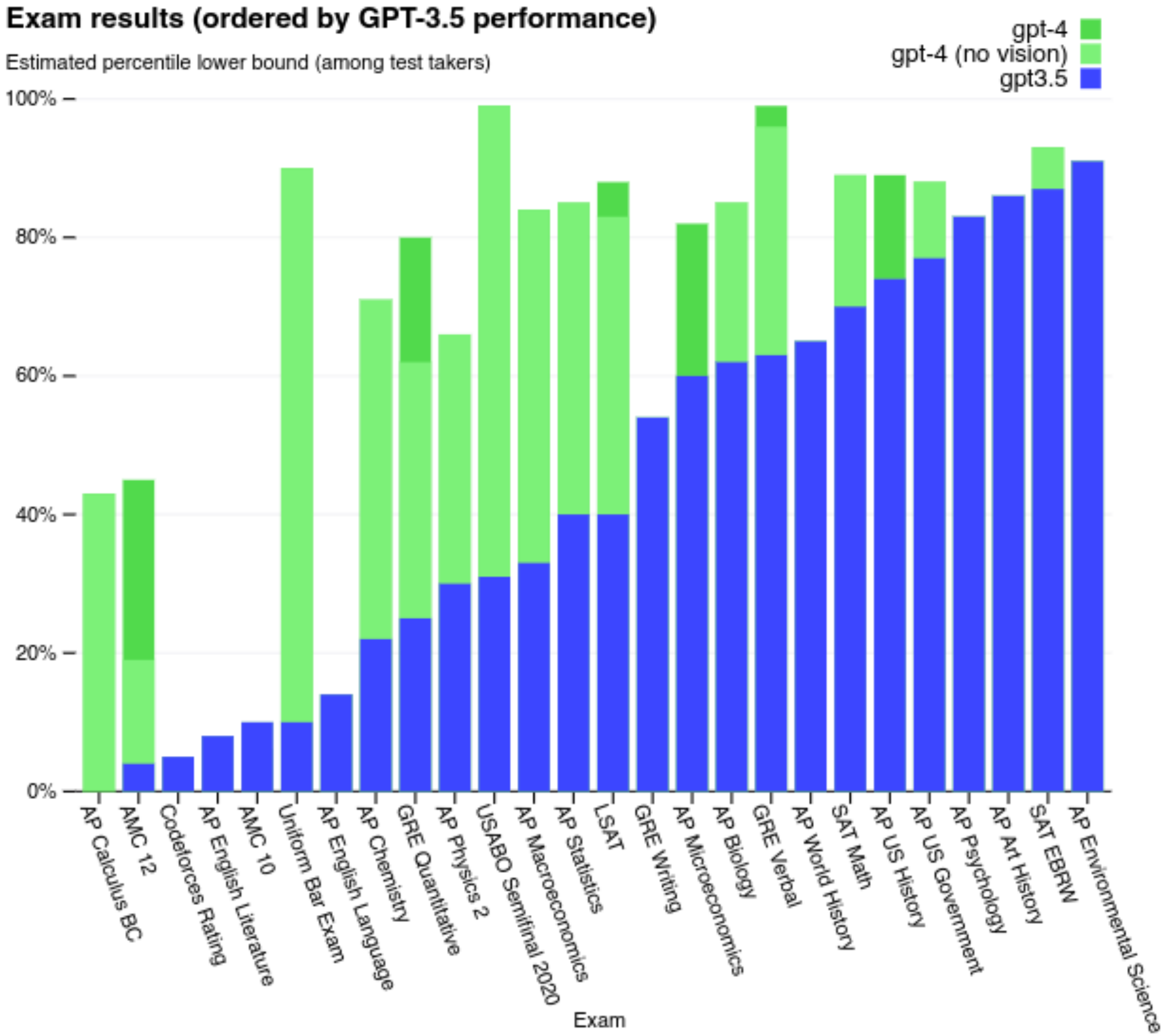


**Figure 4.** GPT performance on academic and professional exams. In each case, we simulate the conditions and scoring of the real exam. Exams are ordered from low to high based on GPT-3.5 performance. GPT-4 outperforms GPT-3.5 on most exams tested. To be conservative we report the lower end of the range of percentiles, but this creates some artifacts on the AP exams which have very wide scoring bins. For example although GPT-4 attains the highest possible score on AP Biology (5/5), this is only shown in the plot as 85th percentile because 15 percent of test-takers achieve that score.

- GPT-4 exhibits human-level performance on the majority of these professional and academic exams. Notably, it passes a simulated version of the Uniform Bar Examination with a score in the top 10% of test takers (Figure 4).

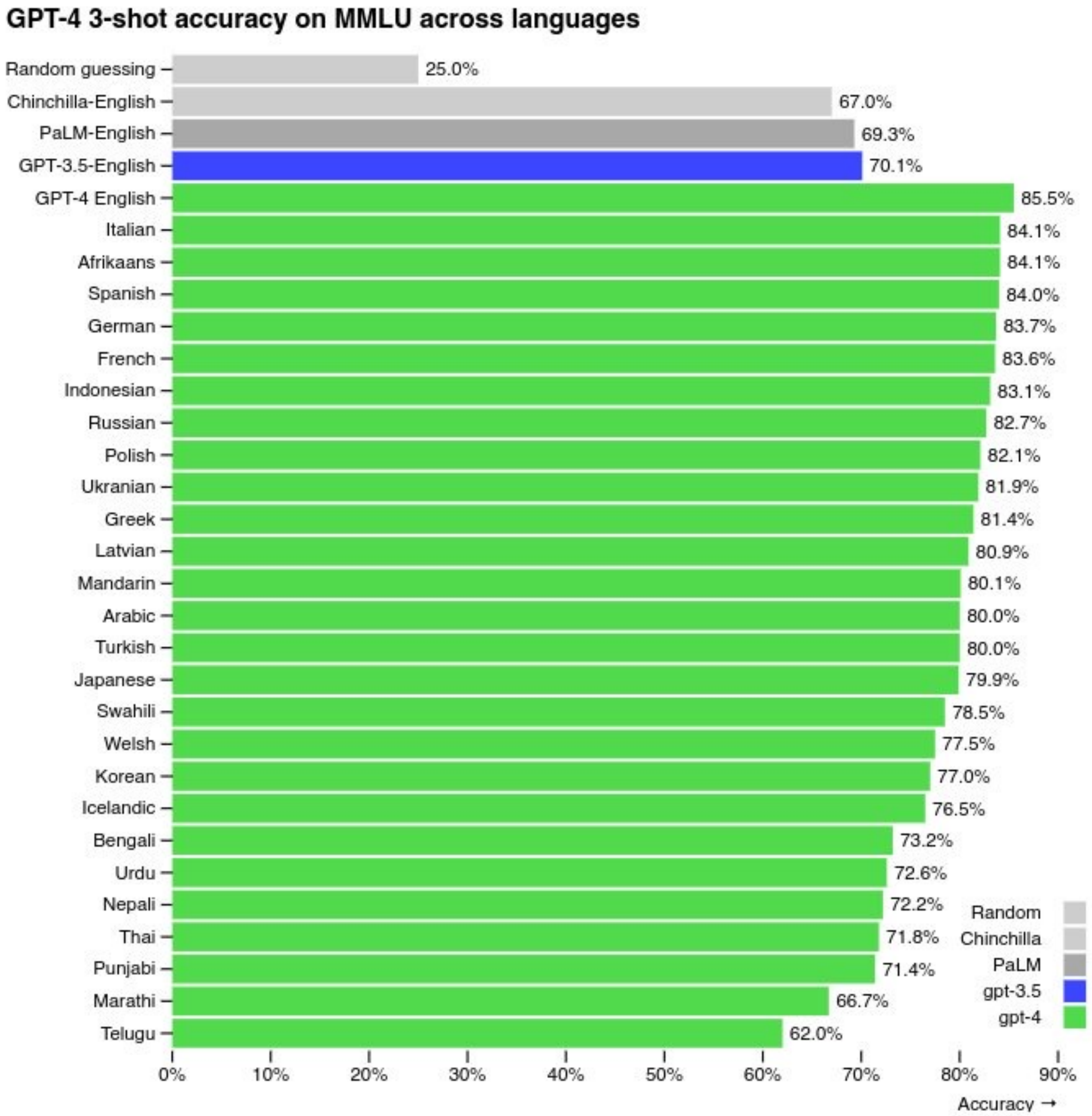


**Figure 5.** Performance of GPT-4 in a variety of languages compared to prior models in English on MMLU. GPT-4 outperforms the English-language performance of existing language models [2, 3] for the vast majority of languages tested, including low-resource languages such as Latvian, Welsh, and Swahili.

- Performance of GPT-4 in a variety of languages compared to prior models in English on MMLU. GPT-4 outperforms the English-language performance of existing language models [2, 3 ] for the vast majority of languages tested, including low-resource languages such as Latvian, Welsh, and Swahili.

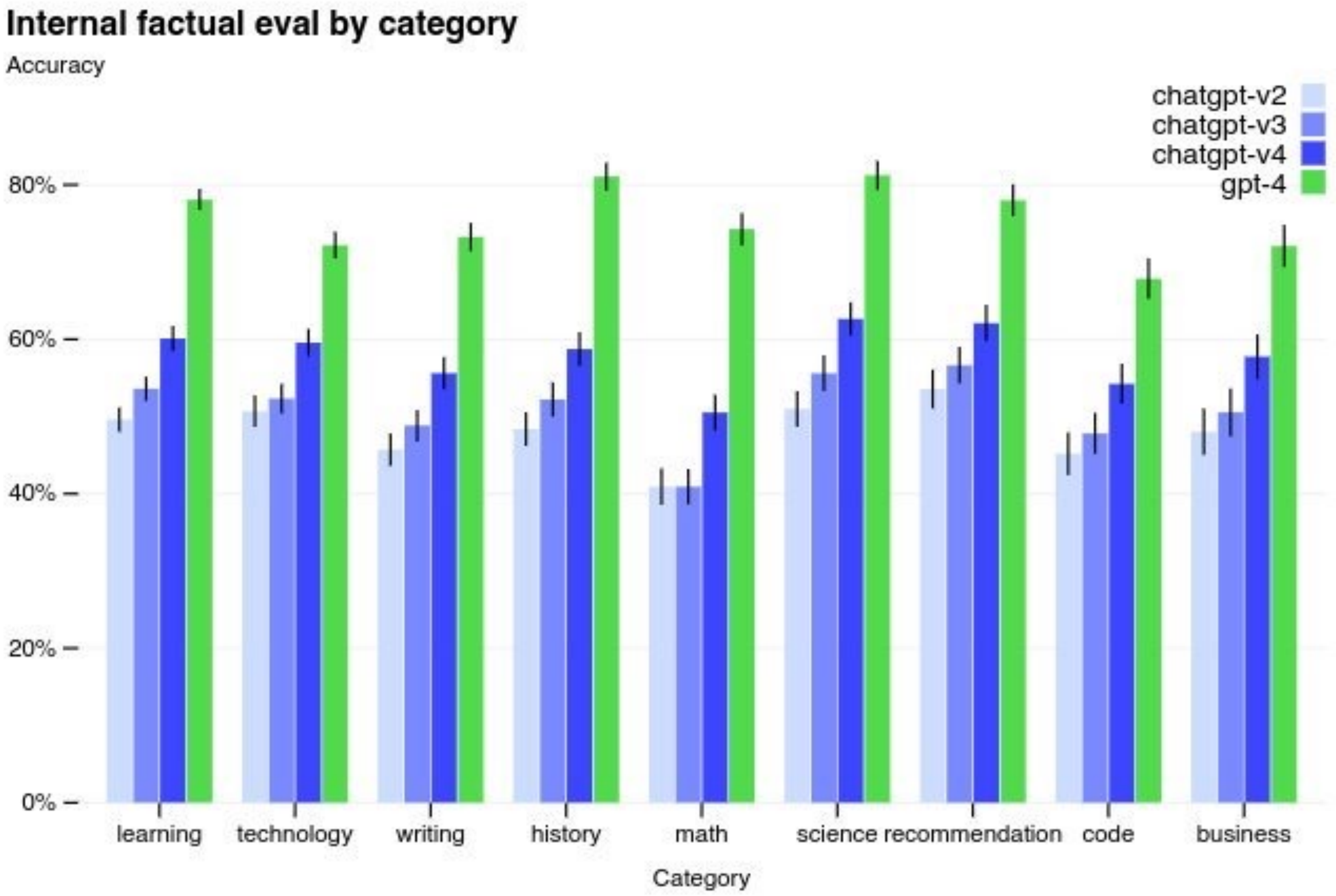


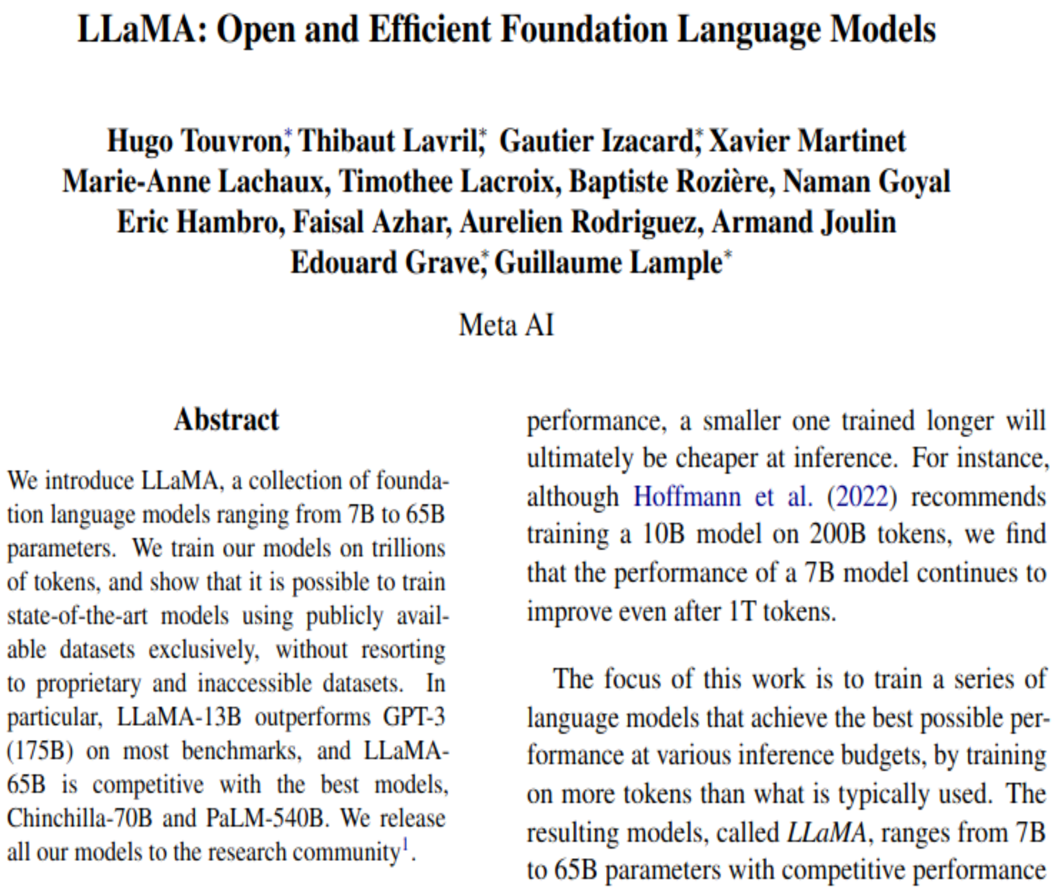
**Figure 6.** Performance of GPT-4 on nine internal adversarially-designed factuality evaluations. Accuracy is shown on the y-axis, higher is better. An accuracy of 1.0 means the model's answers are judged to be in agreement with human ideal responses for all questions in the eval. We compare GPT-4 to three earlier versions of ChatGPT [64] based on GPT-3.5; GPT-4 improves on the latest GPT-3.5 model by 19 percentage points, with significant gains across all topics.

- GPT-4 significantly reduces hallucinations relative to previous GPT-3.5 models (which have them- selves been improving with continued iteration).

- GPT-4 scores 19 percentage points higher than our latest GPT-3.5 on our internal, adversarial designed factuality evaluations.

# LLaMA (Feb 2023)

- LlaMA (Large Language Model Meta AI) is a Generative AI model, specifically a group of foundational Large Language Models developed by Meta AI, a company owned by Meta(Formerly Facebook).
- LLaMA, an auto-regressive language model, is built on the transformer architecture. LLaMA functions by taking a sequence of words as input and predicting the next word, recursively generating text.
- The LLaMA models are available in several sizes: 7B, 13B, 33B, and 65B parameters, and you can access them on Hugging Face (LLaMA models converted to work with Transformers) or on the official repository facebookresearch/llama.



## LLaMA: Open and Efficient Foundation Language Models

Hugo Touvron*, Thibaut Lavril*, Gautier Izacard*, Xavier Martinet
Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal
Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin
Edouard Grave*, Guillaume Lample*

Meta AI

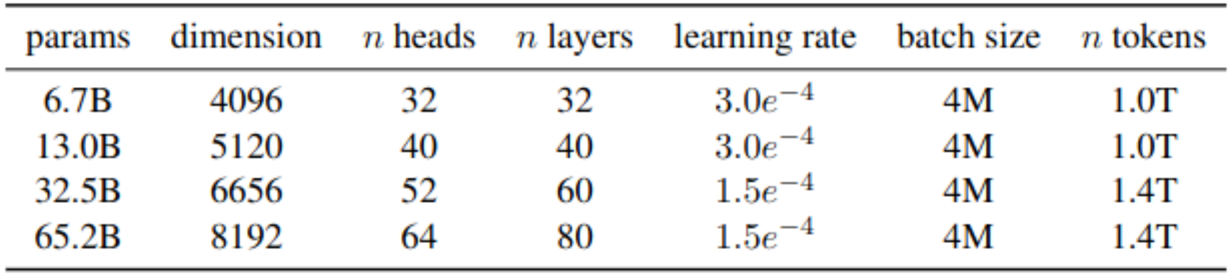
### Abstract

We introduce LLaMA, a collection of foundation language models ranging from 7B to 65B parameters. We train our models on trillions of tokens, and show that it is possible to train state-of-the-art models using publicly available datasets exclusively, without resorting to proprietary and inaccessible datasets. In particular, LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla-70B and PaLM-540B. We release all our models to the research community[1].

performance, a smaller one trained longer will ultimately be cheaper at inference. For instance, although Hoffmann et al. (2022) recommends training a 10B model on 200B tokens, we find that the performance of a 7B model continues to improve even after 1T tokens.

The focus of this work is to train a series of language models that achieve the best possible performance at various inference budgets, by training on more tokens than what is typically used. The resulting models, called *LLaMA*, ranges from 7B to 65B parameters with competitive performance

Source :- (https://arxiv.org/pdf/2302.13971.pdf)



| params | dimension | $n$ heads | $n$ layers | learning rate | batch size | $n$ tokens |
|---|---|---|---|---|---|---|
| 6.7B | 4096 | 32 | 32 | $3.0e^{-4}$ | 4M | 1.0T |
| 13.0B | 5120 | 40 | 40 | $3.0e^{-4}$ | 4M | 1.0T |
| 32.5B | 6656 | 52 | 60 | $1.5e^{-4}$ | 4M | 1.4T |
| 65.2B | 8192 | 64 | 80 | $1.5e^{-4}$ | 4M | 1.4T |

Table 2: **Model sizes, architectures, and optimization hyper-parameters.**

The details of the hyper-parameters for our different models are given in this Table.

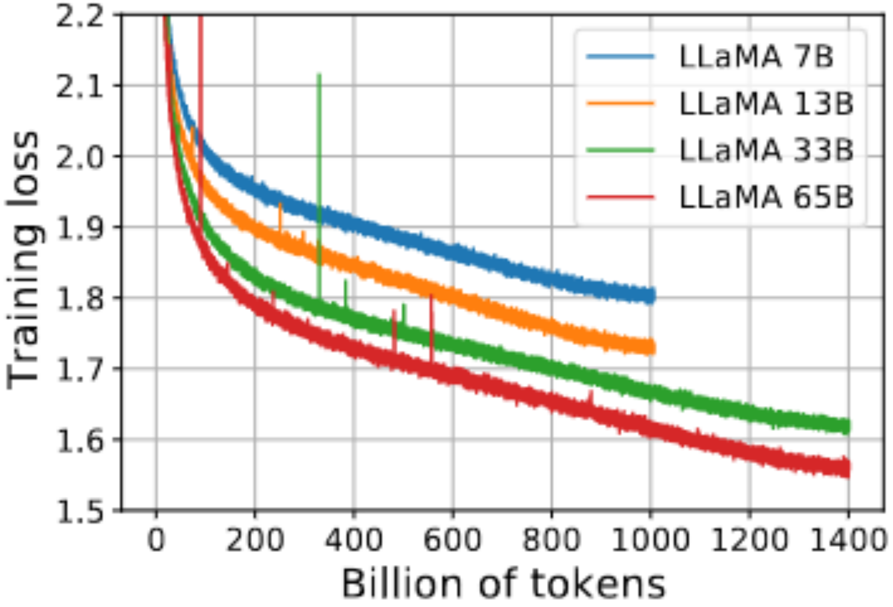
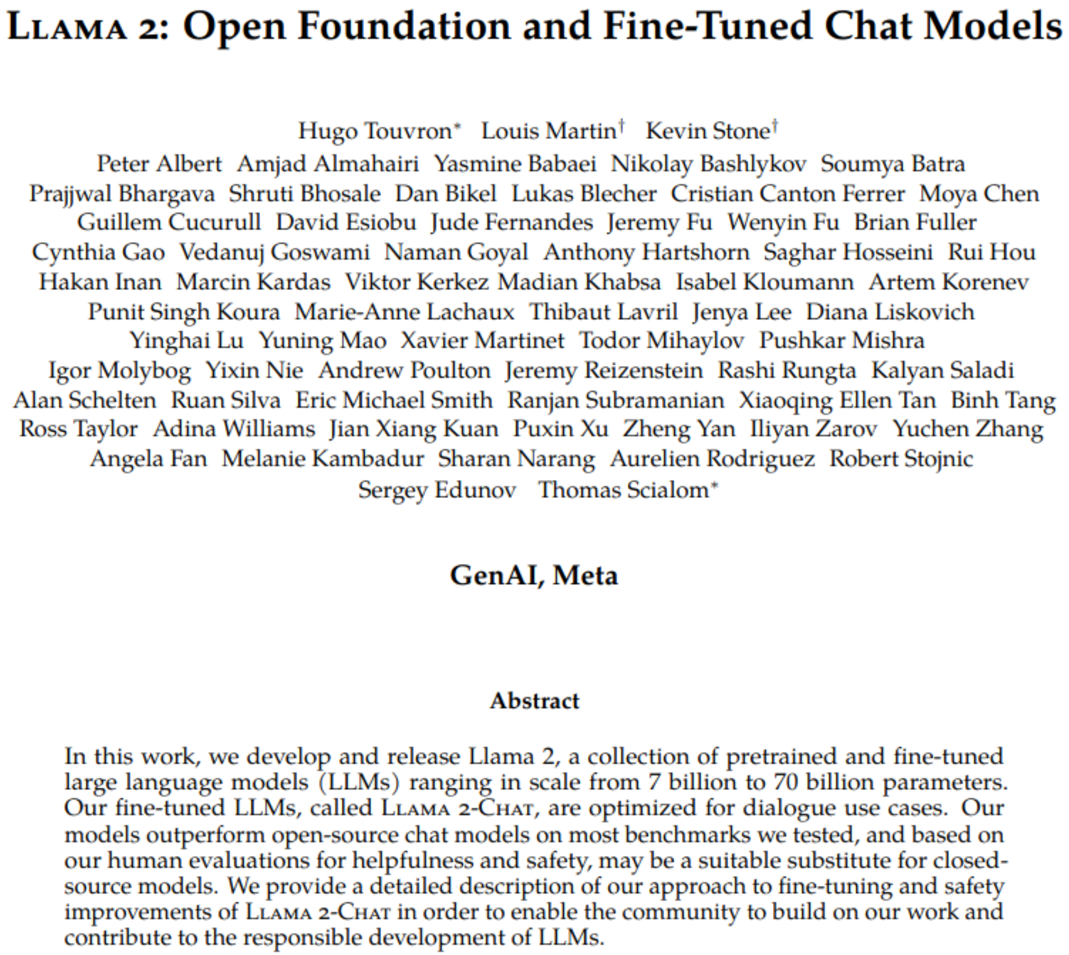


Figure shows Training loss over train tokens for the 7B, 13B, 33B, and 65 models. LLaMA-33B and LLaMA-65B were trained on 1.4T tokens. The smaller models were trained on 1.0T (Trillion) tokens. All models are trained with a batch size of 4M tokens.

Source :- (https://arxiv.org/pdf/2302.13971.pdf)

# LLaMA 2 (July 2023)

- LlaMA 2 surpasses the previous version, LlaMA version 1, which Meta released in July of 2023. It came out in three sizes: 7B, 13B, and 70B parameter models.
- Upon its release, LlaMA 2 achieved the highest score on Hugging Face. Even across all segments (7B, 13B, and 70B), the top-performing model on Hugging Face originates from LlaMA 2, having been fine-tuned or retrained.
- Llama 2-Chat, a fine-tuned version of Llama 2 that is optimized for dialogue use cases. They also release variants of this model with 7B, 13B, and 70B parameters as well.



## Llama 2: Open Foundation and Fine-Tuned Chat Models

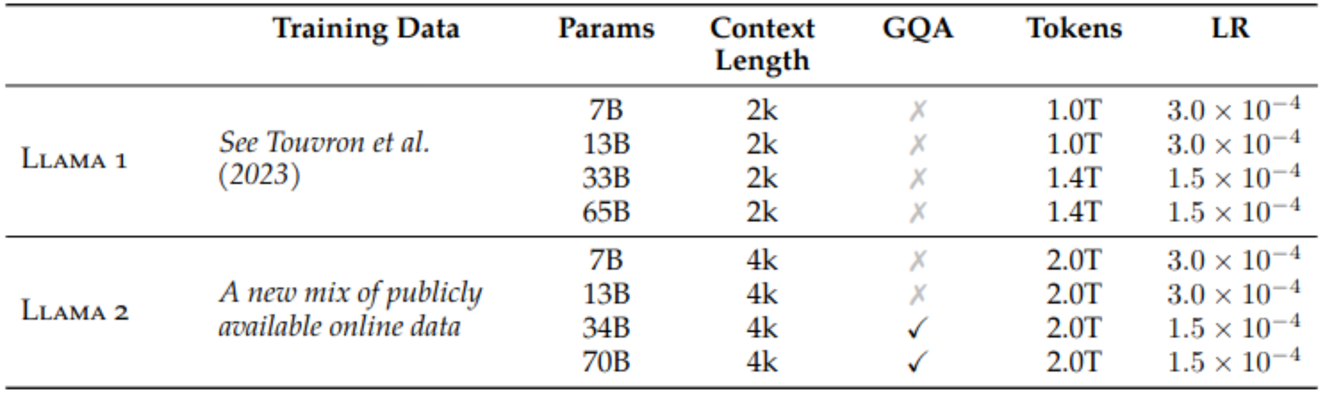
Hugo Touvron* Louis Martin† Kevin Stone†
Peter Albert Amjad Almahairi Yasmine Babaei Nikolay Bashlykov Soumya Batra
Prajjwal Bhargava Shruti Bhosale Dan Bikel Lukas Blecher Cristian Canton Ferrer Moya Chen
Guillem Cucurull David Esiobu Jude Fernandes Jeremy Fu Wenyin Fu Brian Fuller
Cynthia Gao Vedanuj Goswami Naman Goyal Anthony Hartshorn Saghar Hosseini Rui Hou
Hakan Inan Marcin Kardas Viktor Kerkez Madian Khabsa Isabel Kloumann Artem Korenev
Punit Singh Koura Marie-Anne Lachaux Thibaut Lavril Jenya Lee Diana Liskovich
Yinghai Lu Yuning Mao Xavier Martinet Todor Mihaylov Pushkar Mishra
Igor Molybog Yixin Nie Andrew Poulton Jeremy Reizenstein Rashi Rungta Kalyan Saladi
Alan Schelten Ruan Silva Eric Michael Smith Ranjan Subramanian Xiaoqing Ellen Tan Binh Tang
Ross Taylor Adina Williams Jian Xiang Kuan Puxin Xu Zheng Yan Iliyan Zarov Yuchen Zhang
Angela Fan Melanie Kambadur Sharan Narang Aurelien Rodriguez Robert Stojnic
Sergey Edunov Thomas Scialom*

**GenAI, Meta**

**Abstract**

In this work, we develop and release Llama 2, a collection of pretrained and fine-tuned large language models (LLMs) ranging in scale from 7 billion to 70 billion parameters. Our fine-tuned LLMs, called Llama 2-Chat, are optimized for dialogue use cases. Our models outperform open-source chat models on most benchmarks we tested, and based on our human evaluations for helpfulness and safety, may be a suitable substitute for closed-source models. We provide a detailed description of our approach to fine-tuning and safety improvements of Llama 2-Chat in order to enable the community to build on our work and contribute to the responsible development of LLMs.

Source: https://arxiv.org/pdf/2307.09288.pdf



| | Training Data | Params | Context Length | GQA | Tokens | LR |
|---|---|---|---|---|---|---|
| Llama 1 | *See Touvron et al. (2023)* | 7B | 2k | ✗ | 1.0T | $3.0 \times 10^{-4}$ |
| | | 13B | 2k | ✗ | 1.0T | $3.0 \times 10^{-4}$ |
| | | 33B | 2k | ✗ | 1.4T | $1.5 \times 10^{-4}$ |
| | | 65B | 2k | ✗ | 1.4T | $1.5 \times 10^{-4}$ |
| Llama 2 | *A new mix of publicly available online data* | 7B | 4k | ✗ | 2.0T | $3.0 \times 10^{-4}$ |
| | | 13B | 4k | ✗ | 2.0T | $3.0 \times 10^{-4}$ |
| | | 34B | 4k | ✓ | 2.0T | $1.5 \times 10^{-4}$ |
| | | 70B | 4k | ✓ | 2.0T | $1.5 \times 10^{-4}$ |

- Compares the attributes of the new Llama 2 models with the Llama 1 models.

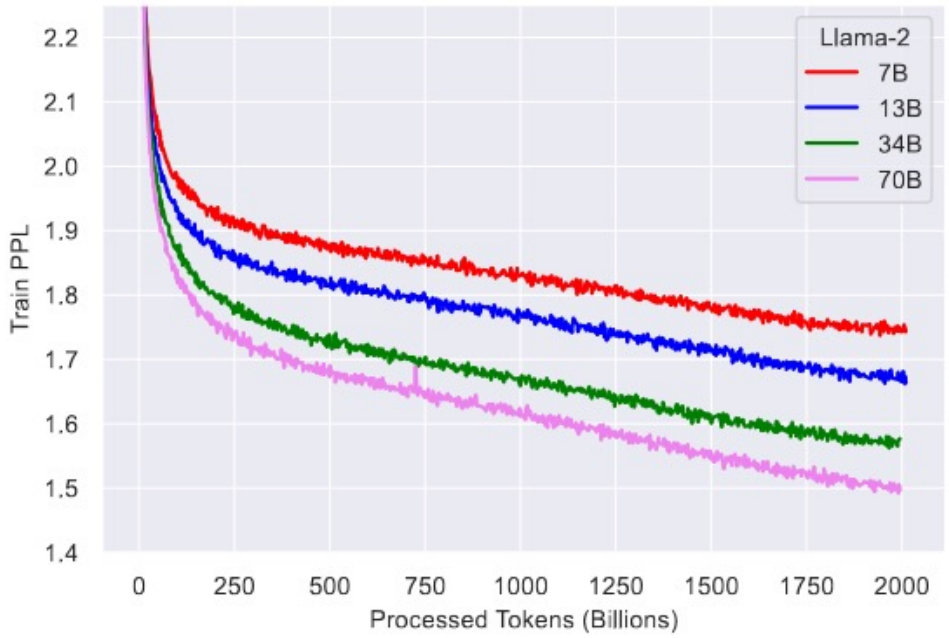
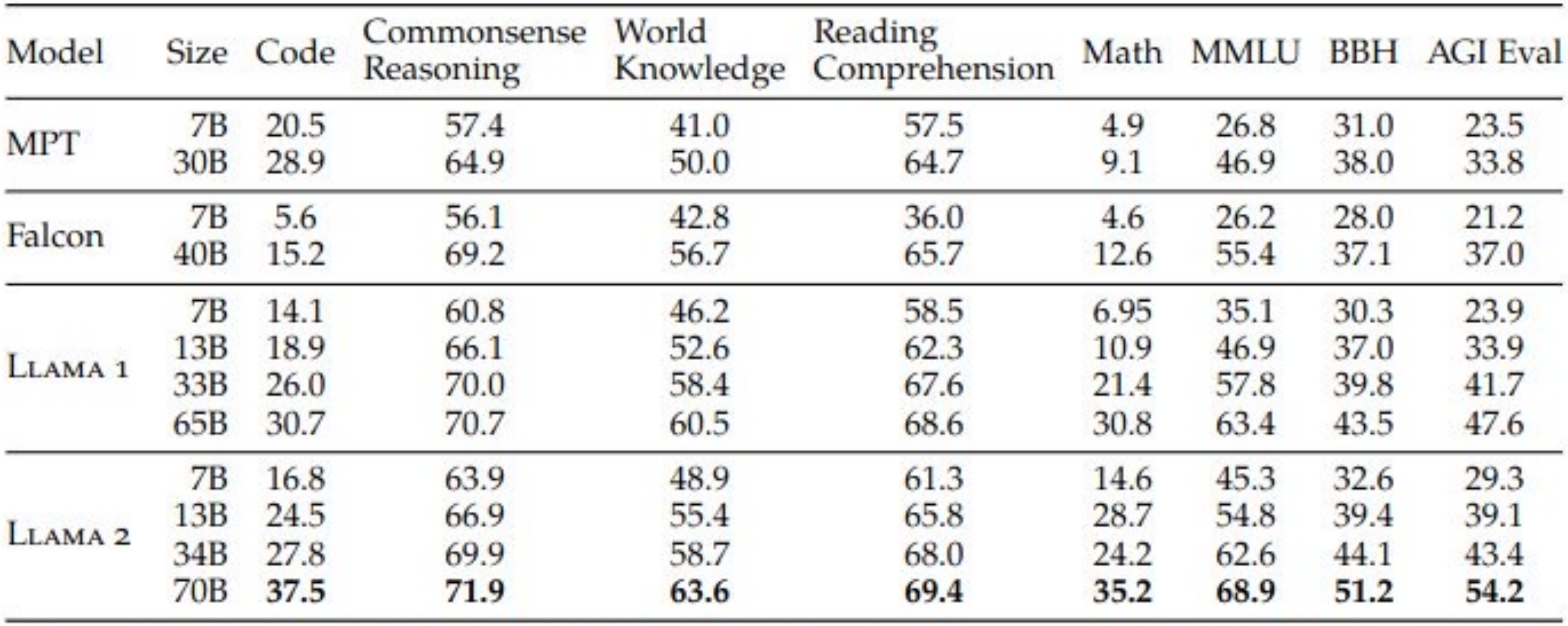


**Figure 5: Training Loss for Llama 2 models.** We compare the training loss of the Llama 2 family of models. We observe that after pretraining on 2T Tokens, the models still did not show any sign of saturation.

- Figure shows the training loss for Llama 2.

Source: https://arxiv.org/pdf/2307.09288.pdf



| Model | Size | Code | Commonsense Reasoning | World Knowledge | Reading Comprehension | Math | MMLU | BBH | AGI Eval |
|---|---|---|---|---|---|---|---|---|---|
| MPT | 7B | 20.5 | 57.4 | 41.0 | 57.5 | 4.9 | 26.8 | 31.0 | 23.5 |
| | 30B | 28.9 | 64.9 | 50.0 | 64.7 | 9.1 | 46.9 | 38.0 | 33.8 |
| Falcon | 7B | 5.6 | 56.1 | 42.8 | 36.0 | 4.6 | 26.2 | 28.0 | 21.2 |
| | 40B | 15.2 | 69.2 | 56.7 | 65.7 | 12.6 | 55.4 | 37.1 | 37.0 |
| Llama 1 | 7B | 14.1 | 60.8 | 46.2 | 58.5 | 6.95 | 35.1 | 30.3 | 23.9 |
| | 13B | 18.9 | 66.1 | 52.6 | 62.3 | 10.9 | 46.9 | 37.0 | 33.9 |
| | 33B | 26.0 | 70.0 | 58.4 | 67.6 | 21.4 | 57.8 | 39.8 | 41.7 |
| | 65B | 30.7 | 70.7 | 60.5 | 68.6 | 30.8 | 63.4 | 43.5 | 47.6 |
| Llama 2 | 7B | 16.8 | 63.9 | 48.9 | 61.3 | 14.6 | 45.3 | 32.6 | 29.3 |
| | 13B | 24.5 | 66.9 | 55.4 | 65.8 | 28.7 | 54.8 | 39.4 | 39.1 |
| | 34B | 27.8 | 69.9 | 58.7 | 68.0 | 24.2 | 62.6 | 44.1 | 43.4 |
| | 70B | **37.5** | **71.9** | **63.6** | **69.4** | **35.2** | **68.9** | **51.2** | **54.2** |

**Table 3: Overall performance on grouped academic benchmarks compared to open-source base models.**

- The results for the Llama 1 and Llama 2 base models, MosaicML Pretrained Transformer (MPT)†† models, and Falcon (Almazrouei et al., 2023)* models on standard academic benchmarks compared. Llama 2 models outperform Llama 1 model, MPT, Falcon.

*https://arxiv.org/pdf/2311.16867.pdf
https://arxiv.org/pdf/2307.09288.pdf

# How Does LLaMA Differ From Other AI Models?

- The paper [1] presents a comprehensive evaluation of LLaMA models, comparing them with other state-of-the-art language models such as GPT-3, GPT-NeoX, Gopher, Chinchilla, and PaLM. The benchmark tests include common sense reasoning, trivia, reading comprehension, question answering, mathematical reasoning, code generation, and general domain knowledge.
- **Common sense reasoning.** The LLaMA-65B model has outperformed SOTA model architectures in PIQA, SIQA, and OpenBookQA reasoning benchmarks. Even smaller model 33B has outperformed all of them in ARC, easy and challenging.
- **Closed-Book Question Answering & Trivia.** The test measures LLM's ability to interpret and respond to realistic, human questions. LLaMA model has consistently outperformed GPT3, Gopher, Chinchilla, and PaLM in Natural Questions and TriviaQA benchmarks.
- **Reading comprehension.** It uses RACE-middle and RACE-high benchmark tests. LLaMA models have outperformed GPT-3 and have similar performance to PaLM 540B.
- **Mathematical Reasoning.** LLaMA was not fine-tuned on any mathematical data, and it performed quite poorly compared to Minerva.
- **Code Generation**. It uses HumanEval and MBPP test benchmarks. LLaMA has outperformed both LAMDA and PaLM in HumanEval@100, MBP@1, and MBP@80.
- **Domain knowledge.** LLaMA models have **performed worse** compared to the massive PaLM 540B parameter model. PaLM has wide domain knowledge due to a larger number of parameters.

[1] https://arxiv.org/pdf/2302.13971.pdf

# Comparision of GPT3.5, Bard, GPT-4, Llama-2

| **Feature** | **ChatGPT (GPT 3.5)** | **Bard** | **Bing Chat (GPT-4)** | **Llama-2** |
|---|---|---|---|---|
| Accuracy | Not as accurate as Bard | Generally more accurate than ChatGPT | Most accurate | Least accurate |
| Versatile | Generally more versatile than Bard | Can generate text, translate languages, and write different kinds of creative content | Not as versatile as ChatGPT or Bard | Less than ChatGPT and Bard both better than bing |
| Company | OpenAI | Google | Microsoft | Meta |
| Primary Purpose | Creative text generation | Conversational AI | Information retrieval | Text generation, answer questions, language translation, etc. |
| Integration | Standalone model | Standalone model | Integrated with bing search engine | Standalone model |
| Ease to use | User-Friendly | User-Friendly | Not as User-Friendly as ChatGPT or Bard | User-Friendly |

Continued.........

| **Feature** | **ChatGPT (GPT 3.5)** | **Bard** | **Bing Chat (GPT-4)** | **Llama-2** |
|---|---|---|---|---|
| Access to online data | No, trained on data available till 2021 | Yes | Yes | Yes |
| Cost | GPT 3.5 free / GPT-4 (20 USD per month) | Free | Free | Free |
| Availability | Publicly available | Publicly available | Publicly available | Publicly available |
| Architecture | Generative pre-trained transformer | Pathways Language models (PaLM2) | Next Generation GPT | Transformer |
| Limitations | May generate less coherent or incorrect text | Not as creative as ChatGPT | May provide limited or incomplete information | Trained on a smaller dataset than ChatGPT and Bard, may not generate text for some topics |

# Benefits of Generative AI

1. Increased efficiency and productivity
2. Faster Results
3. Cost Saving
4. Increased Creativity
5. Enhance decision making
6. Accelerate research and analysis
7. Provide assistance in exploring new Field
8. Personal assistant

# Limitations of generative AI

The output of generative AI can be

- Biased
- Lake of consistency (Hallucination)
- Inaccurate
- Output may not reflect the real world
- Create realistic looking fake images, audio and videos (deepfakes).

# Hallucinations

- Hallucinations are words or phrases that generated by the model that are often nonsensical or grammatically incorrect.

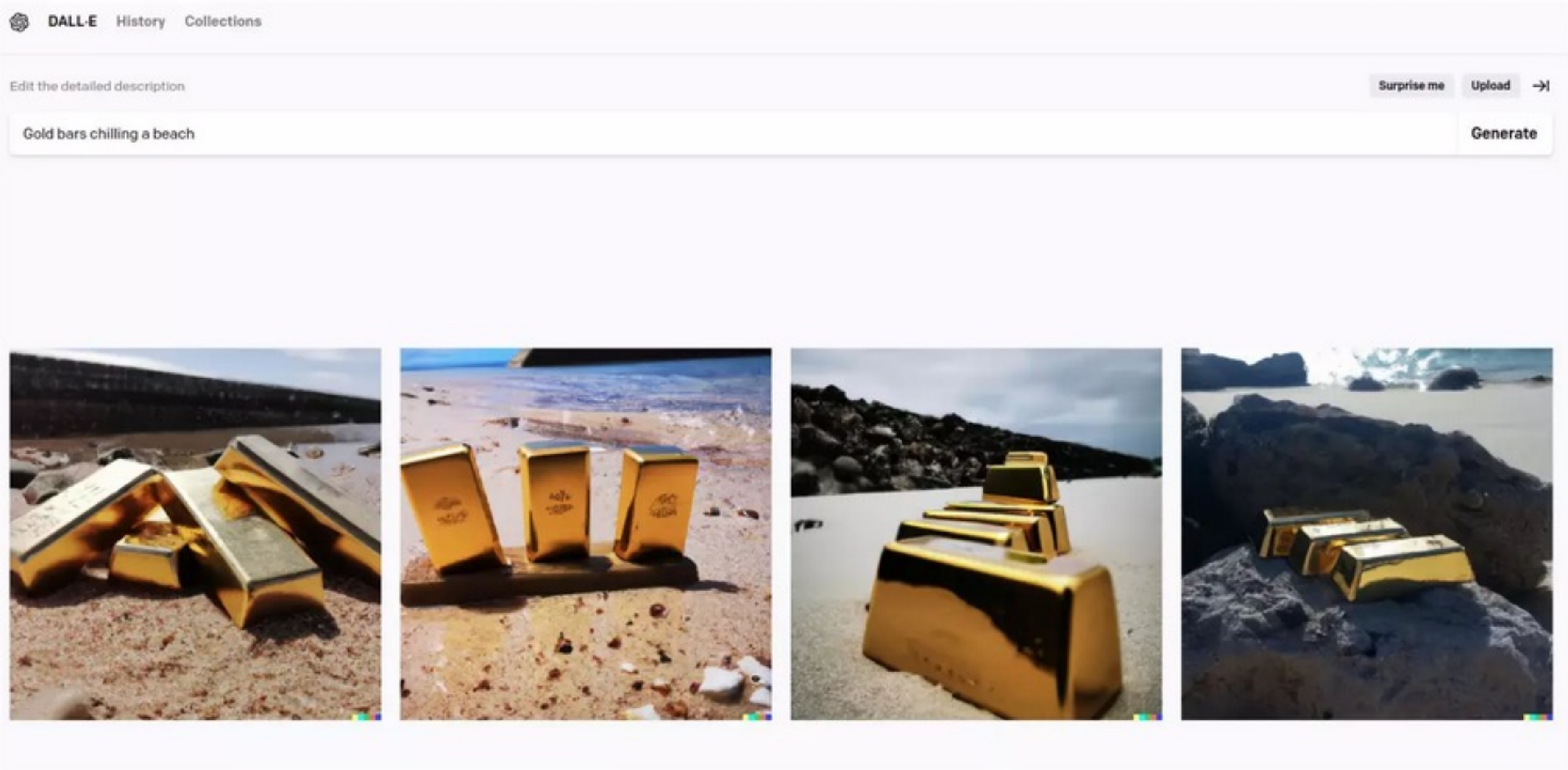
**Challenges**

- The model not trained on enough data
- Model is trained on noisy or dirty data
- The model is not given enough context
- The model is not given enough constraints

# Applications of Generative AI

# Applications

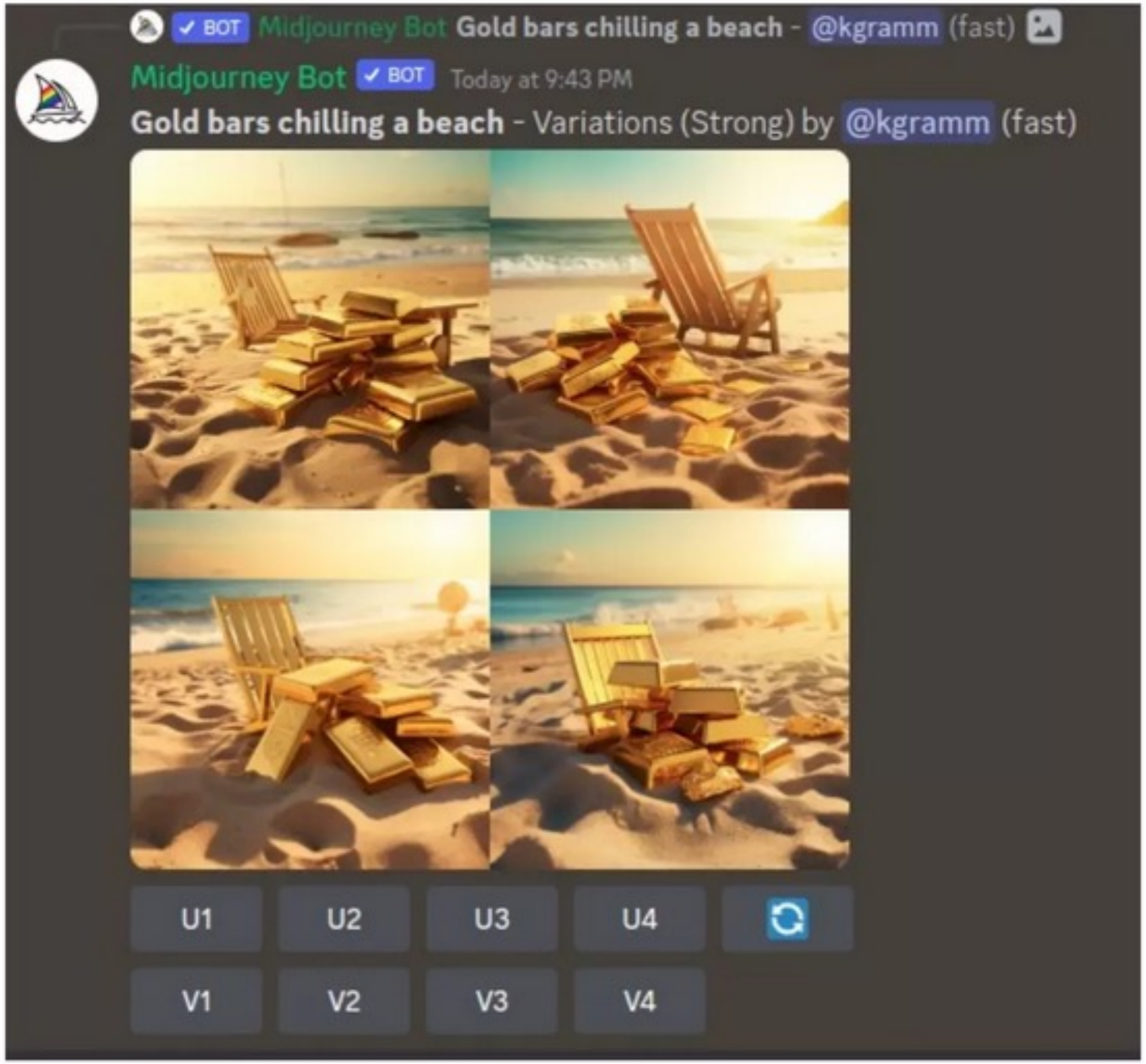
## Dall-E Example 2



Gold bars chilling a beach

# Applications

## Midjourney



✓ BOT Midjourney Bot **Gold bars chilling a beach** - @kgramm (fast)

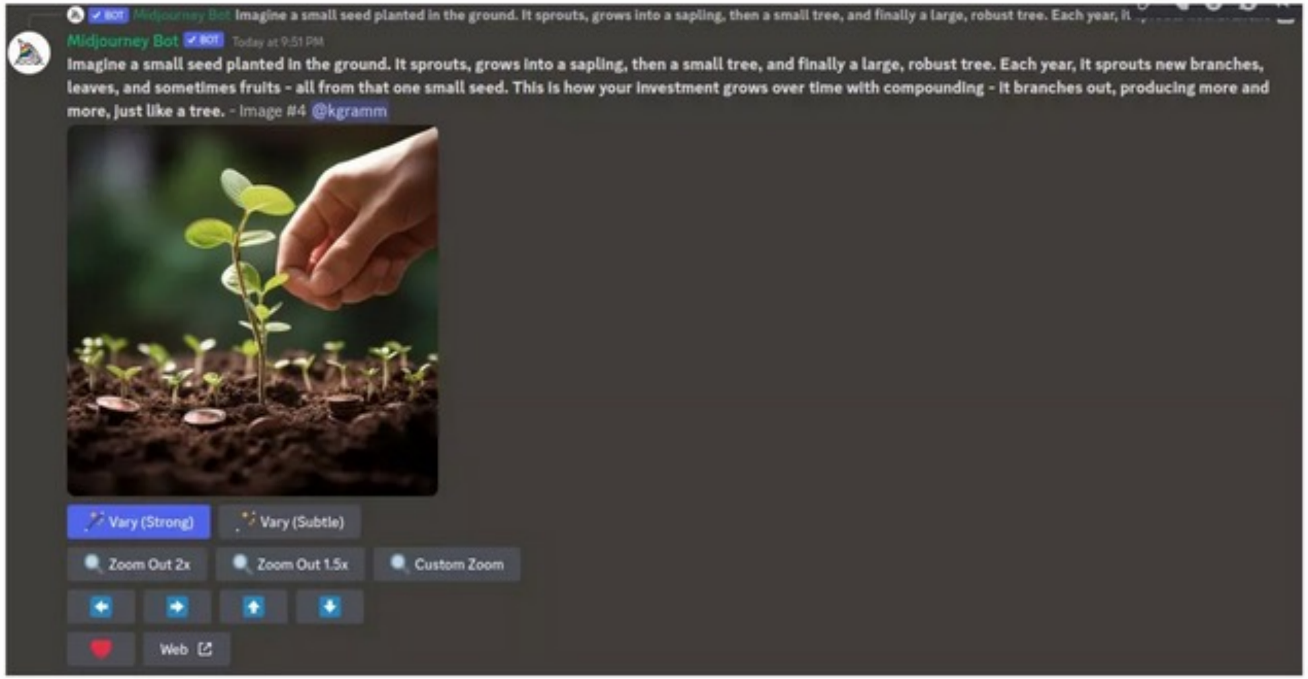
Midjourney Bot ✓ BOT Today at 9:43 PM

**Gold bars chilling a beach** - Variations (Strong) by @kgramm (fast)

# Applications

## Midjourney

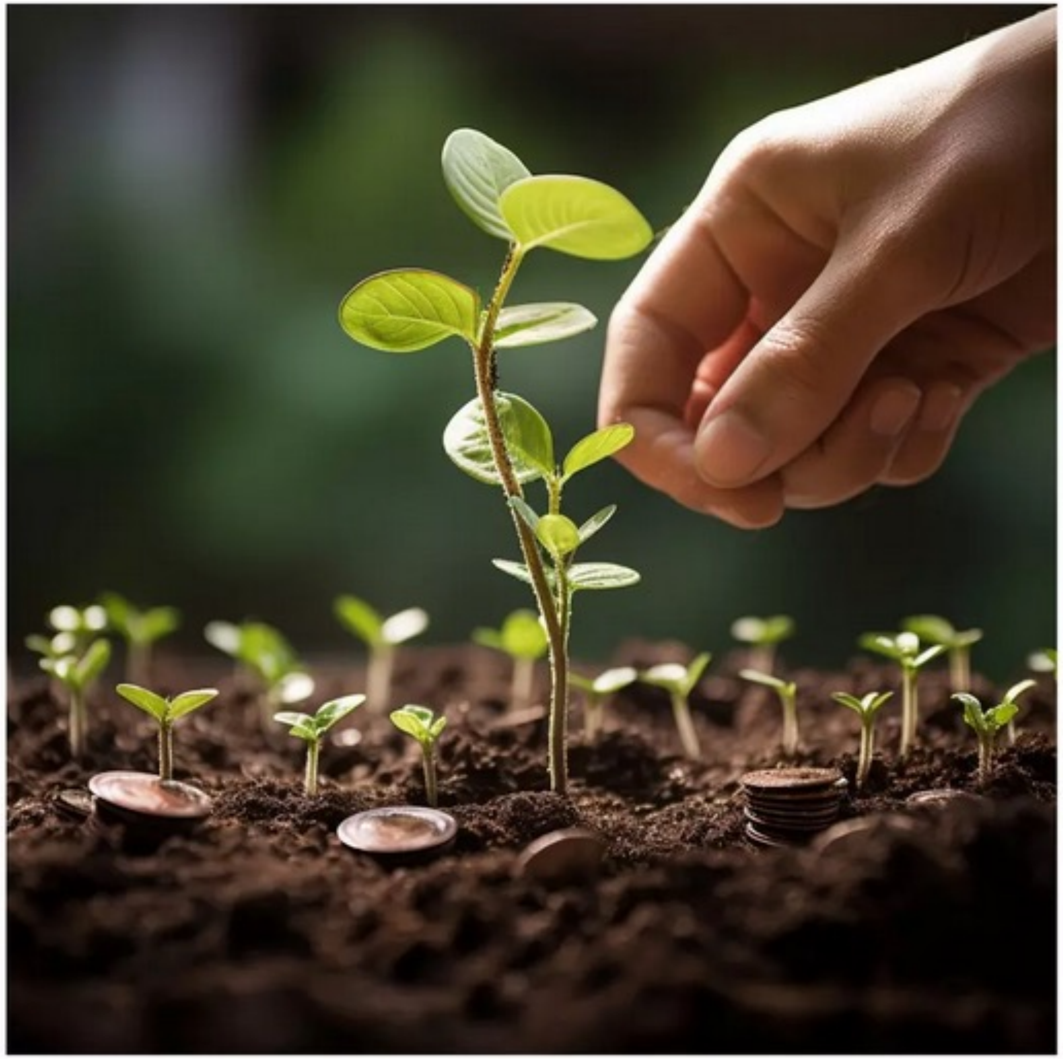
**Prompt**: *Imagine a small seed planted in the ground. It sprouts, grows into a sapling, then a small tree, and finally a large robust tree. Each year, it sprouts new branches, leaves and sometimes fruits – all from that small seed. This is how your investment grows with compounding – It branches out producing more and more just like a tree*



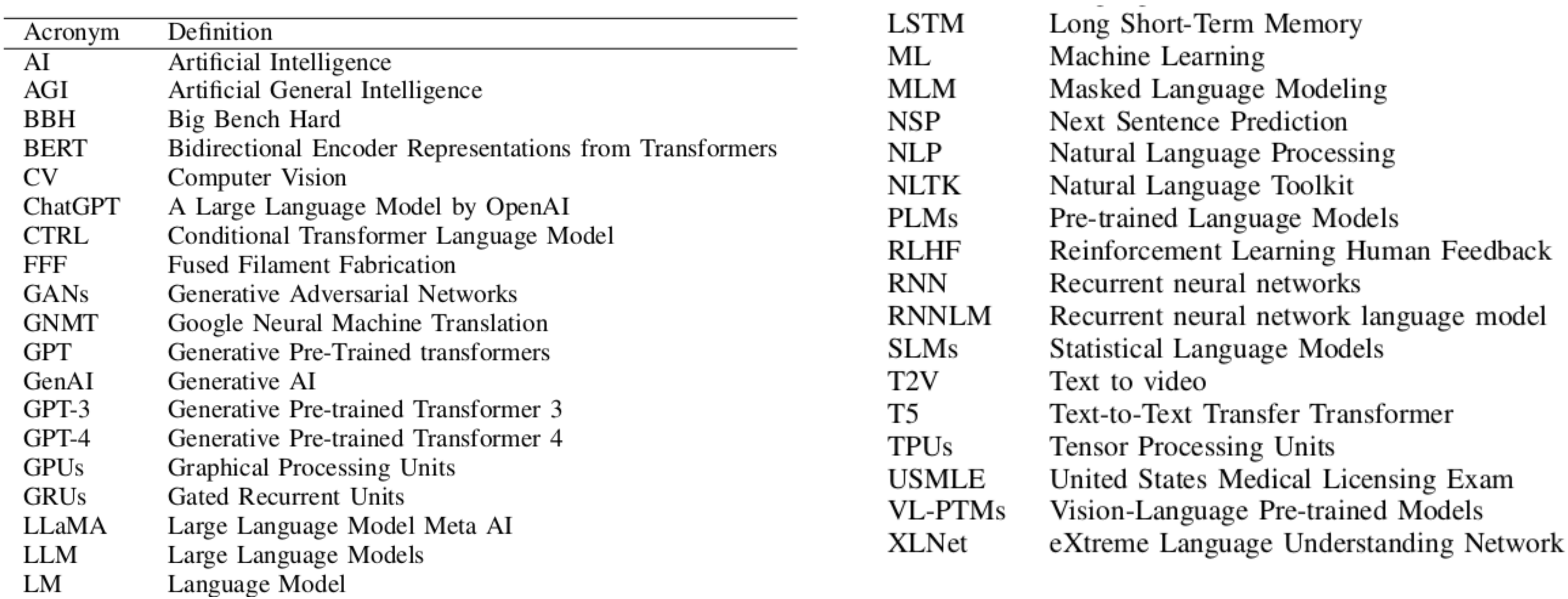
✓ BOT Midjourney Bot Imagine a small seed planted in the ground. It sprouts, grows into a sapling, then a small tree, and finally a large, robust tree. Each year, it

Midjourney Bot ✓ BOT Today at 9:51 PM

**Imagine a small seed planted in the ground. It sprouts, grows into a sapling, then a small tree, and finally a large, robust tree. Each year, it sprouts new branches, leaves, and sometimes fruits - all from that one small seed. This is how your investment grows over time with compounding - it branches out, producing more and more, just like a tree.** - Image #4 @kgramm



# Acronym and their Definitions



| Acronym | Definition |
|---|---|
| AI | Artificial Intelligence |
| AGI | Artificial General Intelligence |
| BBH | Big Bench Hard |
| BERT | Bidirectional Encoder Representations from Transformers |
| CV | Computer Vision |
| ChatGPT | A Large Language Model by OpenAI |
| CTRL | Conditional Transformer Language Model |
| FFF | Fused Filament Fabrication |
| GANs | Generative Adversarial Networks |
| GNMT | Google Neural Machine Translation |
| GPT | Generative Pre-Trained transformers |
| GenAI | Generative AI |
| GPT-3 | Generative Pre-trained Transformer 3 |
| GPT-4 | Generative Pre-trained Transformer 4 |
| GPUs | Graphical Processing Units |
| GRUs | Gated Recurrent Units |
| LLaMA | Large Language Model Meta AI |
| LLM | Large Language Models |
| LM | Language Model |
| LSTM | Long Short-Term Memory |
| ML | Machine Learning |
| MLM | Masked Language Modeling |
| NSP | Next Sentence Prediction |
| NLP | Natural Language Processing |
| NLTK | Natural Language Toolkit |
| PLMs | Pre-trained Language Models |
| RLHF | Reinforcement Learning Human Feedback |
| RNN | Recurrent neural networks |
| RNNLM | Recurrent neural network language model |
| SLMs | Statistical Language Models |
| T2V | Text to video |
| T5 | Text-to-Text Transfer Transformer |
| TPUs | Tensor Processing Units |
| USMLE | United States Medical Licensing Exam |
| VL-PTMs | Vision-Language Pre-trained Models |
| XLNet | eXtreme Language Understanding Network |

Thank You!