

Practical Guide to Generative A.I.

Johnson Shum

Ailog Limited

For education purpose only



▶ ⏪ 🔊 0:35 / 10:23



Introducing Sora — OpenAI's text-to-video model

 **OpenAI**
87.3萬位訂閱者

訂閱

👍 4.8萬 | 🗨️ | ➦ 分享 | ⬇️ 下載 | ⋮



YouTube 首頁

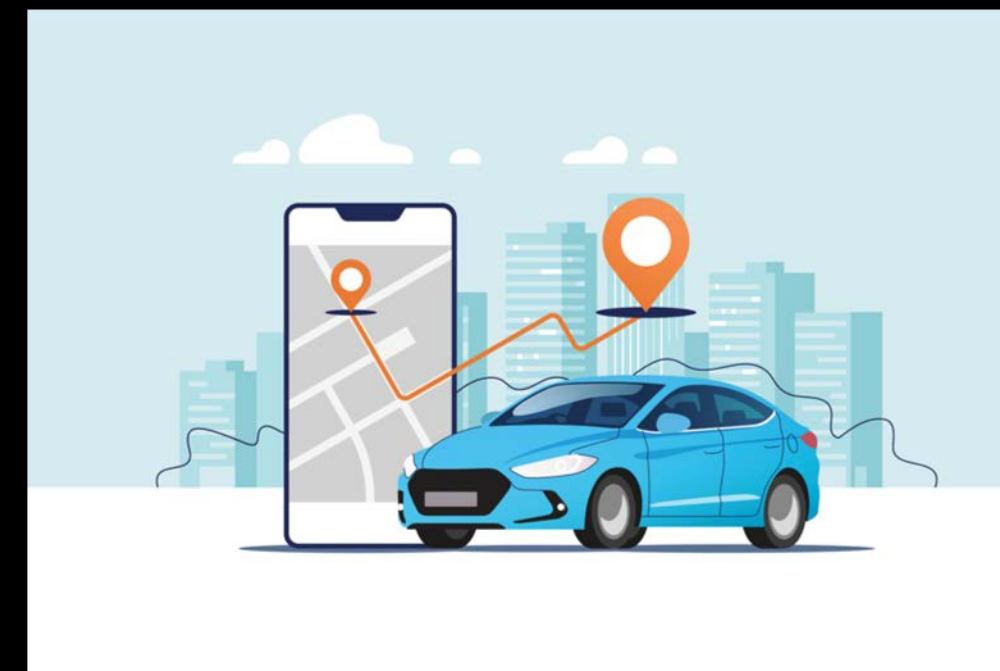
https://www.youtube.com/watch?v=HK6y8DAPN_0

What is AI

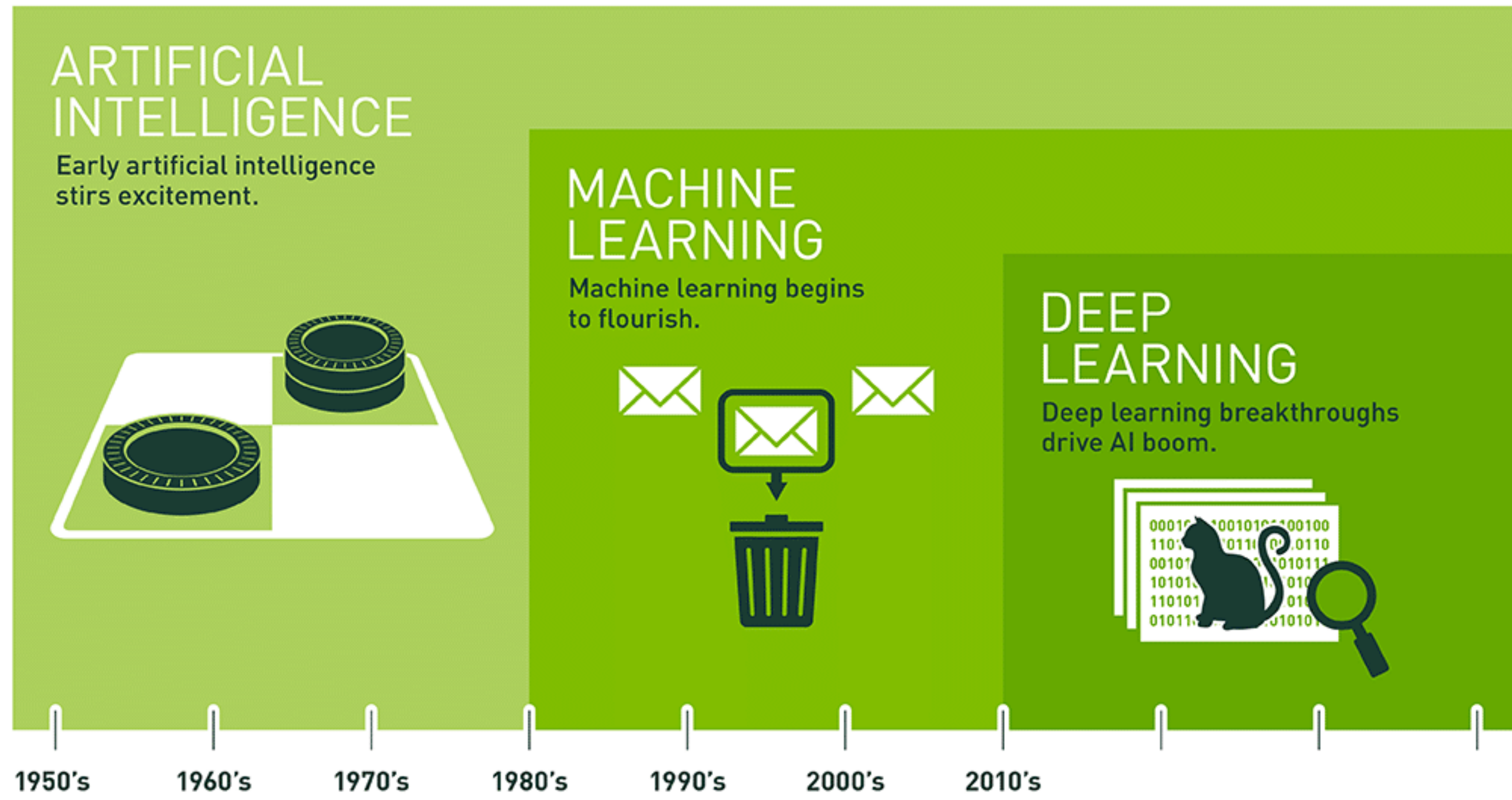
What is AI

Computer or computing system can perform tasks that have typically required human intelligence

- Smart" assistants like Apple's Siri or Amazon's Alexa
- Search and recommendation algorithms for suggesting movies on Netflix for your next binge-watching session
- Route optimization for drivers, shipping and logistics companies, and self-driving cars
- Automated stock trading and investing



What is AI



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

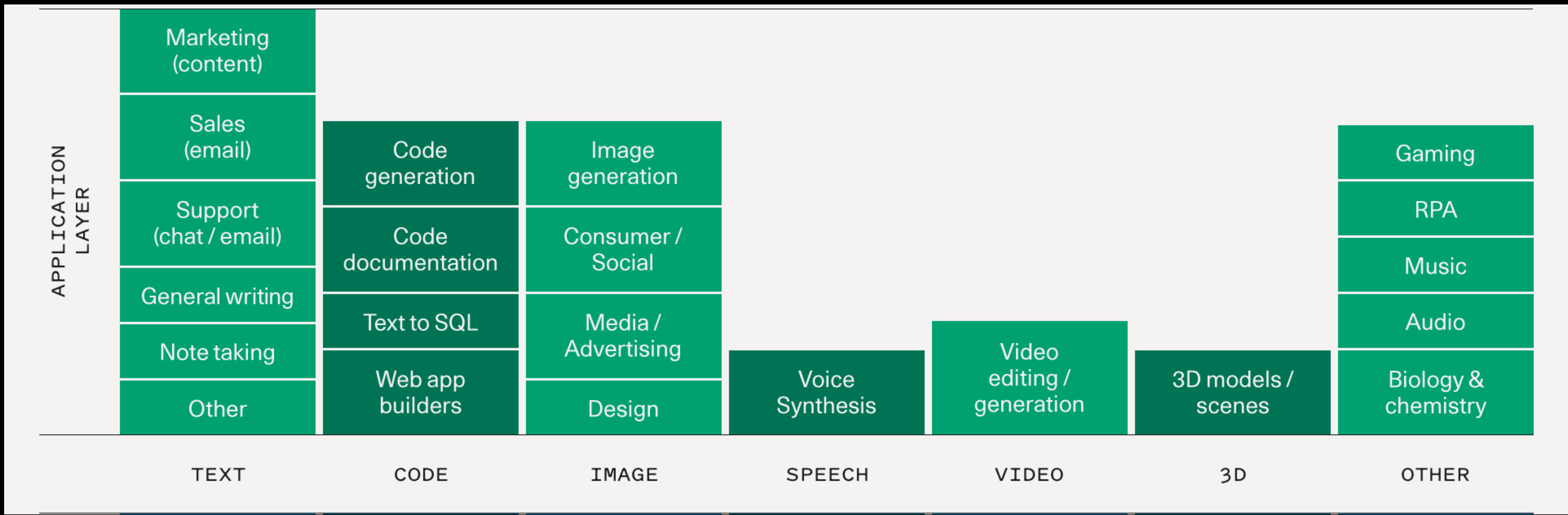
Introduction to Generative AI

What is Generative AI (Gen-AI)

- A class of machine learning algorithms
- Uses neural networks to create text, images, and other content
- Produces content that is significantly different from its training data
- Capable of generating highly complex output
- Represents advancements beyond previous machine learning models



What can generative AI do



What can generative AI do

Outline my talk "Practical Guide to Generative Artificial Intelligence"

Assistant Poe

Title: Practical Guide to Generative Artificial Intelligence

I. Introduction

- A. Briefly introduce the concept of generative artificial intelligence (AI)
- B. Explain the significance and potential applications of generative AI
- C. State the objective of the talk: to provide a practical guide for implementing generative AI

II. Understanding Generative AI

- A. Define generative AI and its key components
- B. Discuss the difference between generative and discriminative models
- C. Highlight popular generative AI algorithms, such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs)

III. Data Collection and Preprocessing

- A. Discuss the importance of high-quality and diverse training data
- B. Explain techniques for collecting and preprocessing data for generative AI models
- C. Address potential challenges and considerations in data collection and preprocessing

IV. Model Architecture and Training

Text



Images



尹光 Wan Kwong 《一人之境》 Official Music Video (原唱：林家謙) (AI Cover) #ai尹光 #尹光 #尹光ai...

Music and voice

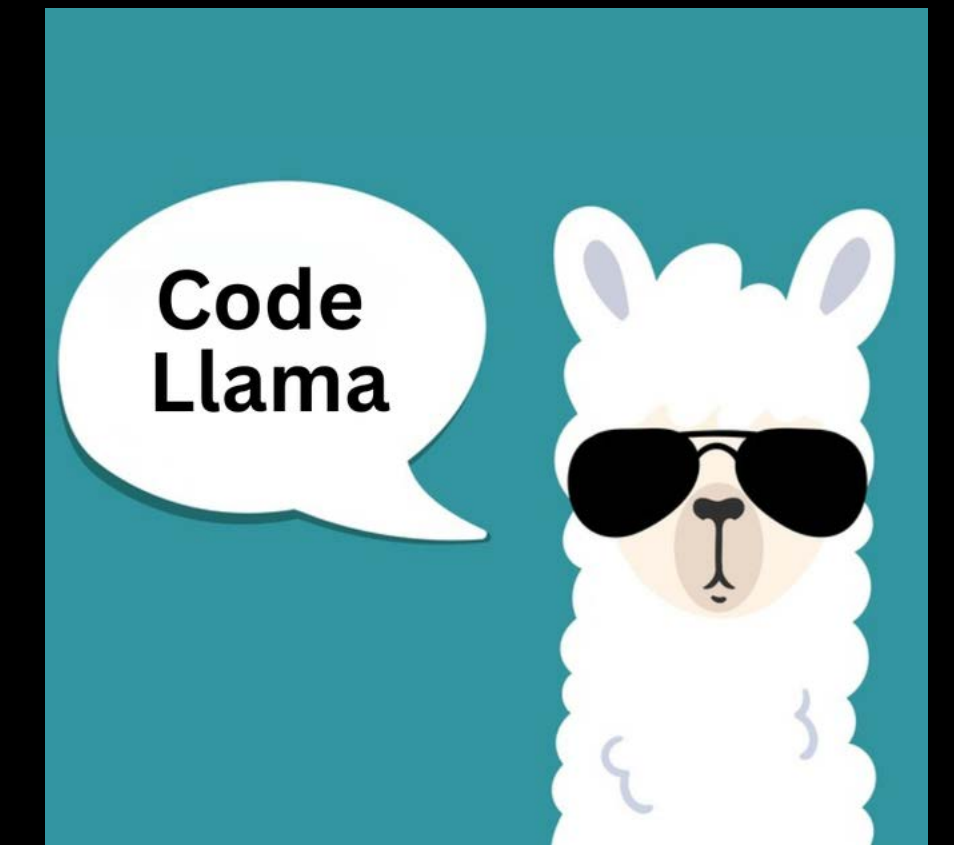
<https://www.youtube.com/watch?v=3JDUR-eIYWA>

What is Foundation Models:

- Base for specific generative AI applications
- Deep learning models trained on massive data sets that are terabytes in size
- Primarily trained to predict missing words in texts (text completion)
- Require fine-tuning for more specific tasks
- Capable of parsing natural language, generating new text and images, and engaging in conversation
- Large language models (LLMs) are the most important type of foundation models for many businesses.

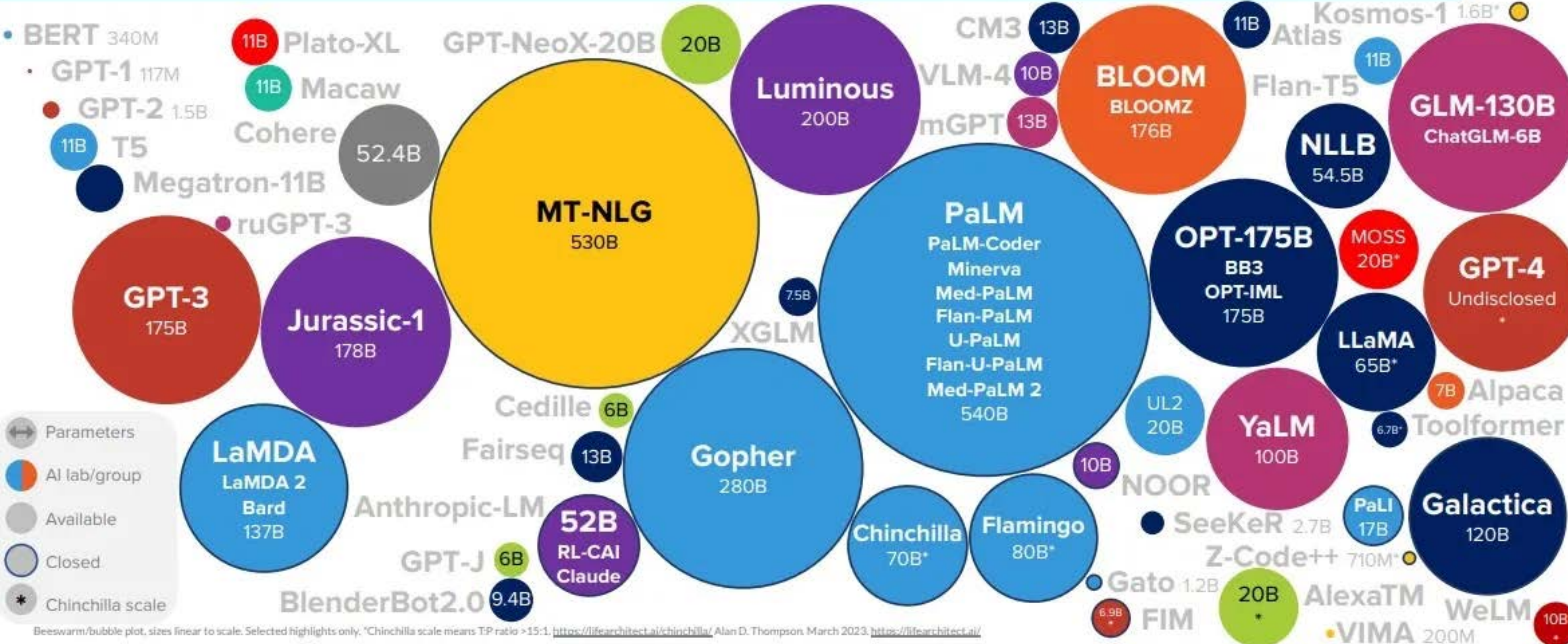
What is Large Language Models (LLM)

- Advanced AI models designed to work with text
- Generate content, summarizing text, and understanding natural language commands
- Trained on massive amounts of text data from the internet, learning patterns and structures in language
- Ability to understand context, generate coherent responses, and mimic human conversation.



Generative AI Model Size

LANGUAGE MODEL SIZES TO MAR/2023



Beeswarm/bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means T:P ratio >15:1. <https://lilresearch.github.io/chinchilla/> Alan D. Thompson, March 2023. <https://lilresearch.github.io/>

Compare to simpler machine learning models

Simpler Machine Learning Models

- Main purpose: **Discovering patterns** in training data and producing query results
- Strength: Sophisticated data **analysis** and impressive feats
- Range of outputs: **Smaller** compared to generative AI
- Knowledge base: More **limited**
- Example: Predicting customer churn likelihood from **data**

Generative AI

- Main purpose: **Creating** content with infinite variety in form and substance
- Strength: **Generating** complex and diverse outputs
- Outputs: **Text**, audio, **image**, video etc.
- Knowledge base: Usually very **large**
- Example: Making churn predictions and taking action through **personalized emails**

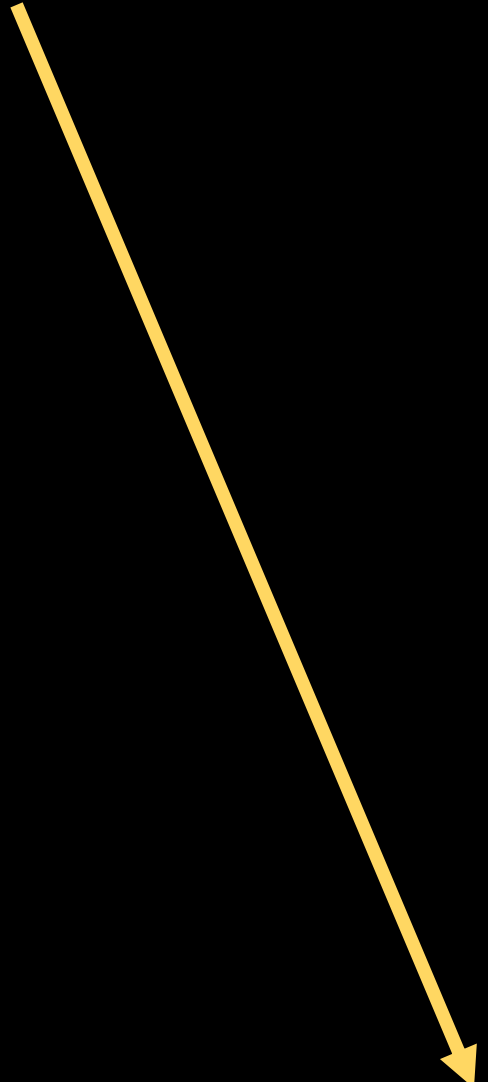
What is Generative Pre-trained Transformer

Generative Pre-trained Transformer (GPT)



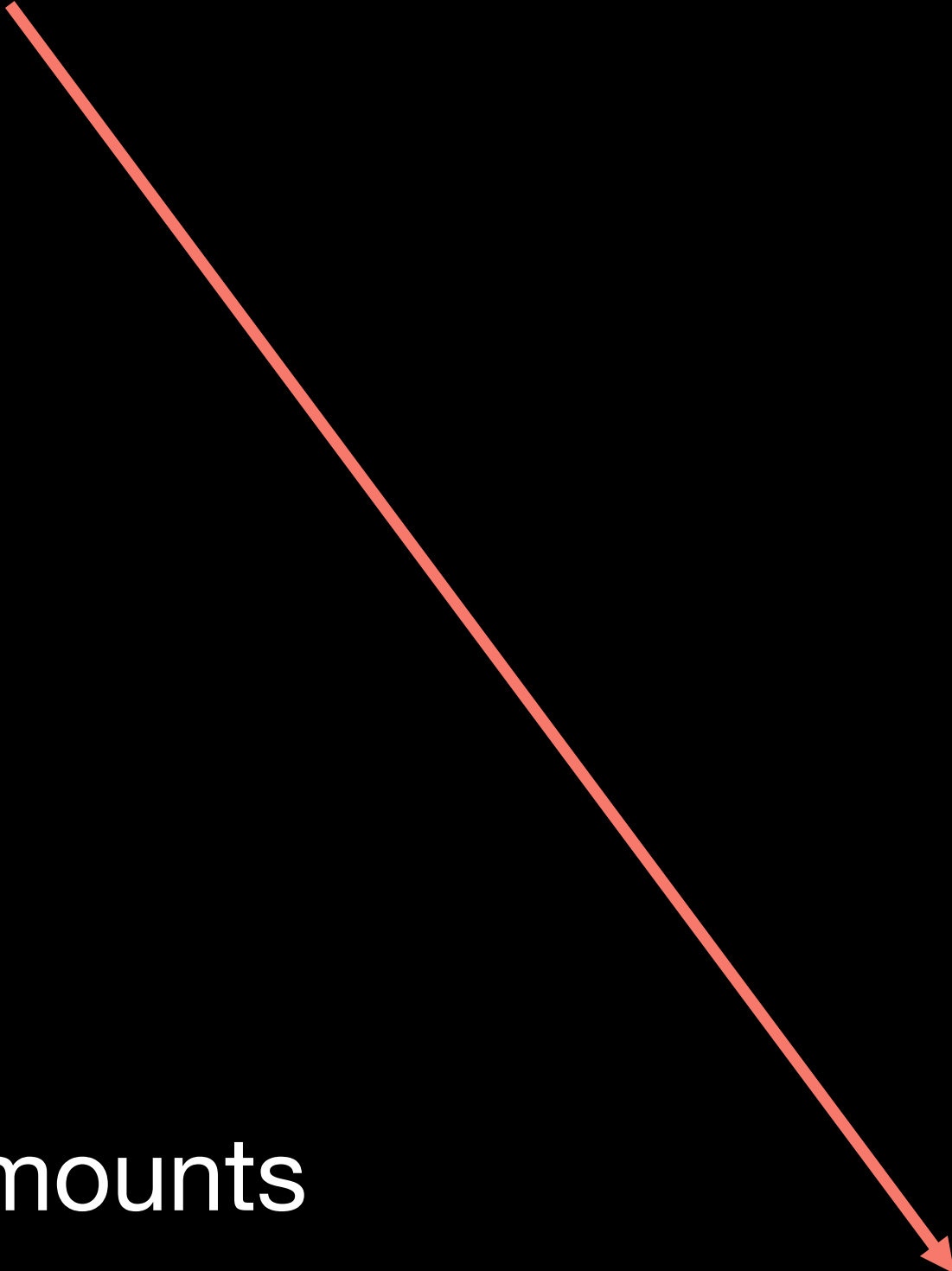
Generative

Means next word prediction



Pre-trained

The LLM is pretrained with massive amounts of text

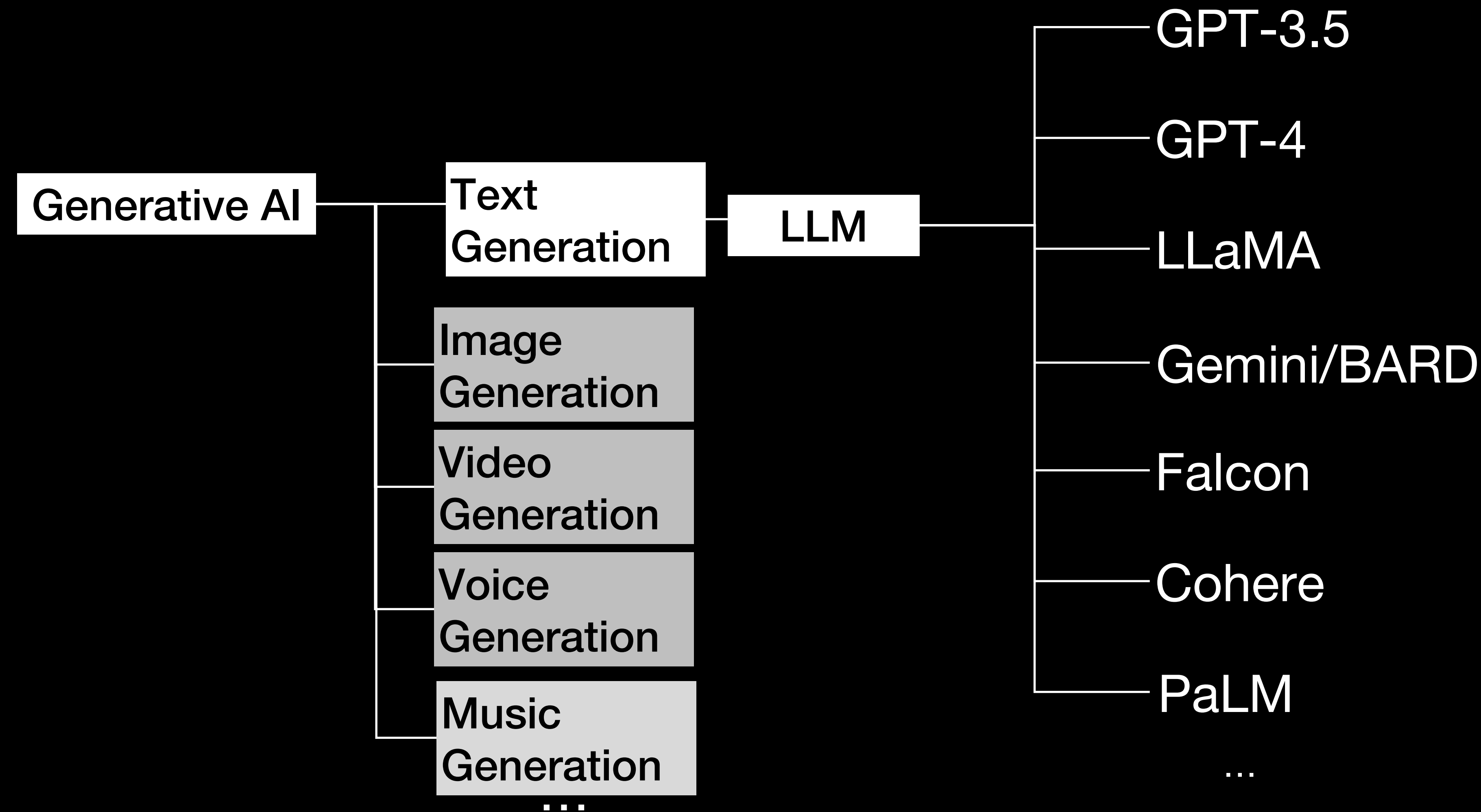


Transformer

The neural network architecture

Large Language Models

Gen-AI and Common Large Language Models (LLM)



Large language models is a branch of generative AI, focus specifically on generating **human-like text** by learning from extensive training data, enabling coherent and contextually relevant text generation

Large Language Models

1. Pretraining

Massive amounts of data from the internet + books + etc.



2. Instruction Fine-tuning

- Teaching the model to respond to instructions.
- Model learns to respond to instructions.
- Helps alignment



3. Reinforcement Learning from Human Feedback

- Similar purpose to instruction tuning.
- Helps produce output that is closer to what humans want or like.



Pre-training

- Initial stage of training for large language models (LLMs).
- The model learns to predict the next word by analyzing massive amounts of data.
- The model acquires grammar, syntax, and general knowledge
- Alignment with human intentions makes the LLM suitable for assistant-like tasks.
- Popular language platforms like Quora or StackOverflow provide more structured data that can help improve alignment.

Instruction fine-tuning

- Instruction fine-tuning is the next stage after pre-training for LLMs
- This stage focuses on training the model to respond well to instructions and align with human intentions.
- High-quality instruction-response pairs are used as training data for this stage, which are more expensive to create and typically sourced from humans
- The size of the instruction dataset is typically smaller compared to the pre-training dataset.

Reinforcement Learning from Human Feedback

- Third stage that some LLMs, like ChatGPT
- RLHF further improves alignment and ensures that the LLM's output reflects human values and preferences.
- Achieving or surpassing human-level performance and driving significant improvements in LLMs.

Token

The minimum unit of processing for LLMs is called a token.



The GPT family of models process text using **tokens**, which are common sequences of characters found in text. The models understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.



BERT uses tokens as the basic input units to process text. These tokens are generated through the process of tokenization, which involves breaking down the input text into smaller subword units or characters.

Tokenizer

GPT-3.5 & GPT-4 GPT-3 (Legacy)

In the context of Language Models (LLM), a token refers to a unit of text that the model processes. It can be as short as a single character or as long as a word or even a subword. In most cases, tokens are typically words or subwords. LLMs break down input text into tokens to understand and generate coherent language. Each token typically corresponds to a unique numerical representation that the model uses for processing. The number of tokens in an LLM can affect computational resources and model performance, as longer texts require more tokens and memory to process.

Clear Show example

Tokens	Characters
115	574

In the context of Language Models (LLM), a token refers to a unit of text that the model processes. It can be as short as a single character or as long as a word or even a subword. In most cases, tokens are typically words or subwords. LLMs break down input text into tokens to understand and generate coherent language. Each token typically corresponds to a unique numerical representation that the model uses for processing. The number of tokens in an LLM can affect computational resources and model performance, as longer texts require more tokens and memory to process.

Text Token IDs

<https://platform.openai.com/tokenizer>

There is no strict definition of a token. It can be a word, a group of words, punctuation, or even a part of a word (sub-text).

According to the ChatGPT LLM tokenizer, some general rules of thumb for defining tokens are:

1 token \approx 4 chars in English

1 token \approx $\frac{3}{4}$ words

100 tokens \approx 75 words

Or

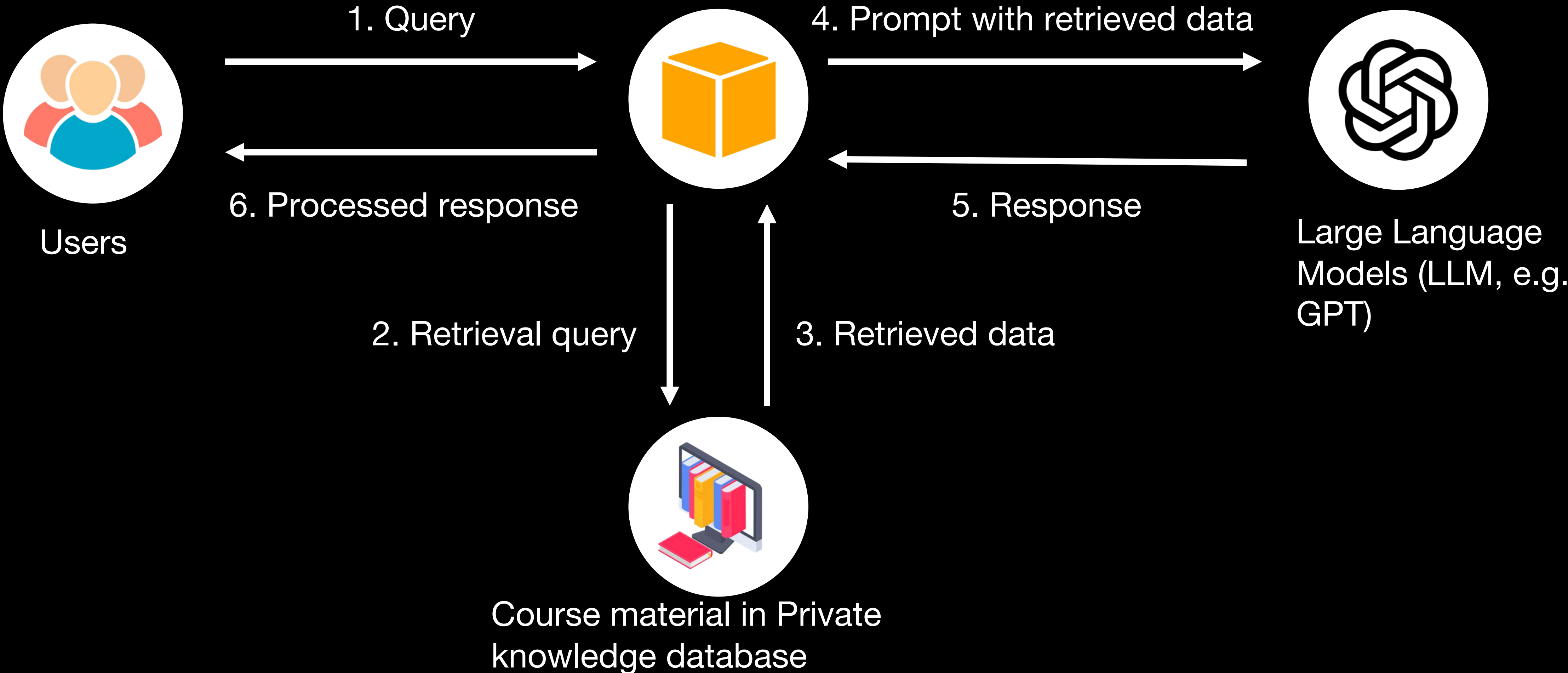
1–2 sentences \approx 30 tokens

1 paragraph \approx 100 tokens

1,500 words \approx 2048 tokens

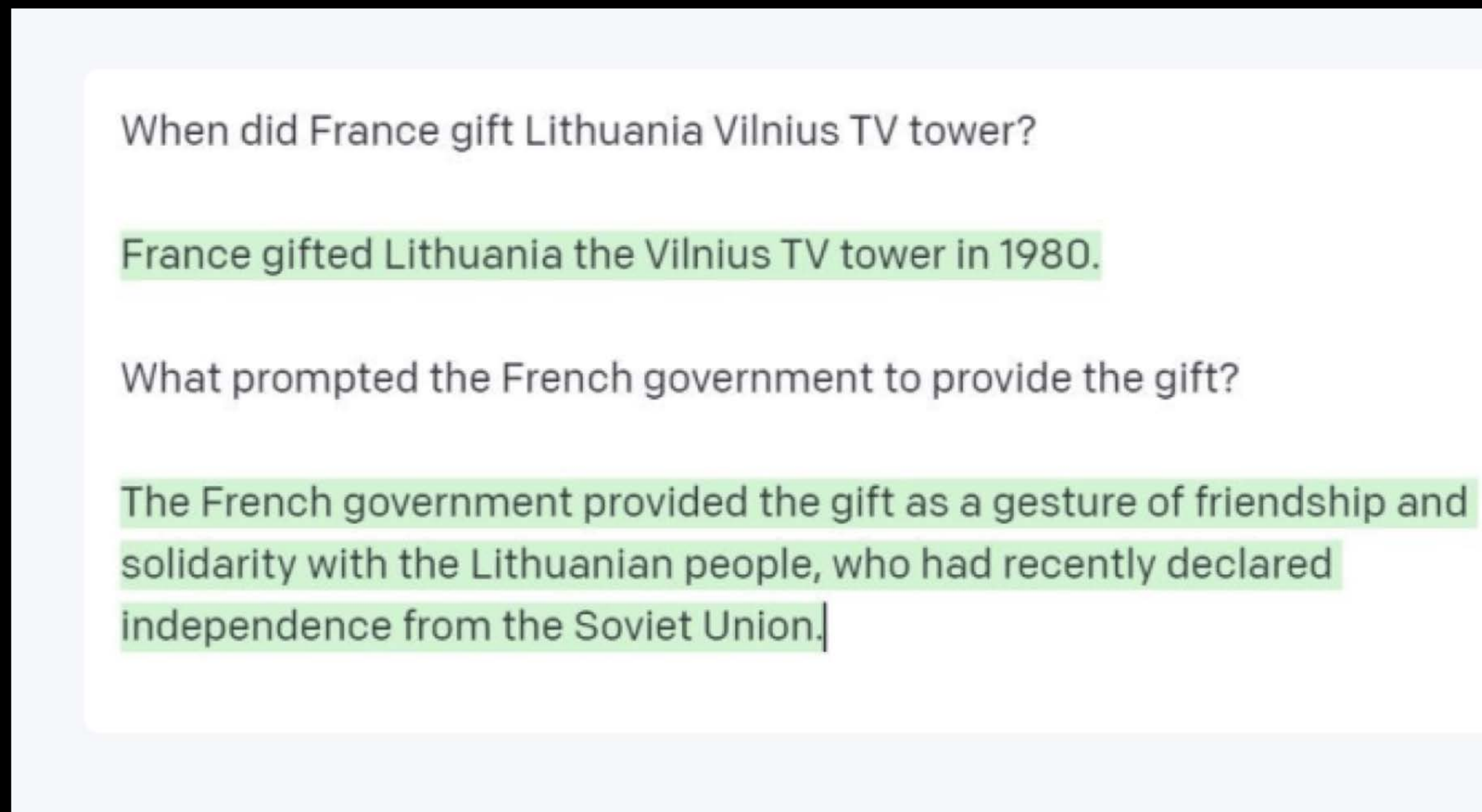
Retrieval-Augmented Generation (RAG)

Allow large language model (LLM) to answer with private knowledge base



AI common problems

GPT 4

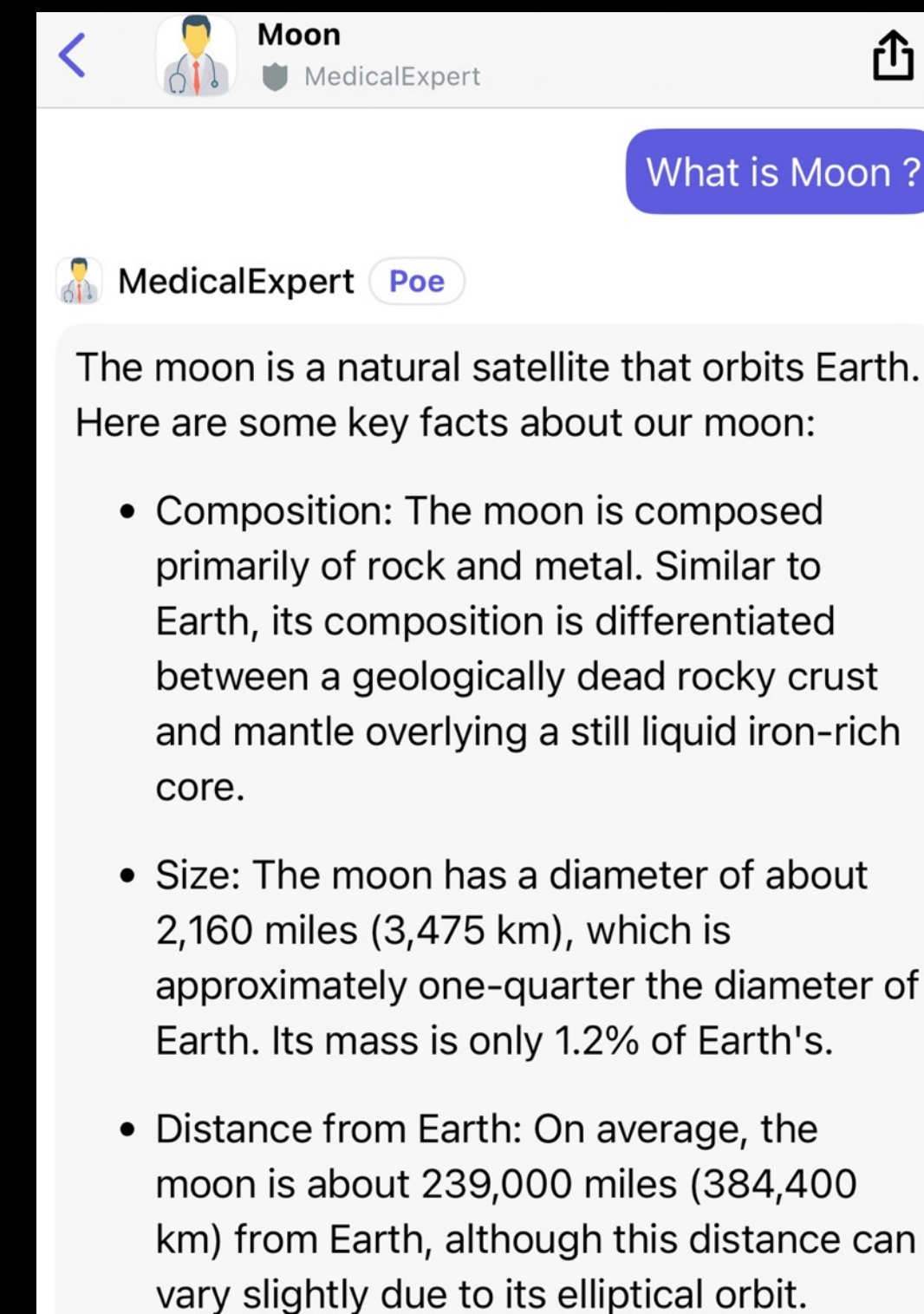


France never gifted the Vilnius TV tower to Lithuania

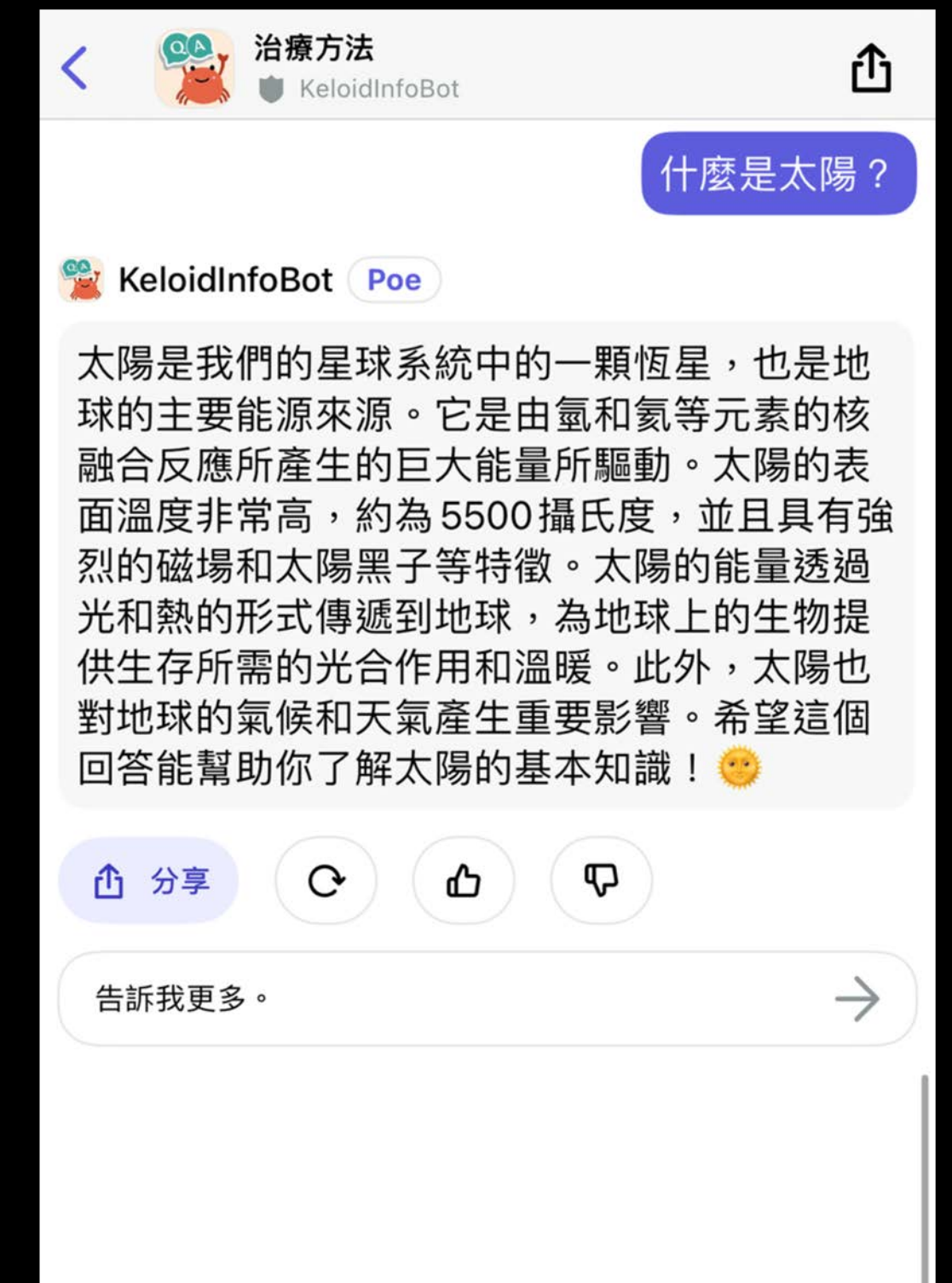
AI hallucination when an LLM **presents false info**

Source: <https://www.tidio.com/blog/ai-hallucinations/>

POE Bot creation (GPT 3.5)



Medical knowledge bots answer unrelated things



AI created answers that are **out of its training scope**

How to evaluate LLMs

LLM benchmarks



Reasoning and Commonsense

Apply logic and knowledge to solve problems



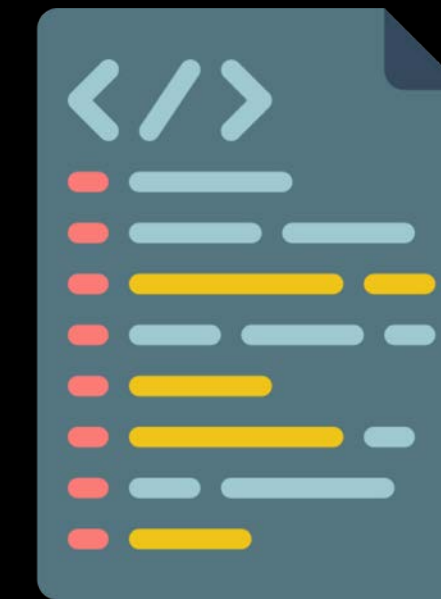
Language Understanding and Question Answering

interpret text and answer questions accurately



Conversation

Engage in dialogue and provide relevant responses



Coding

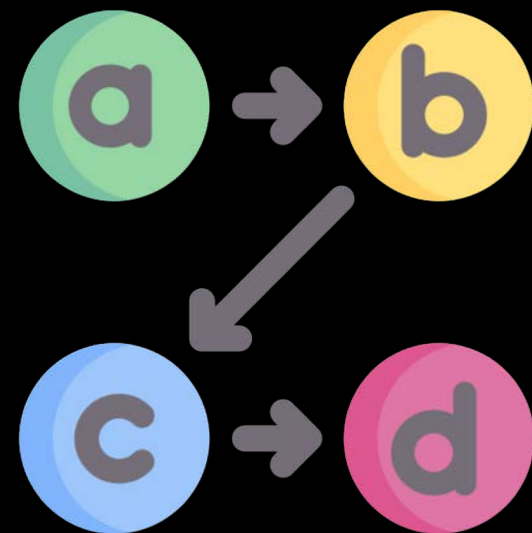
Interpret and generate code

LLM benchmarks



Translation

Accurately translate text



Logic

Logical reasoning like inductive reasoning



Math

Math problems, from arithmetic to calculus



Standardized Tests

SAT, ACT, or other educational assessments

LLM benchmarks

Massive Multitask Language Understanding (MMLU)

Broad Spectrum of Knowledge

- Wide array of subjects covered
- Includes humanities, social sciences, hard sciences, and critical fields
- Total of 57 distinct tasks
- Ensures thorough evaluation across multiple domains of knowledge

Volume of Questions

- 15,908 questions
- Few-shot development set
- Validation set
- Test set

Comprehensive Coverage per Subject

- Each subject represented by a minimum of 100 test examples
- Extensive coverage for deep evaluation
- Ensures broad evaluation of model's capabilities in each subject area

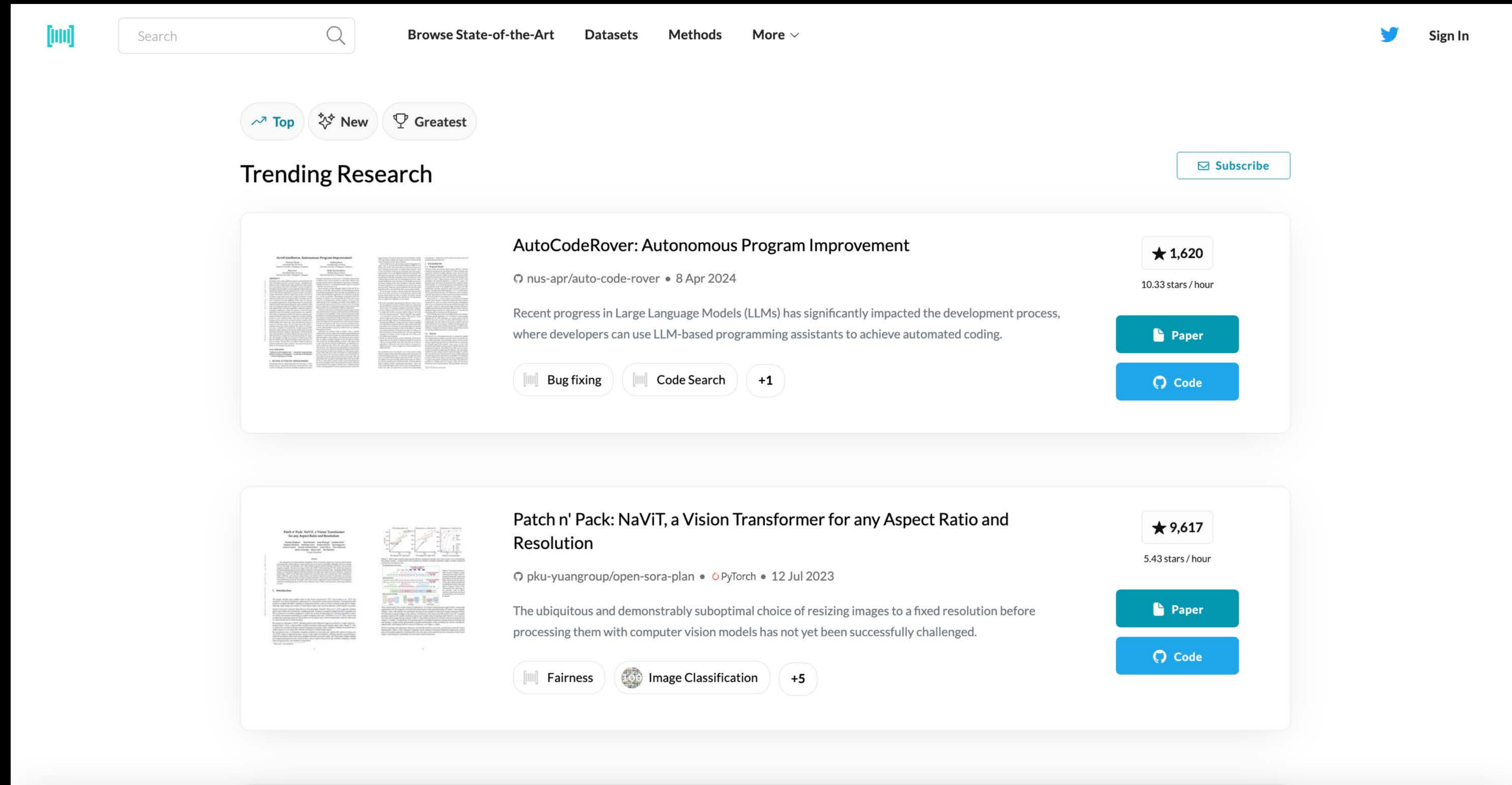
LLM benchmarks

The screenshot shows the LLM Benchmark interface with various filters and a table of model performance metrics. The filters include Model types (pretrained, continuously pretrained, fine-tuned on domain-specific datasets, chat models, base merges and moerges), Precision (float16, bfloat16, 8bit, 4bit, GPTQ), and Model sizes (in billions of parameters). The table lists models such as davidkim205/Rhea-72b-v0.5, MTSAIR/MultiVerse_70B, SF-Foundation/Ein-72B-v0.11, and others, with columns for Average, ARC, HellaSwag, MMLU, and TruthfulQA scores.

T	Model	Average	ARC	HellaSwag	MMLU	TruthfulQA
◆	davidkim205/Rhea-72b-v0.5	81.22	79.78	91.15	77.95	74.5
💬	MTSAIR/MultiVerse_70B	81	78.67	89.77	78.22	75.18
◆	MTSAIR/MultiVerse_70B	80.98	78.58	89.74	78.27	75.09
◆	SF-Foundation/Ein-72B-v0.11	80.81	76.79	89.02	77.2	79.02
◆	SF-Foundation/Ein-72B-v0.13	80.79	76.19	89.44	77.07	77.82
◆	SF-Foundation/Ein-72B-v0.12	80.72	76.19	89.46	77.17	77.78
◆	abacusai/Smaug-72B-v0.1	80.48	76.02	89.27	77.15	76.67
◆	ibivibiv/alpaca-dragon-72b-v1	79.3	73.89	88.16	77.4	72.69

https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

LLM benchmarks



The screenshot shows the 'Papers with Code' website interface. At the top, there is a search bar and navigation links for 'Browse State-of-the-Art', 'Datasets', 'Methods', and 'More'. A 'Sign In' button is located in the top right corner. Below the navigation, there are three filter buttons: 'Top', 'New', and 'Greatest'. The main section is titled 'Trending Research' and features a 'Subscribe' button. Two research papers are highlighted:

- AutoCodeRover: Autonomous Program Improvement**
 - Repository: nus-apr/auto-code-rover
 - Date: 8 Apr 2024
 - Stars: 1,620 (10.33 stars / hour)
 - Description: Recent progress in Large Language Models (LLMs) has significantly impacted the development process, where developers can use LLM-based programming assistants to achieve automated coding.
 - Tags: Bug fixing, Code Search, +1
 - Buttons: Paper, Code
- Patch n' Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution**
 - Repository: pku-yuangroup/open-sora-plan
 - Framework: PyTorch
 - Date: 12 Jul 2023
 - Stars: 9,617 (5.43 stars / hour)
 - Description: The ubiquitous and demonstrably suboptimal choice of resizing images to a fixed resolution before processing them with computer vision models has not yet been successfully challenged.
 - Tags: Fairness, Image Classification, +5
 - Buttons: Paper, Code

<https://paperswithcode.com/>

Massive Multitask Language Understanding (MMLU)

Few-shot Development Set

- Concise set of questions for each subject
- 5 questions per topic
- Crucial role in initial development and testing stages

Validation Set

- Validation set consists of 1,540 questions
- Vital tool for fine-tuning the model
- Used for selecting hyper-parameters
- Critical step in optimizing language model performance

Test Set

- Test set is the most extensive part of the assessment
- Includes 14,079 questions covering various subjects
- Designed for rigorous evaluation of language understanding and application

Massive Multitask Language Understanding (MMLU)

MMLU Rankings

Rank	Company	Model	MMLU	License
1	OpenAI	gpt-4-0314	86%	Proprietary
2	Anthropic	claude-2.0	79%	Proprietary
3	Anthropic	claude-1	77%	Proprietary
4	Mistral	mistral-medium	75%	Proprietary
5	01 AI	yi-34b-chat	74%	Yi License
6	Anthropic	claude-instant-1	73%	Proprietary
7	Google	gemini-pro	72%	Proprietary
8	Google	gemini-pro-dev-api	72%	Proprietary
9	Mistral	mixtral-8x7b-instruct-v0.1	71%	Apache 2.0
10	OpenAI	gpt-3.5-turbo-0314	70%	Proprietary
11	TII	falcon-180b-chat	68%	Falcon-180B TII License
12	Alibaba	qwen-14b-chat	67%	Qianwen LICENSE
13	Upstage AI	solar-10.7b-instruct-v1.0	66%	CC-BY-NC-4.0
14	OpenChat	openchat-3.5	64%	Apache-2.0
15	UC Berkeley	starling-1m-7b-alpha	64%	CC-BY-NC-4.0
16	Microsoft	wizardlm-70b	64%	Llama 2 Community
17	Meta	llama-2-70b-chat	63%	Llama 2 Community
18	UW	guanaco-65b	62%	Non-commercial
19	HuggingFace	zephyr-7b-beta	61%	MIT
20	LMSYS	vicuna-33b	59%	Non-commercial

Analysis: BRACAI

Source: <https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard>

HellaSwag

HellaSwag

ActivityNet

WikiHow

MC Complete the following statement

Question: Make chocolate avocado mousse. Swap out packaged chocolate mousse or pudding for homemade avocado mousse. Avocado is a heart healthy-vegetables and this recipe will give you a serving of vegetables, making it a great alternative.

The Choices:

[A] Melt 1/2 cup of dark or semi sweet chocolate chips in the microwave for 30-60 seconds. Stir until they are all melted and smooth.

[B] Here's how to make a simple but nutritious avocado mousse :
[substeps] Wash and slice half of your avocado. Next, wash and slice the avocado.

[C] Cool the mousse on the stove or in the microwave for at least an hour, at the minimum. Create a butternut mousse.

[D] Make your own slices of avocado. Avocado also goes great in salads and served with tacos as an added treat.

[70,000 multiple-choice questions](#)

HellaSwag

Rank	Model	Accuracy↑	Paper	Code	Result	Year
1	CompassMTL 567M with Tailor	96.1	Task Compass: Scaling Multi-task Pre-training with Task Prefix			2022
2	CompassMTL 567M	95.6	Task Compass: Scaling Multi-task Pre-training with Task Prefix			2022
3	DeBERTa-Large 304M (classification-based)	95.6	Two is Better than Many? Binary Classification as an Effective Approach to Multi-Choice Question Answering			2022
4	GPT-4 (10-shot)	95.3	GPT-4 Technical Report			2023
5	DeBERTa-Large 304M	94.7	Two is Better than Many? Binary Classification as an Effective Approach to Multi-Choice Question Answering			2022
6	Unicorn 11B (fine-tuned)	93.9	UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark			2021
7	DeBERTa++	93	DeBERTa: Decoding-enhanced BERT with Disentangled Attention			2020
8	ELECTRA-Large 335M (fine-tuned on DiscoSense and HellaSwag)	91.5	DiscoSense: Commonsense Reasoning with Discourse Connectives			2022
9	DBRX Instruct 132B (10-shot)	89				2024
10	TheBloke/llama-2-70b-Guanaco-QLoRA-fp16 (10-shot)	88.3				
11	ALBERT-XXL 235M	88				2021
12	PaLM 2-L (1-shot)	87.4	PaLM 2 Technical Report			2023
13	ELECTRA-Large 335M (fine-tuned on HellaSwag)	86.9	DiscoSense: Commonsense Reasoning with Discourse Connectives			2022
14	PaLM 2-M (1-shot)	86.7	PaLM 2 Technical Report			2023
15	MUPPET Roberta Large	86.4	Muppet: Massive Multi-task Representations with Pre-Finetuning			2021
16	LLaMA 65B + CFG (0-shot)	86.3	Stay on topic with Classifier-Free Guidance			2023
17	Falcon-180B (0-shot)	85.9	The Falcon Series of Open Language Models			2023
18	PaLM 2-S (1-shot)	85.6	PaLM 2 Technical Report			2023
19	GPT-3.5 (10-shot)	85.5	GPT-4 Technical Report			2023
20	RoBERTa-Large Ensemble	85.5	RoBERTa: A Robustly Optimized BERT Pretraining Approach			2019
21	LLaMA 30B + CFG (0-shot)	85.3	Stay on topic with Classifier-Free Guidance			2023
22	LLaMA 2 70B (0-shot)	85.3	Llama 2: Open Foundation and Fine-Tuned Chat Models			2023
23	HyKAS+CSKG	85.0	Towards Generalizable Neuro-Symbolic Systems for Commonsense Question Answering			2019
24	LLaMA 65B (0-shot)	84.2	LLaMA: Open and Efficient Foundation Language Models			2023

Wino Grande

- A benchmark for commonsense reasoning
- Recent advances in neural language models have achieved around 90% accuracy on variants of WSC.
- WinoGrande is introduced as a large-scale dataset of 44k problems, inspired by WSC but adjusted to improve scale and difficulty.

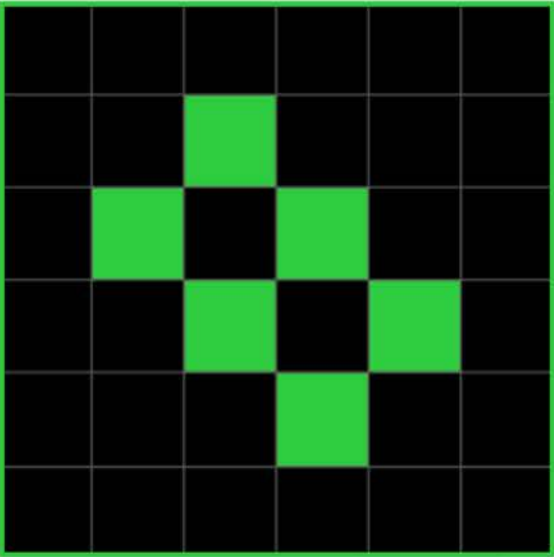
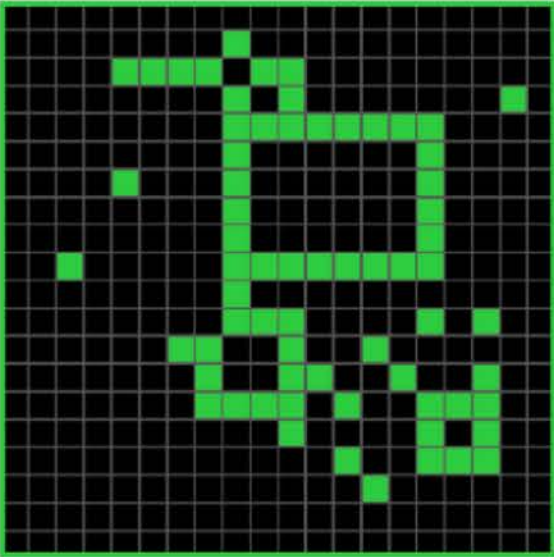
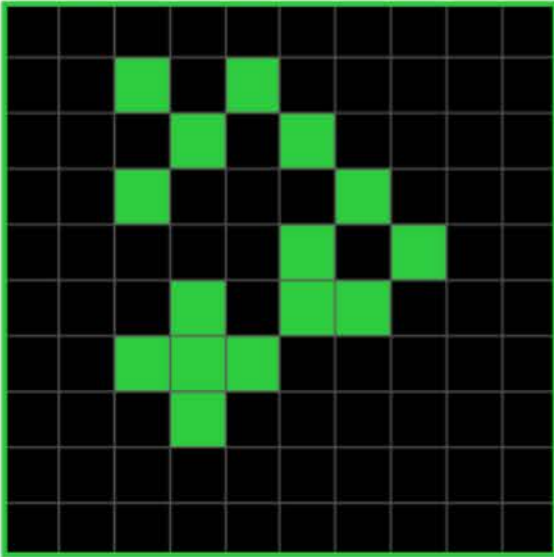
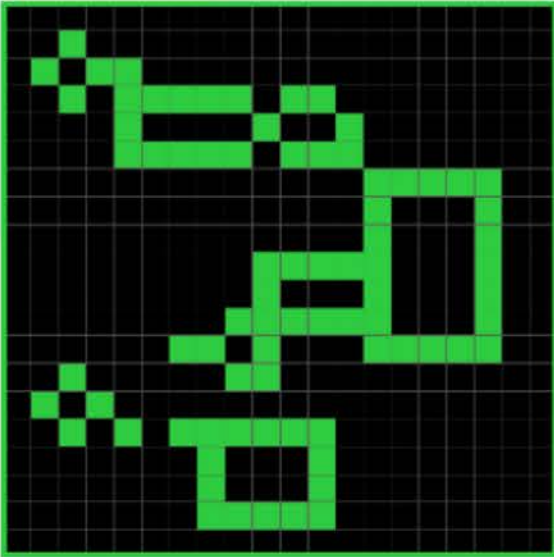




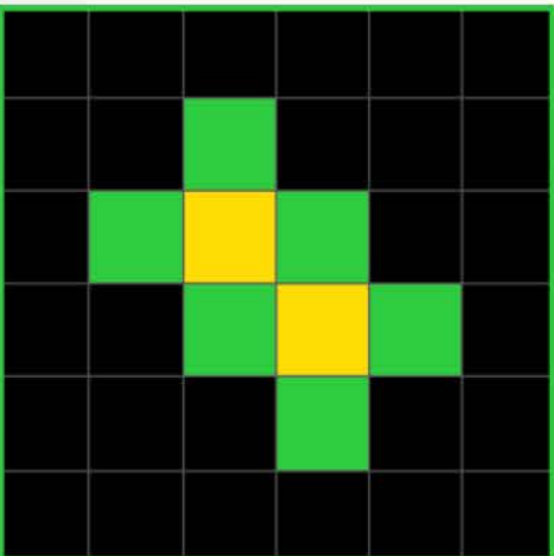
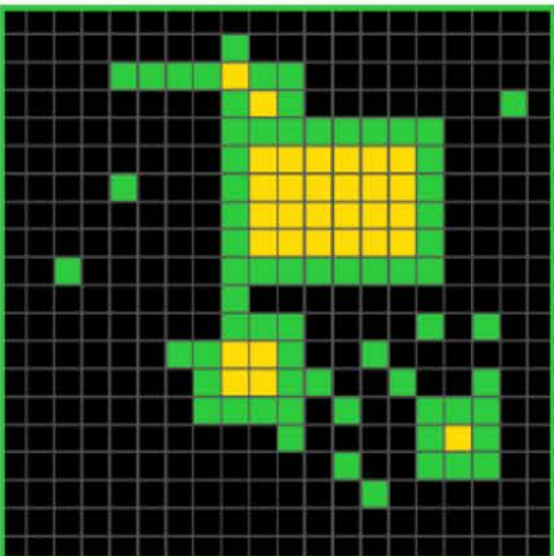
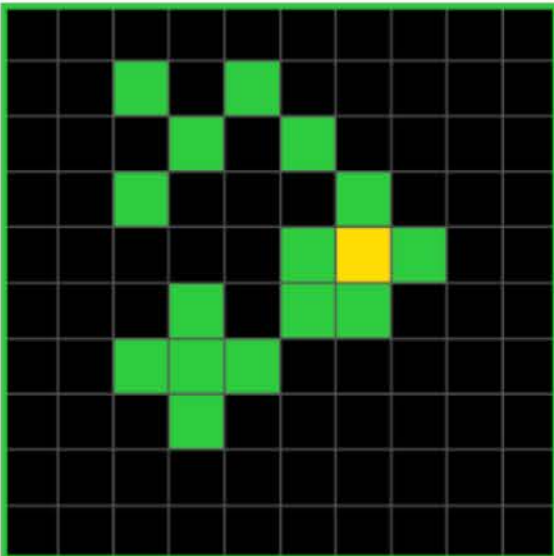
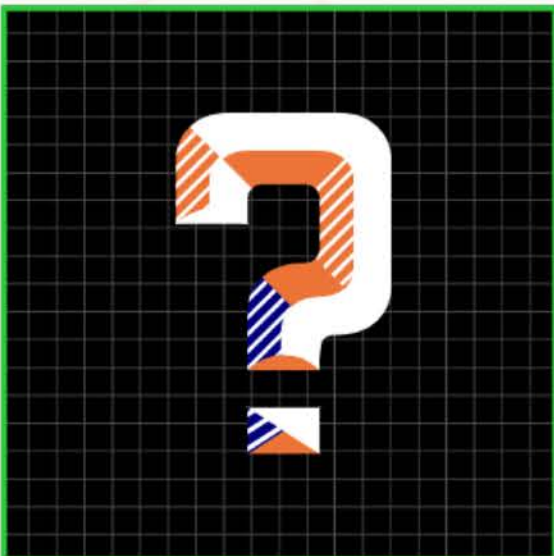
	Twin sentences	Options (answer)
x	The monkey loved to play with the balls but ignored the blocks because he found them <u>exciting</u> . The monkey loved to play with the balls but ignored the blocks because he found them <u>dull</u> .	balls / blocks balls / blocks
x	William could only climb beginner walls while Jason climbed advanced ones because he was very <u>weak</u> . William could only climb beginner walls while Jason climbed advanced ones because he was very <u>strong</u> .	William / Jason William / Jason
✓	Robert woke up at 9:00am while Samuel woke up at 6:00am, so he had <u>less</u> time to get ready for school. Robert woke up at 9:00am while Samuel woke up at 6:00am, so he had <u>more</u> time to get ready for school.	Robert / Samuel Robert / Samuel
✓	The child was screaming after the baby bottle and toy fell. Since the child was <u>hungry</u> , it stopped his crying. The child was screaming after the baby bottle and toy fell. Since the child was <u>full</u> , it stopped his crying.	baby bottle / toy baby bottle / toy

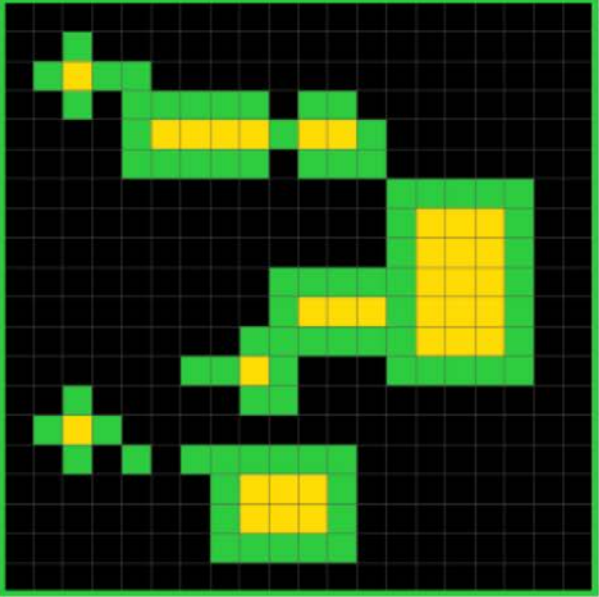
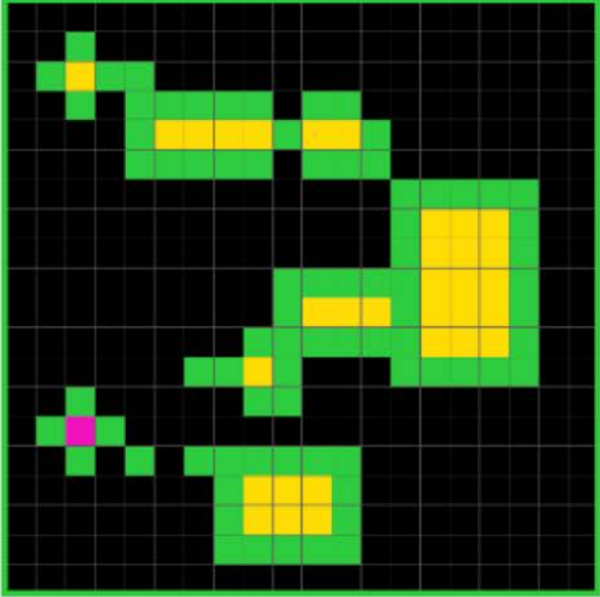


Wino Grande

Rank	Model	Accuracy ↑	Paper	Code	Result	Year
1	ST-MoE-32B 269B (fine-tuned)	96.1	ST-MoE: Designing Stable and Transferable Sparse Expert Models			2022
2	Unicorn 11B (fine-tuned)	91.3	UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark			2020
3	T5-XXL 11B (fine-tuned)	91				2021
4	CompassMTL 567M with Tailor	90.5	Task Compass: Scaling Multi-task Pre-training with Task Prefix			2022
5	CompassMTL 567M	89.6	Task Compass: Scaling Multi-task Pre-training with Task Prefix			2022
6	UnifiedQA 11B (fine-tuned)	89.4	UnifiedQA: Crossing Format Boundaries With a Single QA System			2020
7	Claude 3 Opus (5-shot)	88.5	The Claude 3 Model Family: Opus, Sonnet, Haiku			2024
8	GPT-4 (5-shot)	87.5	GPT-4 Technical Report			2023
9	ExDeBERTa 567M	87	Task Compass: Scaling Multi-task Pre-training with Task Prefix			2022
10	PaLM 2-L (1-shot)	83.0	PaLM 2 Technical Report			2023
11	ST-MoE-L 4.1B (fine-tuned)	81.7	ST-MoE: Designing Stable and Transferable Sparse Expert Models			2022
12	GPT-3.5 (5-shot)	81.6	GPT-4 Technical Report			2023

Rank	Model	Accuracy ↑	Paper	Code	Result	Year
13	PaLM 540B (0-shot)	81.1	PaLM: Scaling Language Modeling with Pathways			2022
14	Camelidae-8x34B	80.9	Parameter-Efficient Sparsity Crafting from Dense to Mixture-of-Experts for Instruction Tuning on General Tasks			2024
15	PaLM 2-M (1-shot)	79.2	PaLM 2 Technical Report			2023
16	RoBERTa-Winogrande 355M (fine-tuned)	79.1	WinoGrande: An Adversarial Winograd Schema Challenge at Scale			2019
17	PaLM 2-S (1-shot)	77.9	PaLM 2 Technical Report			2023
18	Mixtral 8x7B (0-shot)	77.2	Mixtral of Experts			2024
19	PaLM 62B (0-shot)	77.0	PaLM: Scaling Language Modeling with Pathways			2022
20	PaLM-cont 62B (0-shot)	77.0	PaLM: Scaling Language Modeling with Pathways			2022
21	LLaMA 65B (0-shot)	77.0	LLaMA: Open and Efficient Foundation Language Models			2023
22	LLaMA 33B (0-shot)	76.0	LLaMA: Open and Efficient Foundation Language Models			2023
23	Mistral 7B (0-shot)	75.3	Mistral 7B			2023
24	Claude 3 Sonnet (5-shot)	75.1	The Claude 3 Model Family: Opus, Sonnet, Haiku			2024

Arc Challenge (25-shot)

Example 1: Input	Example 2: Input	Example 3: Input	Test: Input
			
			
Example 1: Output	Example 2: Output	Example 3: Output	Test: Output
			

Test: Output	Test: Output
	
	

Arc Challenge (25-shot)

Rank	Model	Accuracy↑	Extra Training Data	Paper	Code	Result	Year
1	GPT-4 (few-shot, k=25)	96.4	×	GPT-4 Technical Report			2023
2	PaLM 2 (few-shot, CoT, SC)	95.1	×	PaLM 2 Technical Report			2023
3	Claude 2 (few-shot, k=5)	91	×	Model Card and Evaluations for Claude Models			2023
4	Claude 1.3 (few-shot, k=5)	90	×	Model Card and Evaluations for Claude Models			2023
5	PaLM 540B (Self Improvement, Self Consistency)	89.8	×	Large Language Models Can Self-Improve			2022
6	PaLM 540B (Self Consistency)	88.7	×	Large Language Models Can Self-Improve			2022
7	PaLM 540B (Self Improvement, CoT Prompting)	88.3	×	Large Language Models Can Self-Improve			2022
8	PaLM 540B (Self Improvement, Standard-Prompting)	87.2	×	Large Language Models Can Self-Improve			2022
9	PaLM 540B (Standard-Prompting)	87.1	×	Large Language Models Can Self-Improve			2022
10	ST-MoE-32B 269B (fine-tuned)	86.5	×	ST-MoE: Designing Stable and Transferable Sparse Expert Models			2022
11	Claude Instant 1.1 (few-shot, k=5)	85.7	×	Model Card and Evaluations for Claude Models			2023
12	GPT-3.5 (few-shot, k=25)	85.2	×	GPT-4 Technical Report			2023

Rank	Model	Accuracy↑	Extra Training Data	Paper	Code	Result	Year
13	PaLM 540B (CoT Prompting)	85.2	×	Large Language Models Can Self-Improve			2022
14	PaLM 2-L (1-shot)	69.2	×	PaLM 2 Technical Report			2023
15	GAL 120B (zero-shot)	67.9	✓	Galactica: A Large Language Model for Science			2022
16	Camelidae-8x34B	65.2	×	Parameter-Efficient Sparsity Crafting from Dense to Mixture-of-Experts for Instruction Tuning on General Tasks			2024
17	PaLM 2-M (1-shot)	64.9	×	PaLM 2 Technical Report			2023
18	FLAN 137B (few-shot, k=13)	63.8	×	Finetuned Language Models Are Zero-Shot Learners			2021
19	FLAN 137B (zero-shot)	63.1	×	Finetuned Language Models Are Zero-Shot Learners			2021
20	PaLM 2-S (1-shot)	59.6	×	PaLM 2 Technical Report			2023
21	LLaMA 33B (zero-shot)	57.8	×	LLaMA: Open and Efficient Foundation Language Models			2023
22	ST-MoE-L 4.1B (fine-tuned)	56.9	×	ST-MoE: Designing Stable and Transferable Sparse Expert Models			2022
23	LLaMA 65B (zero-shot)	56.0	✓	LLaMA: Open and Efficient Foundation Language Models			2023
24	Mistral 7B (0-shot)	55.5	×	Mistral 7B			2023

Comparison


	LLaMA 2 70B	GPT - 3.5	Mixtral 8x7B
MMLU (MCQ in 57 subjects)	69.9%	70.0%	70.6%
HellaSwag (10-shot)	87.1%	85.5%	86.7%
ARC Challenge (25-shot)	85.1%	85.2%	85.8%
WinoGrande (5-shot)	83.2%	81.6%	81.2%
MBPP (pass@1)	49.8%	52.2%	60.7%
GSM-8K (5-shot)	53.6%	57.1%	58.4%
MT Bench (for Instruct Models)	6.86	8.32	8.30

Common websites

<https://huggingface.co/>

Hugging Face Search models, datasets, users... Models Datasets Spaces Posts Docs Pricing Log In Sign Up

NEW Try Cohere Command R+ on HuggingChat



The AI community building the future.

The platform where the machine learning community collaborates on models, datasets, and applications.

Tasks Libraries Datasets Languages Licenses Other

Filter Tasks by name

Multimodal

- Text-to-Image Image-to-Text
- Text-to-Video Visual Question Answering
- Document Question Answering Graph Machine Learning

Computer Vision

- Depth Estimation Image Classification
- Object Detection Image Segmentation
- Image-to-Image Unconditional Image Generation
- Video Classification Zero-Shot Image Classification

Natural Language Processing

- Text Classification Token Classification
- Table Question Answering Question Answering
- Zero-Shot Classification Translation
- Summarization Conversational
- Text Generation Text2Text Generation
- Sentence Similarity

Audio

- Text-to-Speech Automatic Speech Recognition
- Audio-to-Audio Audio Classification
- Voice Activity Detection

Tabular

- Tabular Classification Tabular Regression

Reinforcement Learning

- Reinforcement Learning Robotics

Models 469,541 Filter by name

- meta-llama/Llama-2-70b
Text Generation • Updated 4 days ago • ↓ 25.2k • ♥ 64
- stabilityai/stable-diffusion-xl-base-0.9
Updated 6 days ago • ↓ 2.01k • ♥ 393
- openchat/openchat
Text Generation • Updated 2 days ago • ↓ 1.3k • ♥ 136
- l1lyasviel/ControlNet-v1-1
Updated Apr 26 • ♥ 1.87k
- cerspense/zeroscope_v2_XL
Updated 3 days ago • ↓ 2.66k • ♥ 334
- meta-llama/Llama-2-13b
Text Generation • Updated 4 days ago • ↓ 328 • ♥ 64
- tiiuae/falcon-40b-instruct
Text Generation • Updated 27 days ago • ↓ 288k • ♥ 899
- WizardLM/WizardCoder-15B-V1.0
Text Generation • Updated 3 days ago • ↓ 12.5k • ♥ 332
- CompVis/stable-diffusion-v1-4
Text-to-Image • Updated about 17 hours ago • ↓ 448k • ♥ 5.72k
- stabilityai/stable-diffusion-2-1
Text-to-Image • Updated about 17 hours ago • ↓ 782k • ♥ 2.81k
- Salesforce/xgen-7b-8k-inst
Text Generation • Updated 4 days ago • ↓ 6.18k • ♥ 57

Open source LLMs

Llama 2



LLAMA 2

Llama 2

Llama 2 was trained on **40% more data** than Llama 1, and has double the context length.

Llama 2

MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B	Model architecture:	Data collection for helpfulness and safety:
13B	Pretraining Tokens: 2 Trillion	Supervised fine-tuning: Over 100,000
70B	Context Length: 4096	Human Preferences: Over 1,000,000

Llama 2

Comparison of AI Models

capella

MODEL	PARAMETERS	TRAINING METHOD
Llama 2	7-70 Billion	Reinforcement Learning from Human Feedback (RLHF)
GPT-3.5	175 Billion	Supervised Fine-Tuning
Bard	137 Billion	Supervised Fine-Tuning
GPT-4	1.7 Trillion	Supervised Fine-Tuning

Benchmark (Higher is better)	MPT (7B)	Falcon (7B)	Llama-2 (7B)	Llama-2 (13B)	MPT (30B)	Falcon (40B)	Llama-1 (65B)	Llama-2 (70B)
MMLU	26.8	26.2	45.3	54.8	46.9	55.4	63.4	68.9
TriviaQA	59.6	56.8	68.9	77.2	71.3	78.6	84.5	85.0
Natural Questions	17.8	18.1	22.7	28.0	23.0	29.5	31.0	33.0
GSM8K	6.8	6.8	14.6	28.7	15.2	19.6	50.9	56.8
HumanEval	18.3	N/A	12.8	18.3	25.0	N/A	23.7	29.9
AGIEval (English tasks only)	23.5	21.2	29.3	39.1	33.8	37.0	47.6	54.2
BoolQ	75.0	67.5	77.4	81.7	79.0	83.1	85.3	85.0
HellaSwag	76.4	74.1	77.2	80.7	79.9	83.6	84.2	85.3
OpenBookQA	51.4	51.6	58.6	57.0	52.0	56.6	60.2	60.2
QuAC	37.7	18.8	39.7	44.8	41.1	43.3	39.8	49.3
Winogrande	68.3	66.3	69.2	72.8	71.0	76.9	77.0	80.2

MMLU: multilingual language models across various linguistic tasks and levels

TriviaQA: answer trivia questions

Natural Questions: answer questions accurately based on a given context. It consists of real user queries from the web

GSM8K: ability to understand and reason about general and specific information (include multiple-choice questions)

HumanEval: coherence, correctness, and relevance of model-generated text

AGIEval: evaluating the robustness and resistance of language models to adversarial attacks or biases in the input

BookQ: book-based question answering.

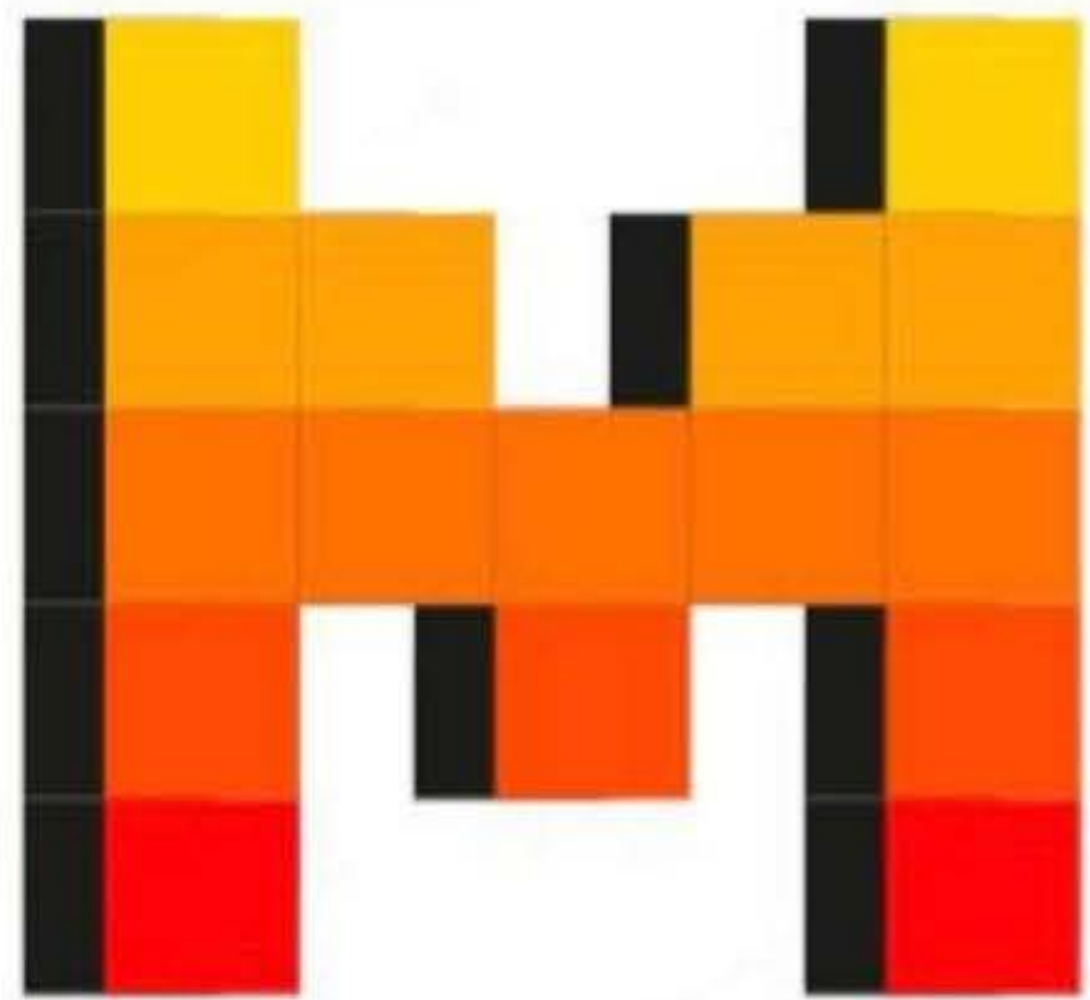
HellaSwag: Commonsense reasoning.

OpenBookQA: answer open-domain book questions

QuAC: answer conversational question evaluation

Winograd: understand and reason about contextual ambiguities in natural language

Mistral AI



**MISTRAL
AI_**

Mistral AI

Background

- Mistral AI is a French company
- Founded in April 2023
- Founded by former employees of Meta Platforms and Google DeepMind
- Produce open-source large language models
- Believe open-source software

Products

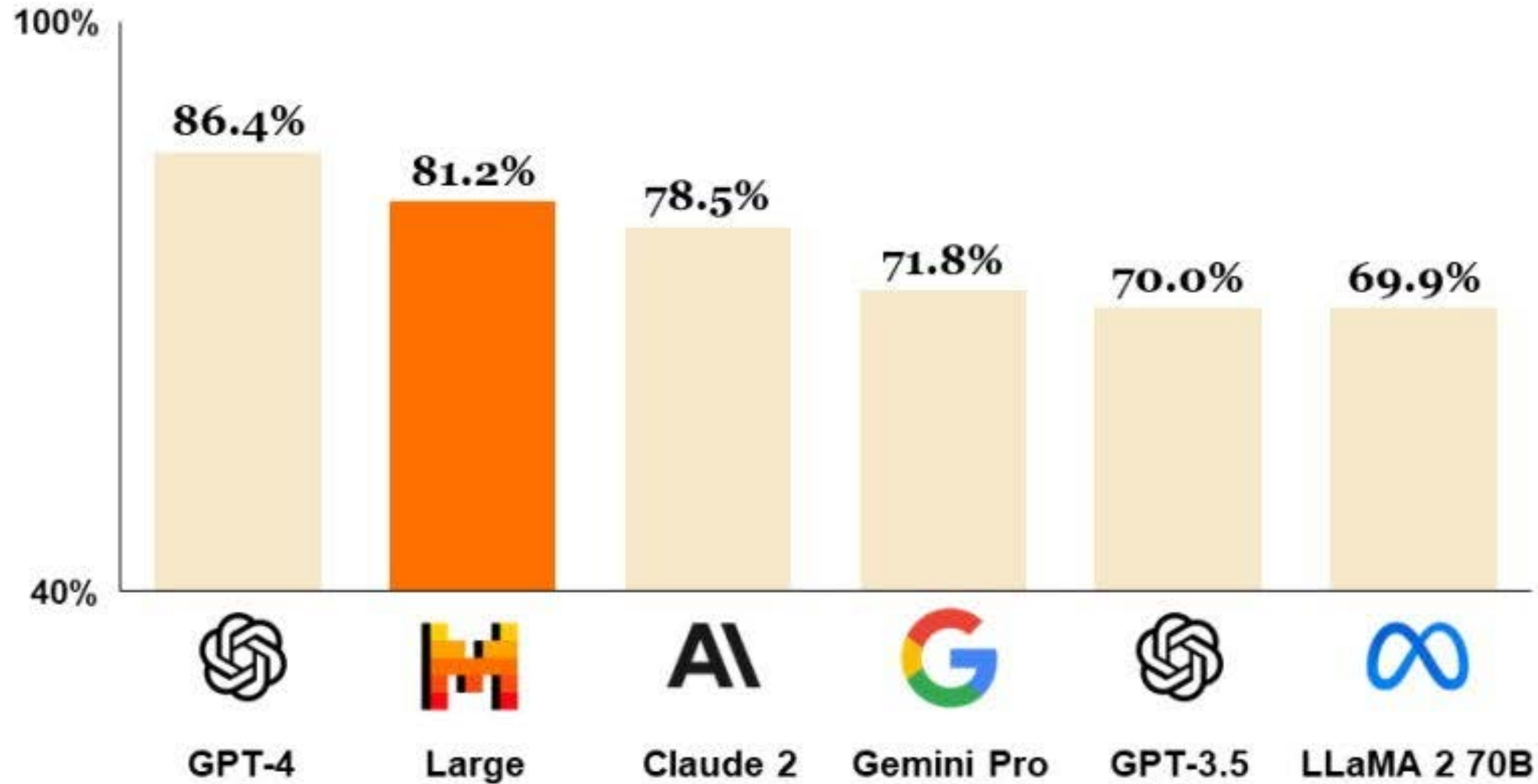
- Mistral 7B
- Mixtral 8x7B
- Mistral Medium
- Mistral Large
- Mixtral 8x22B

Mistral AI

8x7B model


- Handle contexts with up to 32,000 tokens
- The Mixtral 8x7B model has a total of 46.7 billion parameters
- The model's computational cost is comparable to a model with only 12.9 billion parameters.
- The model supports English, French, Italian, German, and Spanish
- Mixtral 8x7B Instruct, scores 8.3 on MT-Bench, making it the leading open-source instruction following model with performance comparable to GPT 3.5.

Mistral Large




MMLU test data by Mistral AI

Open source LLM



MODEL: Mixtral-8x7b-32768 ▼

Welcome ▼ |  ▼


Gemma-7b-It: Available

Llama2-70b-4096: Available

✓ Mixtral-8x7b-32768: Available

555.13 T/s


What is generative AI?

 MODIFY ▼

Generative AI is a subset of artificial intelligence that uses machine learning models to generate new data similar to the data it was trained on. It can create various types of content, such as text, images, audio, and video.

There are several types of generative models, including:

1. Generative Adversarial Networks (GANs): GANs consist of two neural networks - a generator and a discriminator - that are trained together. The generator creates new data, while the discriminator evaluates the generated data and provides feedback to the generator. Over time, the generator becomes better at producing data that can fool the discriminator.
2. Variational Autoencoders (VAEs): VAEs are a type of autoencoder that can generate new data by learning a probabilistic distribution of the input data. They consist of an encoder that maps the input data to a latent space, and a decoder that maps the latent space back to the input space.
3. Transformers: Transformers are a type of neural network architecture that has been successful in natural language processing tasks such as machine translation and text generation. They can be


 LPU INFERENCE ENGINE

TOTAL REQUESTS: 35,625,820

© Groq Inc. 2024
All rights reserved

[API Access](#)
[Why Groq](#)
[About Us](#)
[GroqLabs](#)
[GroqCloud](#)

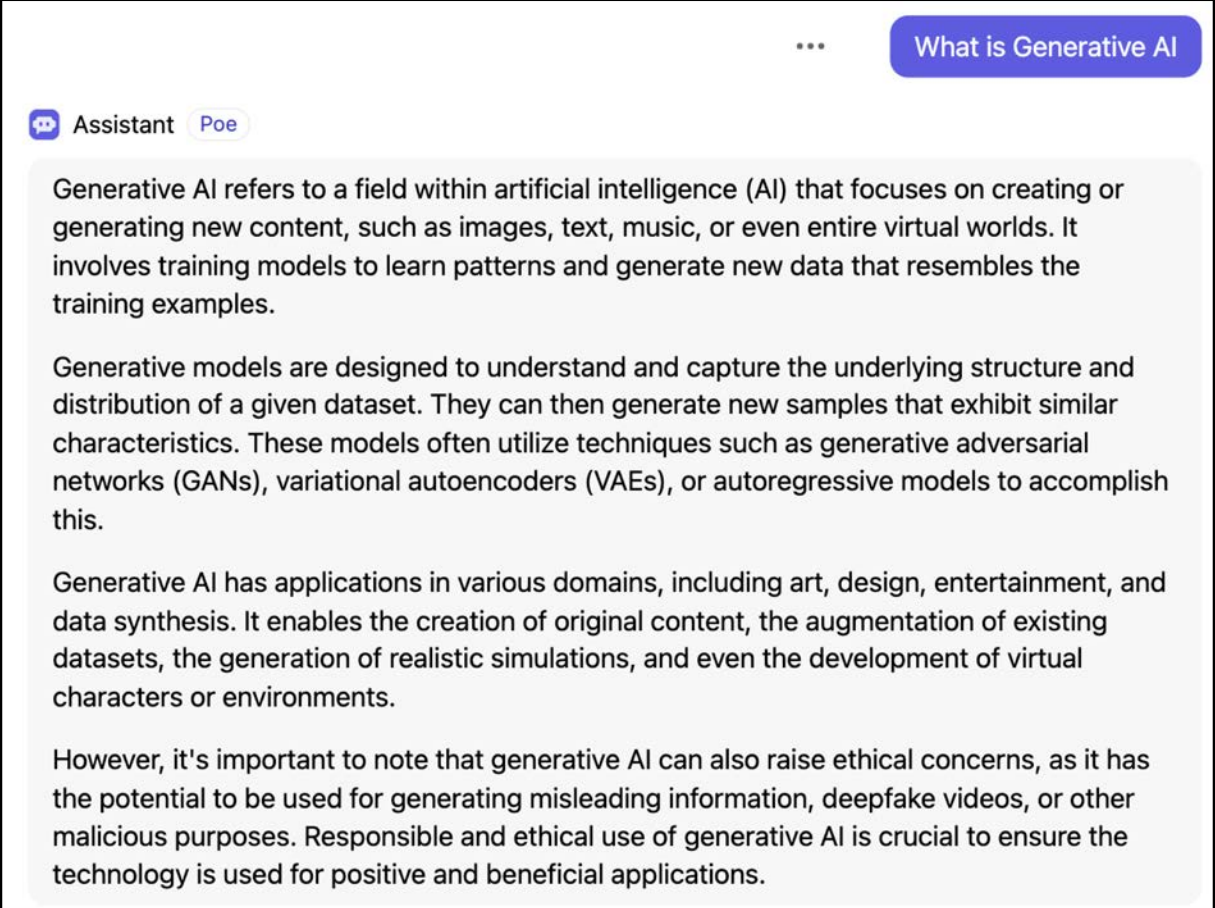
[Careers](#)
[Contact Us](#)
[Privacy Policy](#)
[Terms & Conditions](#)
[Security](#)
[Support](#)



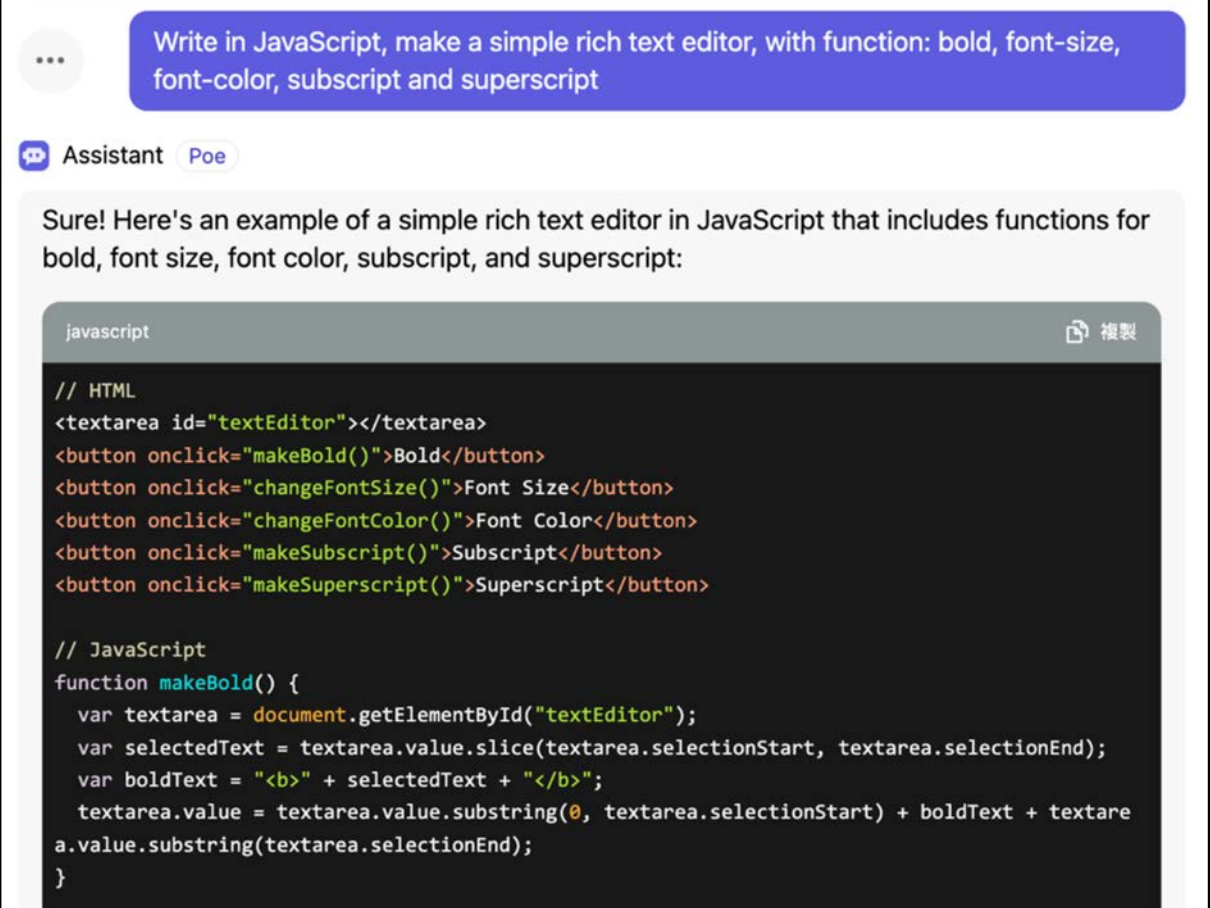
This is a Mistral AI-based chatbot. Accuracy, correctness or appropriateness cannot be guaranteed. | alpha release v0.9.24

You may try them on <https://groq.com>

The Power of Generative AI



Text generation and summarization



Computer programming, writing, and debugging code



Chatbots that can understand natural language and have personalities



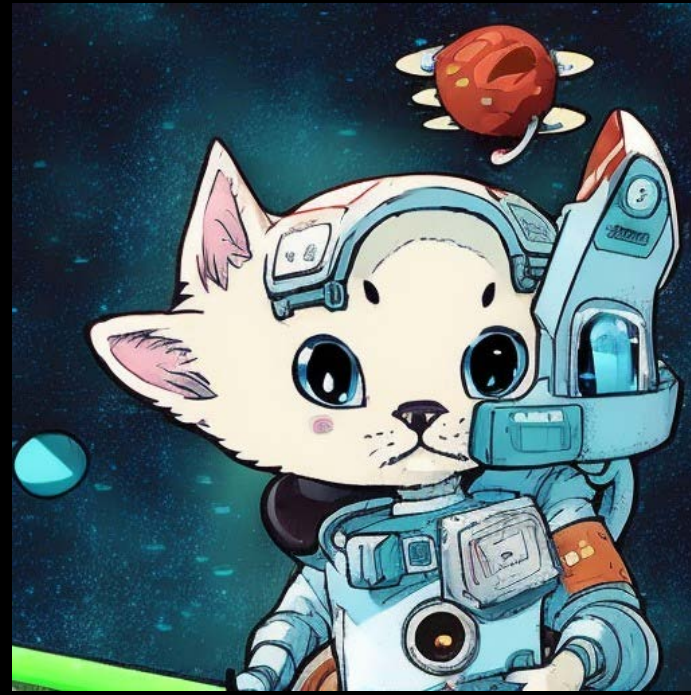
Generative design and artwork

AI image generation

Image generation



Text-to-Image



A drawing cat astronaut push random button on a spaceship, professional, UHD, HD, highly detailed, Eiichiro Oda styl



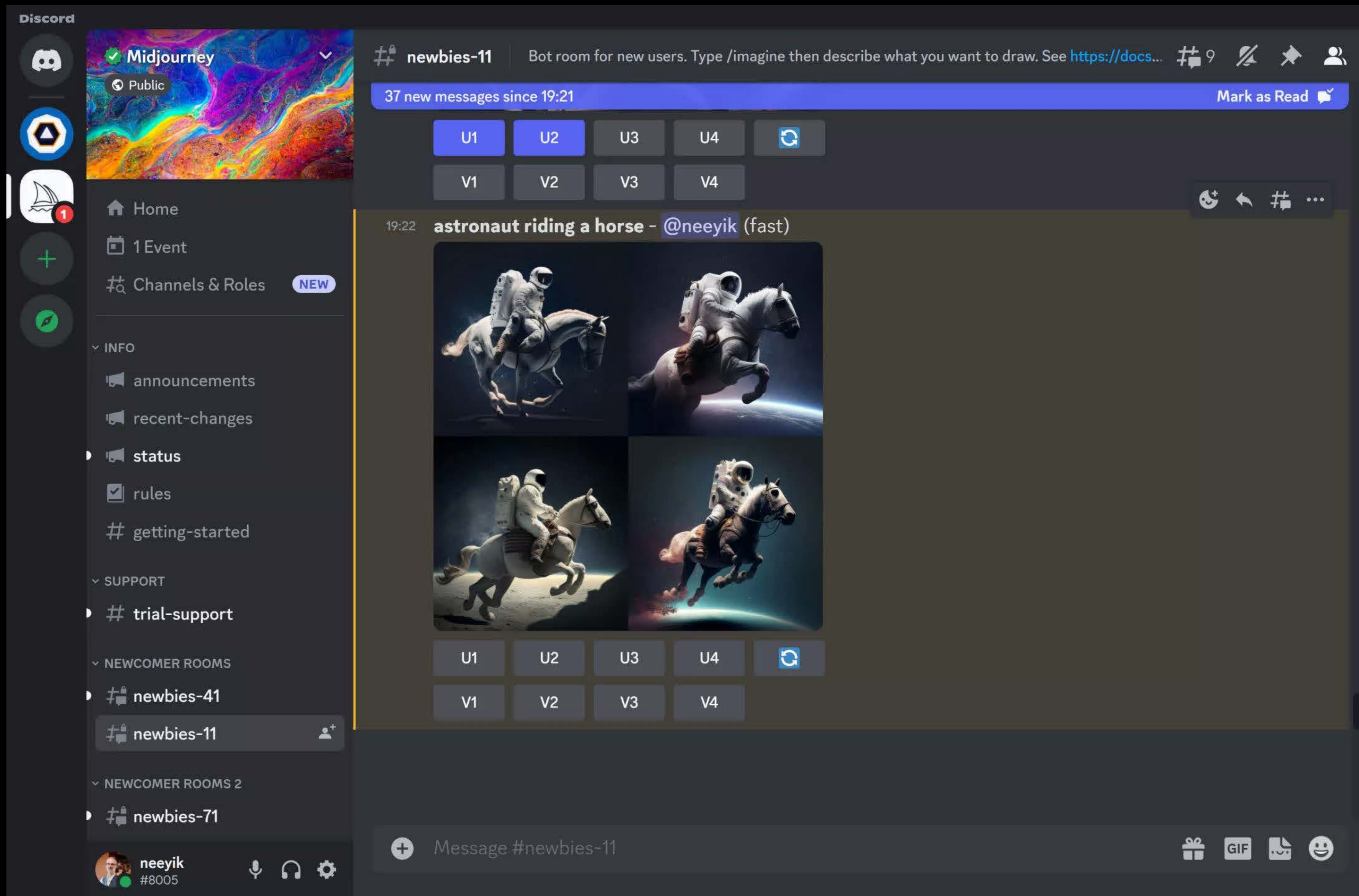
Astronaut looking at a nebula , digital art , trending on artstation , hyperdetailed , matte painting , CGSociety



An hourglass on a wooden table, with sand flowing from the top chamber to the bottom chamber. As the last grain of sand falls, the hourglass flips automatically, starting the process again. This image symbolizes the cyclic nature of beginnings and endings, and the continuous flow of time.

AI image generation

Midjourney





The screenshot shows a Discord chat window for a server named "Midjourney". The chat is in a "newbies-11" bot room. A message from user "@neeyik" at 19:22 contains the prompt "astronaut riding a horse - @neeyik (fast)". Below the prompt are four AI-generated images arranged in a 2x2 grid. Each image shows an astronaut in a white suit riding a white horse against a dark, starry background. The images are variations of the same prompt. Below the images are buttons labeled U1, U2, U3, U4, V1, and V4, which are used for upscaling and variations. The chat interface also shows a sidebar with server navigation options and a message input field at the bottom.

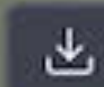


- Create by Midjourney
- Text to image
- Access only on Discord
- Cloud-base

AI image generation

DALL·E

 Can you make a stock photo of an Irish man working remotely?

 **DALL·E 3**
Created 4 images 📄 👍 🗑️

- Create by OpenAI
- Text to image
- Run on ChatGPT
- Cloud-base

AI image generation

Dreamstudio.ai

DreamStudio
by stability.ai

Shortcuts ⌘ K 2.38K

Generate Edit

Style
Choose style >





✓ Prompt
Pen and ink drawing of a mystical underwater world, with schools of fish, coral reefs, and a mermaid, highly detailed linework, high-contrast, stylized

> Negative prompt

> Upload image

Settings
1:1
Image count 4
Advanced

Dream 0.91



Text-to-image

DreamStudio
by stability.ai

Shortcuts ⌘ K 2.38K

Generate Edit

Style
Choose style >

✓ Prompt
Baked salmon fillet with a perfectly crispy skin and tender, flaky flesh, served with a side of steamed vegetables and quinoa, healthy, flavorful, high detail, food photography

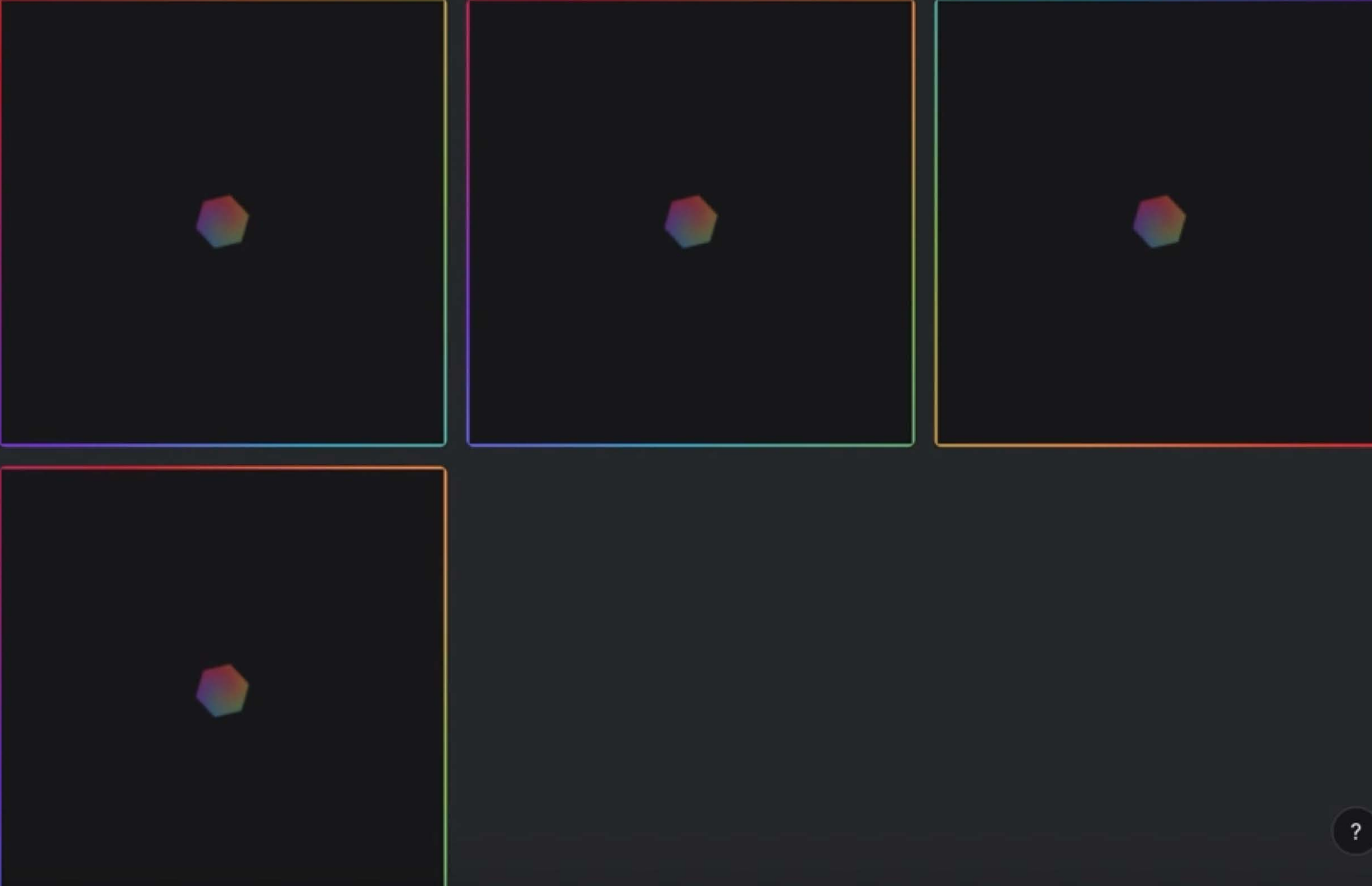
> Negative prompt

> Upload image

Settings
1:1
Image count 4
Advanced

Dream 0.91

Baked salmon fillet with a perfectly crispy skin and tender, flaky flesh, served with a side of ...



The interface shows a text-to-image generation tool. The prompt is "Baked salmon fillet with a perfectly crispy skin and tender, flaky flesh, served with a side of steamed vegetables and quinoa, healthy, flavorful, high detail, food photography". The settings include a 1:1 aspect ratio and an image count of 4. The generated images are displayed in a grid, each with a different colored border (red, green, blue, yellow).

AI image generation



Image-to-image Outpainting

DreamStudio
by stability.ai

Shortcuts ⌘ K 2.37K

Generate Edit

Add a dream +

Upload ↑

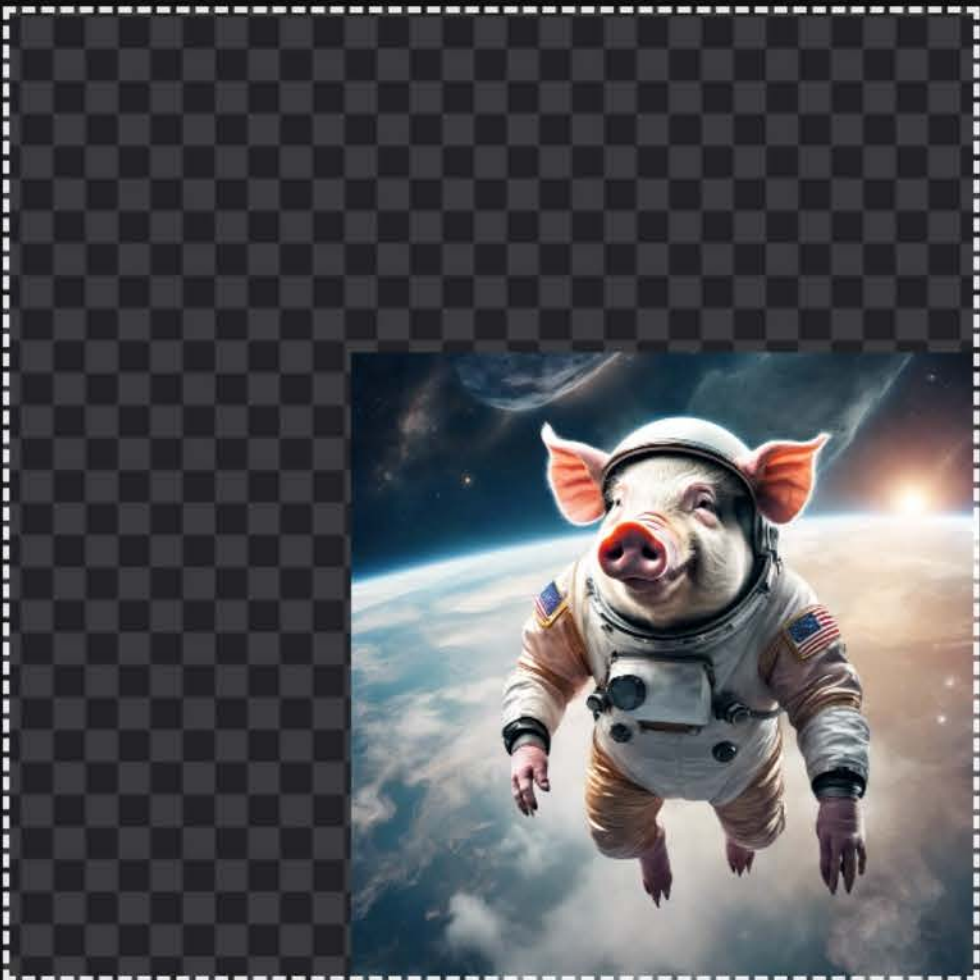
Settings

Snapping

Auto flatten

Background fill # 121212

Smart export



Beautiful Universe, an alien in UFO is looking at t...

Abandoned prison on a cliff, with a stormy sea b...

↑

pig1.png

Tip
Use the V key to switch to the select tool

?

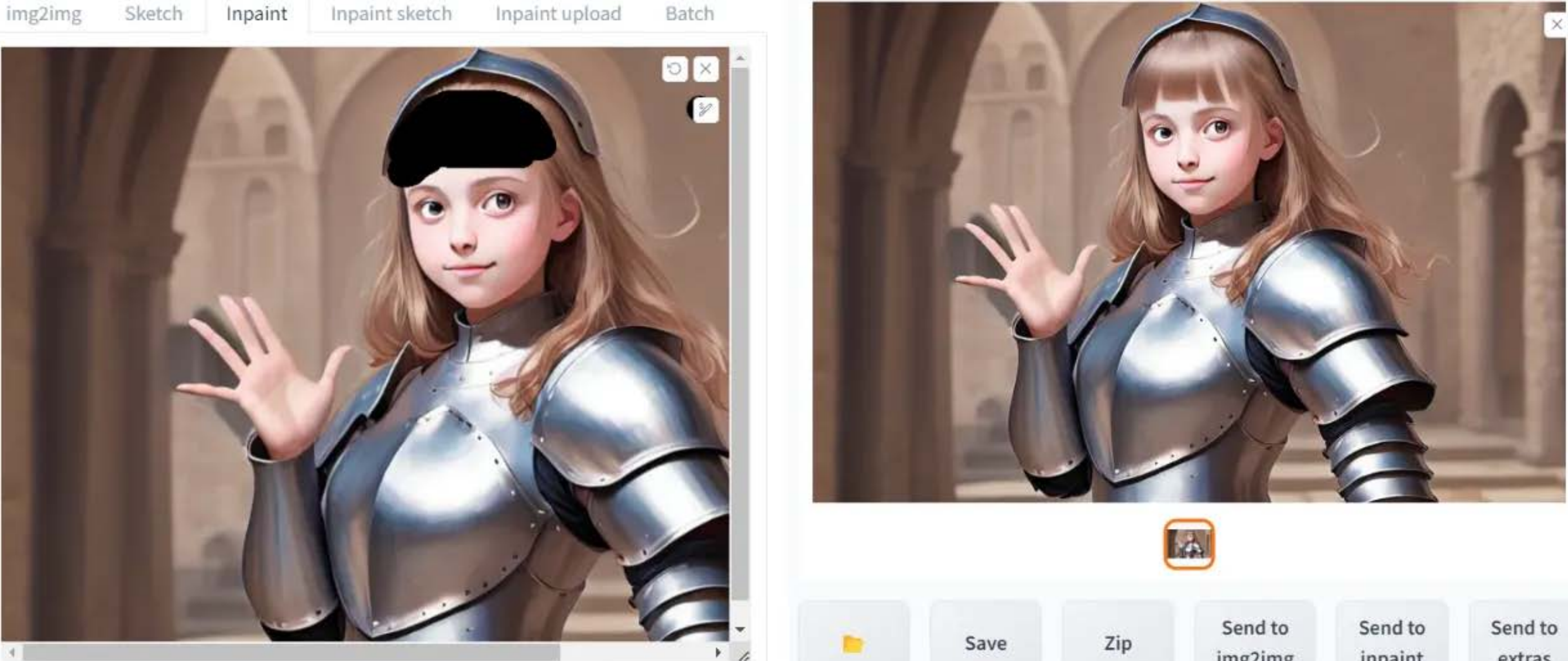
Inpaint

Inpaint



Inpaint

img2img Sketch **Inpaint** Inpaint sketch Inpaint upload Batch



Copy image to:

beautiful bangs hair
Steps: 20, Sampler: Euler a, CFG scale: 7, Seed: 121489600, Size: 768x512, Model hash: 2336dbf342,
Model: dreamshaper_631BakedVae, Denoising strength: 0.75, Clip skip: 2, Mask blur: 7, Noise multiplier:

AI image generation

Stable Diffusion

Stable Diffusion checkpoint
custommerge_0.5-weighted.ckpt [1e98c454]

txt2img | **img2img** | Extras | PNG Info | Checkpoint Merger | Train | Settings

realistic, young, (((1girl))), dynamic pose, (business suit:1.25), (skirt suit:0.8), (raised eyebrows:1.2), (flustered:1.2) (full face blush:1.2), (upturned eyes:1.2), (emerald eyes:1.2), (looking up:1.2), looking at viewer, (detailed face+eyes:1.2), (blonde hair:1.3), (western comic style), (detailed corporate office background:1.2), by (Alphonse Mucha:0.8) digital painting, (skirt lift:1.2), sunlight, window, (smooth body:1.3), (age 20:1.3), cinematic lighting

low resolution, bad quality, blurry, preview image, preview, over exposure, (depth of field:1.3), dof, illustration, painting, implants, silicone, (((anime))), (((long torso))), (((tall))), (((mutated hands and fingers))), (((deformed))), (((portrait))), malformed face, preview, thumbnail, sample, mutation, mutated, closeup, (ugly), animal, text, (((bad anatomy))), (((deformed))), ((poorly drawn face)), grayscale, old, long neck, (((big head))), two heads, (((boy))), (((amputated))), (((big eyes)))

81/150

Generate

Style 1: None | Style 2: None

Sampling Steps: 30

Sampling method: Euler a Euler LMS Heun DPM2 DPM2 a DPM fast DPM adaptive LMS Karras DPM2 Karras DPM2 a Karras DDIM PLMS

Width: 768 | Height: 1024


Restore faces Tiling Highres. fix

Firstpass width: 0 | Firstpass height: 0 | Denoising strength: [slider]

Batch count: 2 | Batch size: 3

CFG Scale: 14

Seed: [input]



Save | Send to img2img | Send to inpaint | Send to extras

Make Zip when Save?

realistic, young, (((1girl))), dynamic pose, (business suit:1.25), (skirt suit:0.8), (raised eyebrows:1.2), (flustered:1.2) (full face blush:1.2), (upturned eyes:1.2), (emerald eyes:1.2), (looking up:1.2), looking at viewer, (detailed face+eyes:1.2), (blonde

AI image generation

One of the github repository for stable Diffusion

AUTOMATIC1111 / **stable-diffusion-webui** Public

Notifications Fork 25k Star 129k

<> Code Issues 2k Pull requests 11 Discussions Actions Projects Wiki Security Insights

master 14 Branches 29 Tags Go to file Code

AUTOMATIC1111 Merge branch 'release_candidate' ✓ adadb4e · 3 days ago 7,318 Commits

.github	update ruff to 0.3.3	last month
configs	support for sdxl-inpaint model	4 months ago
embeddings	add embeddings dir	2 years ago
extensions-builtin	Merge pull request #15319 from catboxanon/feat/ssmd_c...	3 weeks ago
extensions	delete the submodule dir (why do you keep doing this)	2 years ago
html	Merge branch 'dev' into extra-networks-buttons	last month
javascript	Merge branch 'dev' into extra-networks-buttons	last month
localizations	Remove old localizations from the main repo.	2 years ago
models	Add support for the Variations models (unclip-h and uncli...	last year
modules	Merge pull request #15492 from w-e-w/update-restricted...	4 days ago
scripts	fix upscaler 2 images do not match	2 weeks ago

About

Stable Diffusion web UI

web ai deep-learning torch pytorch unstable image-generation gradio diffusion upscaling text2image image2image img2img ai-art txt2img stable-diffusion

Readme AGPL-3.0 license Cite this repository Activity 129k stars 1k watching 25k forks Report repository

Releases 22

<https://github.com/AUTOMATIC1111/stable-diffusion-webui>

AI image generation

Base Model

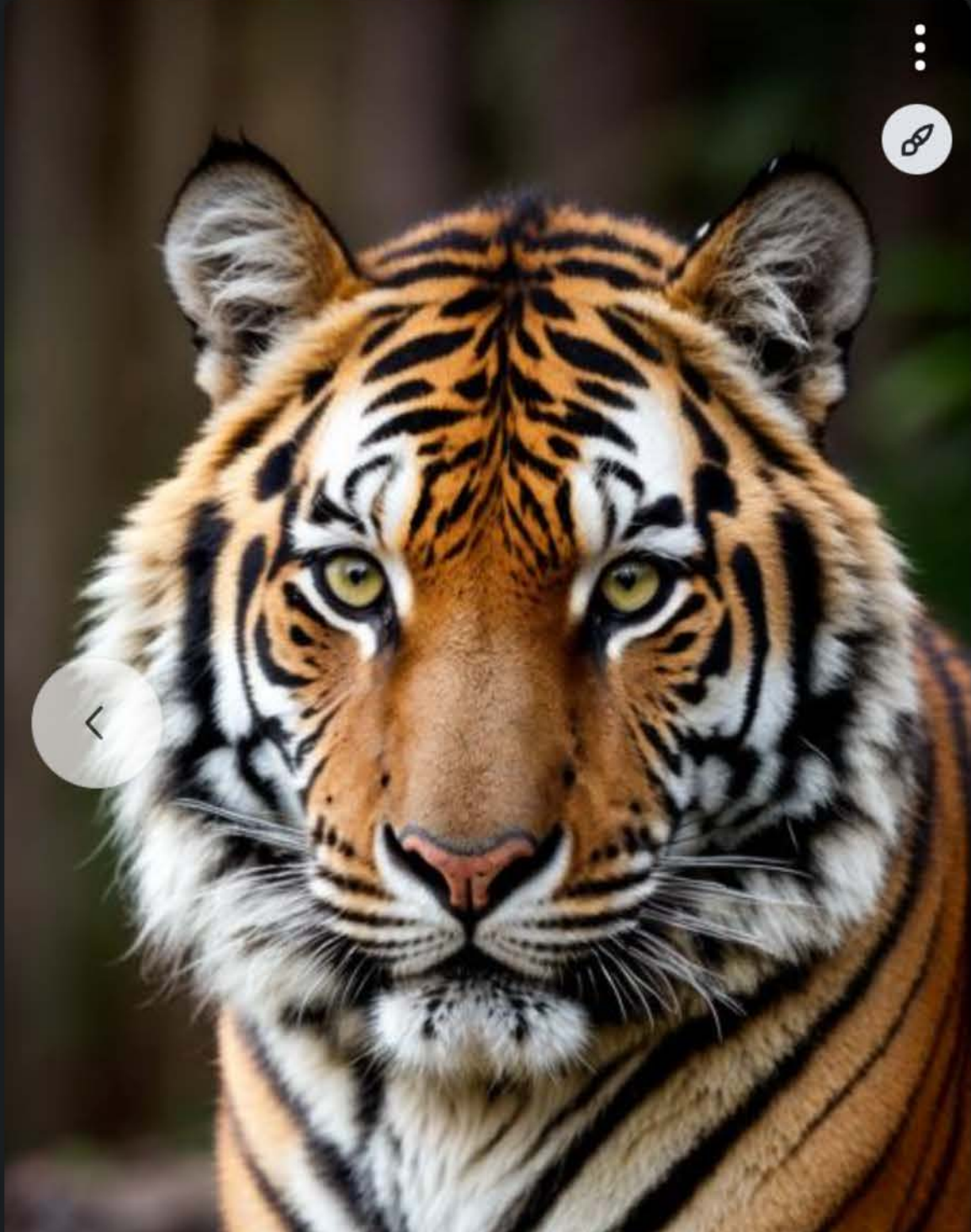
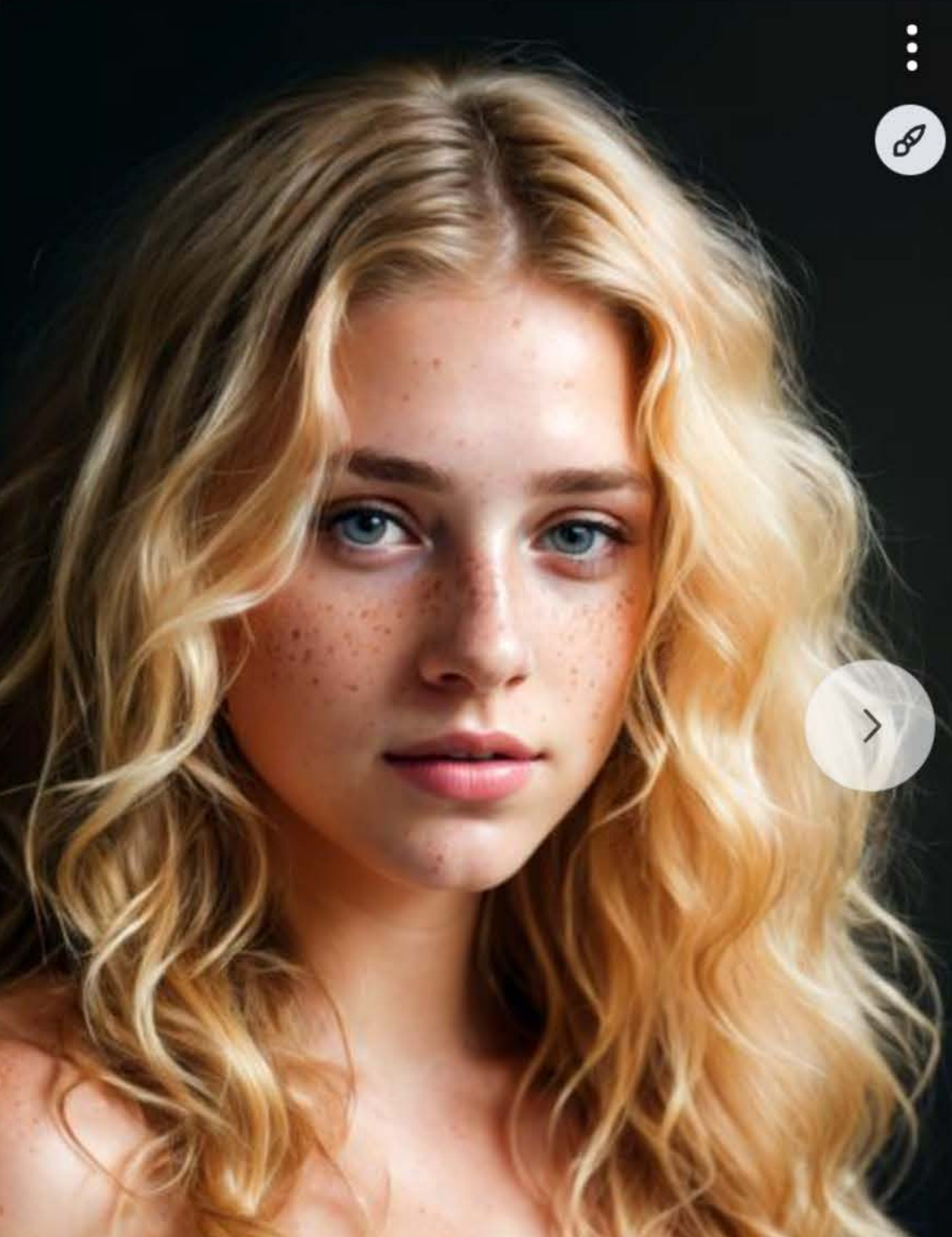
CIVITAI [Do It >](#) Models / Create Sign In

[Home](#) [Models](#) [Images](#) [Videos](#) [Posts](#) [Articles](#) [Bounties](#) [Events](#) [Builds](#)

CyberRealistic Revamp

Updated: Apr 5, 2024 **BASE MODEL** **PHOTOREALISTIC** **WOMAN** **MAN** **REALISTIC**

v3.2


[Download \(1.99 GB\)](#) [Play](#) [Share](#) [Like](#)

Verified: 10 days ago SafeTensor

Details	
Type	CHECKPOINT MERGE
Stats	836
Reviews	Very Positive (115)
Uploaded	Apr 5, 2024
Base Model	SD 1.5
Hash	AUTOV2 FE9B443031

2 Files

About this version

 **Cyberdelia** Joined Dec 7, 2022 [Tip](#) [Follow](#)

AI image generation

LORA

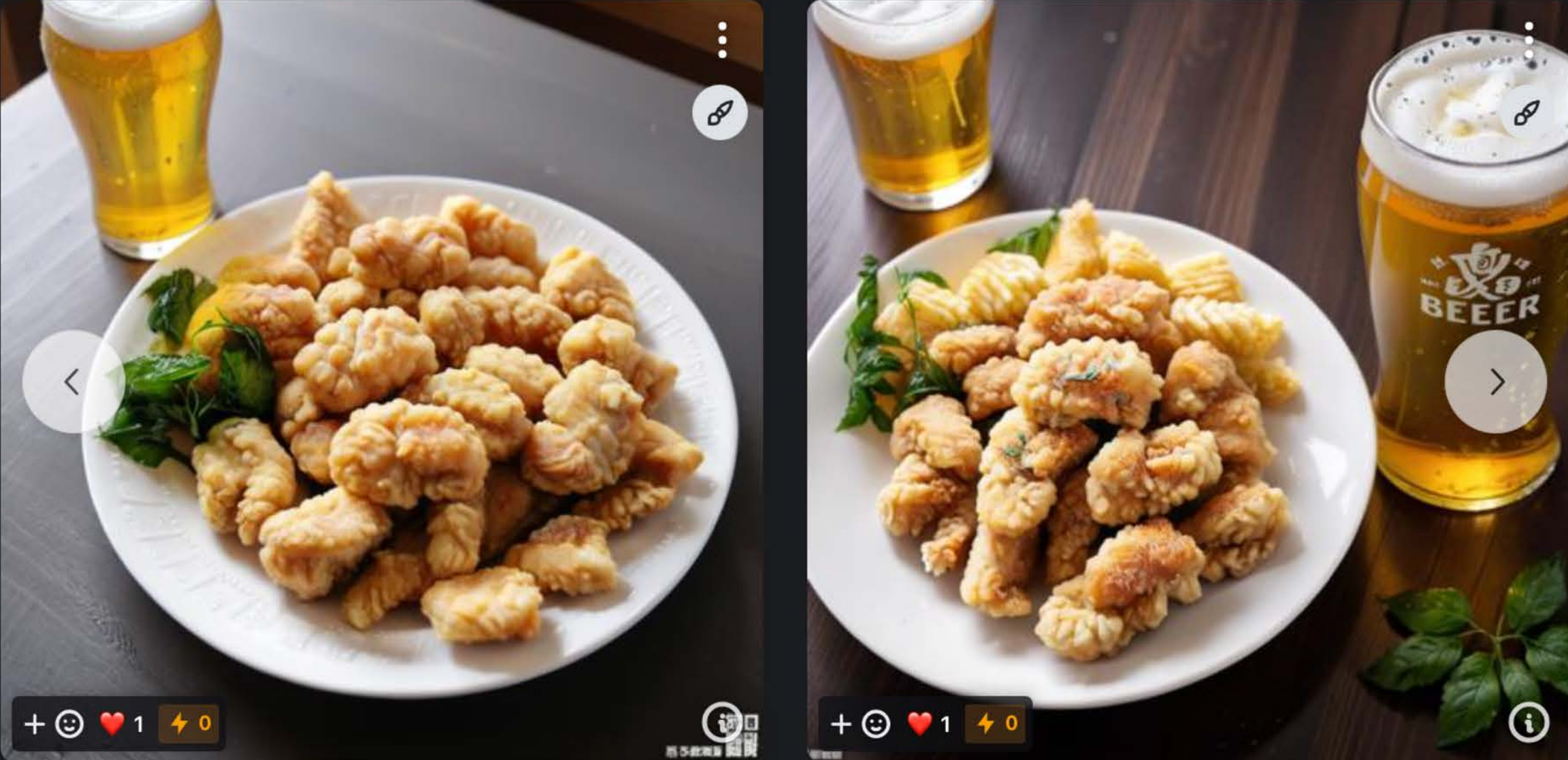
CIVITAI [Do It >](#) Models / Create Sign In

[Home](#) [Models](#) [Images](#) [Videos](#) [Posts](#) [Articles](#) [Bounties](#) [Events](#) [Builds](#)

Taiwanese fried chicken | SDXL

Updated: Apr 15, 2024 OBJECTS FOOD 3 14 1 1 0

[v1.0](#)



Using this LORA, you can obtain an image of Taiwanese Salt and Pepper Chicken.

Salt and Pepper Chicken is a popular snack commonly found in Taiwanese street stalls and night markets. The

[Create](#) [Download](#) [Play](#) [Share](#) [Like](#)

Verified: 6 hours ago SafeTensor

Details

Type	LORA
Stats	14 1
Reviews	Positive (3)
Uploaded	Apr 15, 2024
Base Model	SDXL 1.0
Training	STEPS: 1,750 EPOCHS: 50
Trigger Words	TWGENSOUGE
Hash	AUTOV2 5B45775CDA

1 File

Common websites

<https://civitai.com/>

The screenshot displays the CivitAI website interface. At the top, the CivitAI logo is on the left, followed by a navigation menu with 'Models', 'Search Civitai', and a search icon. On the right, there are 'Create' and 'Sign In' buttons. Below the navigation is a horizontal menu with icons for Home, Models, Images, Videos, Posts, Articles, Bounties, Events, and Builds. The main content area features a 'Featured Images' section with a descriptive paragraph and a link to 'Explore all images'. Below this, a grid of AI-generated images is shown, each with a user name and engagement metrics (likes, hearts, comments, etc.).

CIVITAI 🗨️

Models Search Civitai / 🔍

Create Sign In

Home Models Images Videos Posts Articles Bounties Events Builds

Featured Images

All sorts of cool pictures created by our community, from simple shapes to detailed landscapes or human faces. A virtual canvas where you can unleash your creativity or get inspired.

[Explore all images →](#)

jhonny09 101 40 10 0

PrincessArt 414 177 36 12 100

Number4

DutchyDutch

kilobait3

DiceAiDevelopment

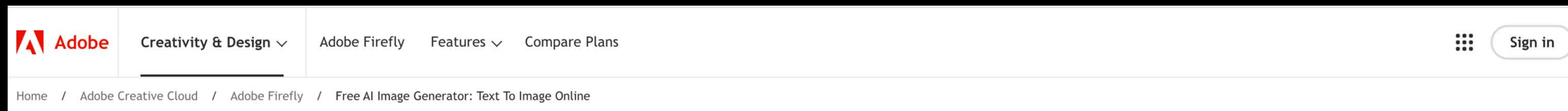
Common websites

<https://www.canva.com/ai-image-generator/>

The screenshot displays the Canva AI Image Generator website. At the top, the Canva logo is on the left, followed by navigation links: Design spotlight, Business, Education, Plans and pricing, and Learn. A search bar on the right contains the text "Search thousands of templates". Further right are "Log in" and "Sign up" buttons. Below the navigation, a breadcrumb trail shows "Home > Text to Image". The main heading is "Free Online AI Image Generator". Below this, a paragraph reads: "Dream it up, then add it to your design. Watch your words and phrases transform into beautiful images with the best AI image generator apps available at your fingertips. Stand out with an image perfect for your project." A purple "Generate AI Images" button is centered below the text. Below this is a preview of the Canva AI Image Generator interface. The preview shows the Canva logo, a "Share" button with an upload icon, and three user profile icons. The main area of the preview features a purple "Storyboard" banner with the Canva logo and the word "Storyboard". Below the banner is a "Try an example" input field with a plus sign, and a "Styles" section with a "See all" link. At the bottom of the preview, there is an "Add sandstorm" button. A second purple "Generate AI Images" button is centered below the preview.

Common websites

<https://www.adobe.com/products/firefly/features/text-to-image.html>



AI Image Generator: Create images from text.

Do you dream of seeing a long-haired dachshund with flowing rainbow hair? Or flowers growing out of concrete in a lost city? Whatever you imagine, if you can describe it, you can create it fast using the AI image generator in Adobe Firefly.

[Generate Image Now](#)

Try it

Dog in a sweater, primary colo...

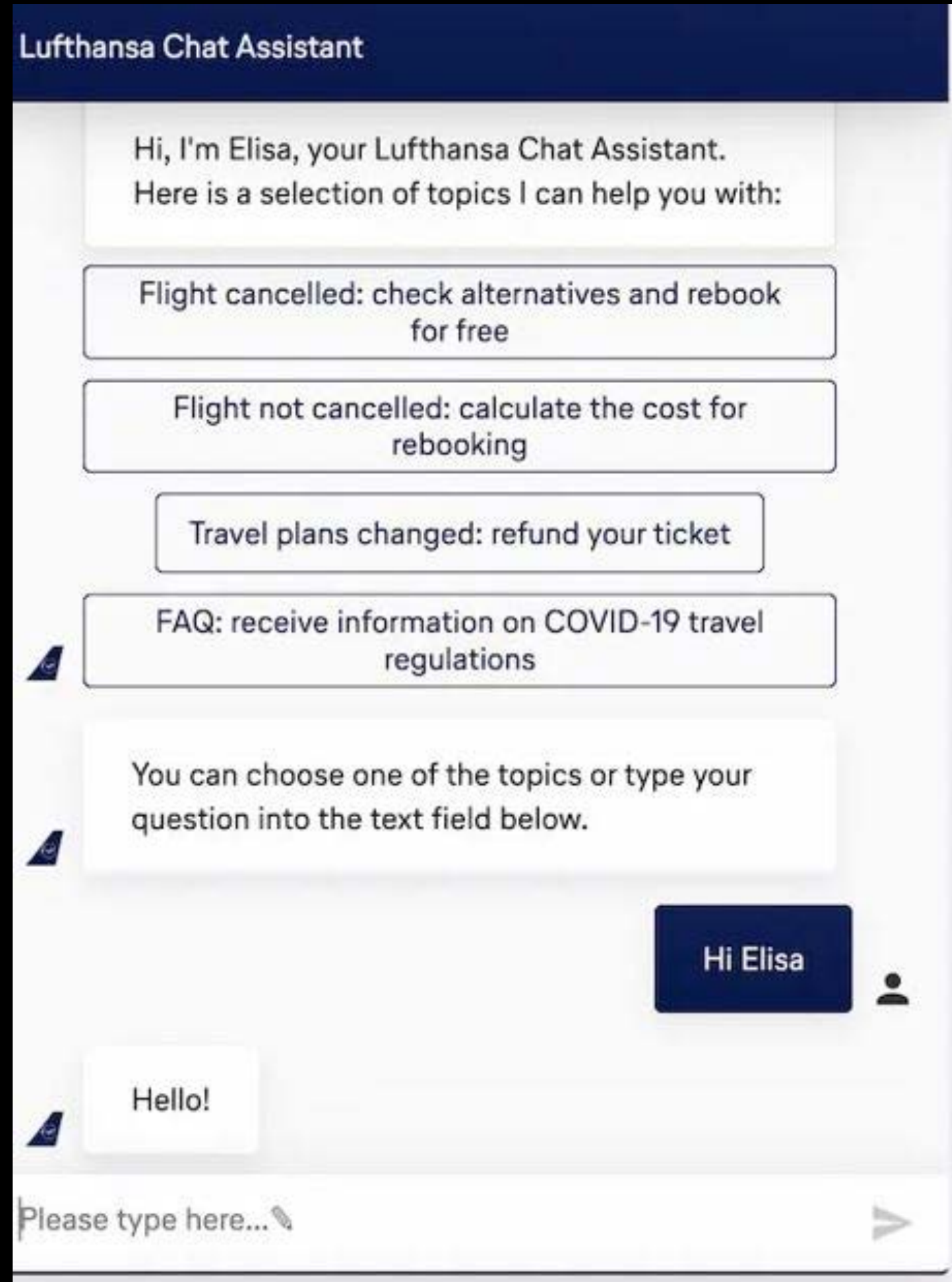
[Generate](#)



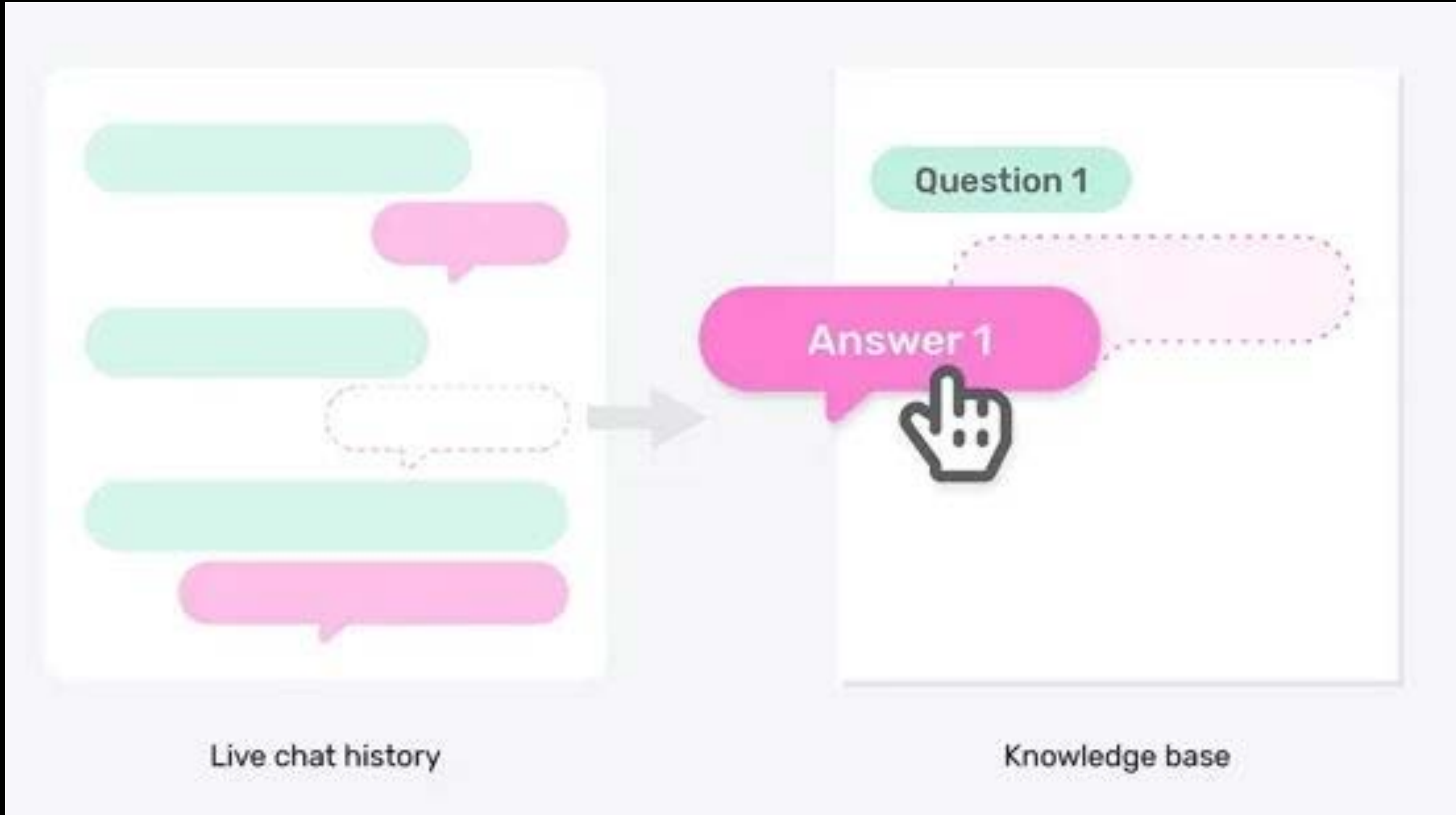
AI ChatBots

Traditional Chatbots

How A Rule-Based Chatbot Works

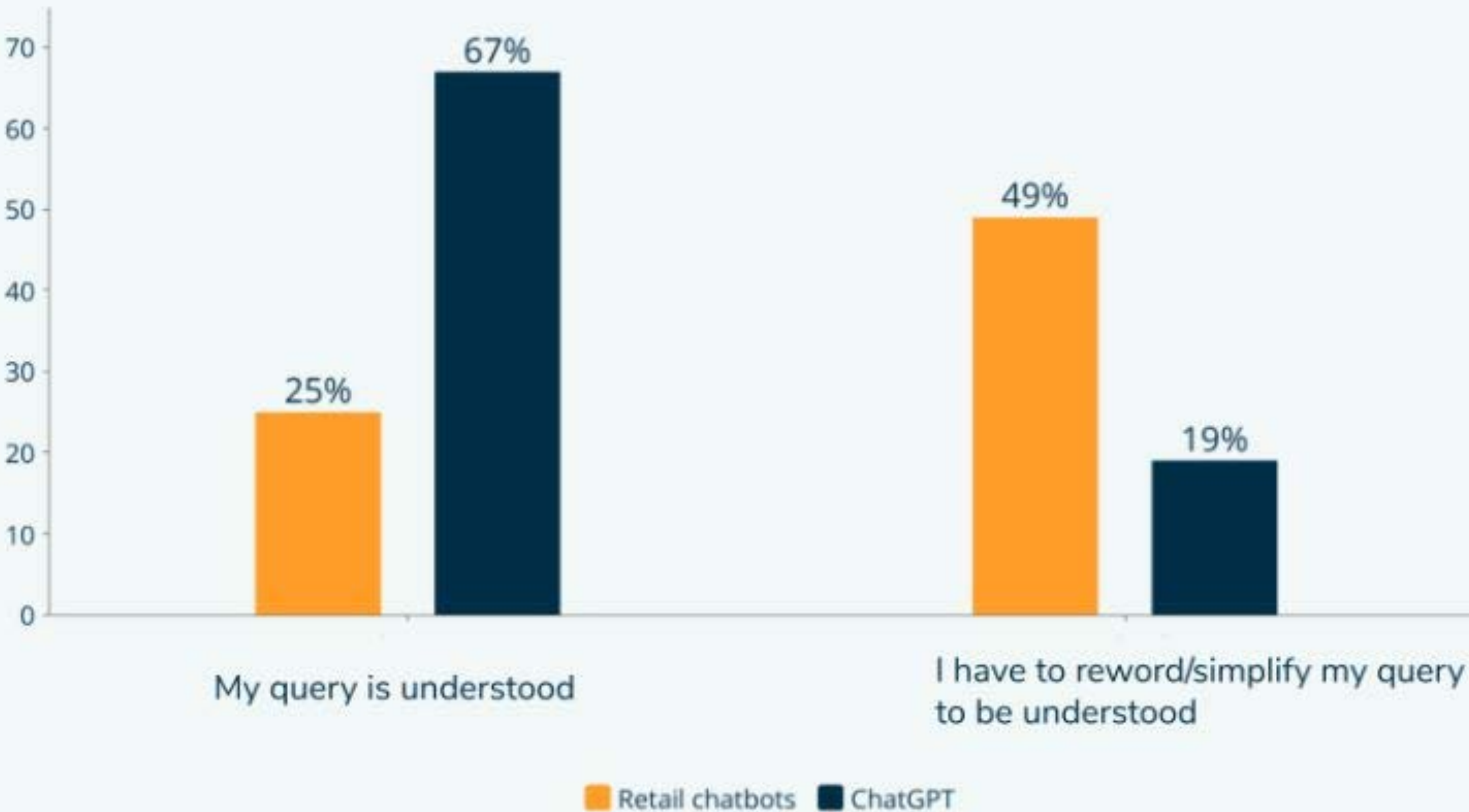


Generative AI Chatbots



Generative AI Chatbots

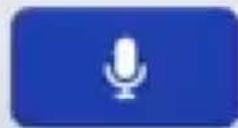
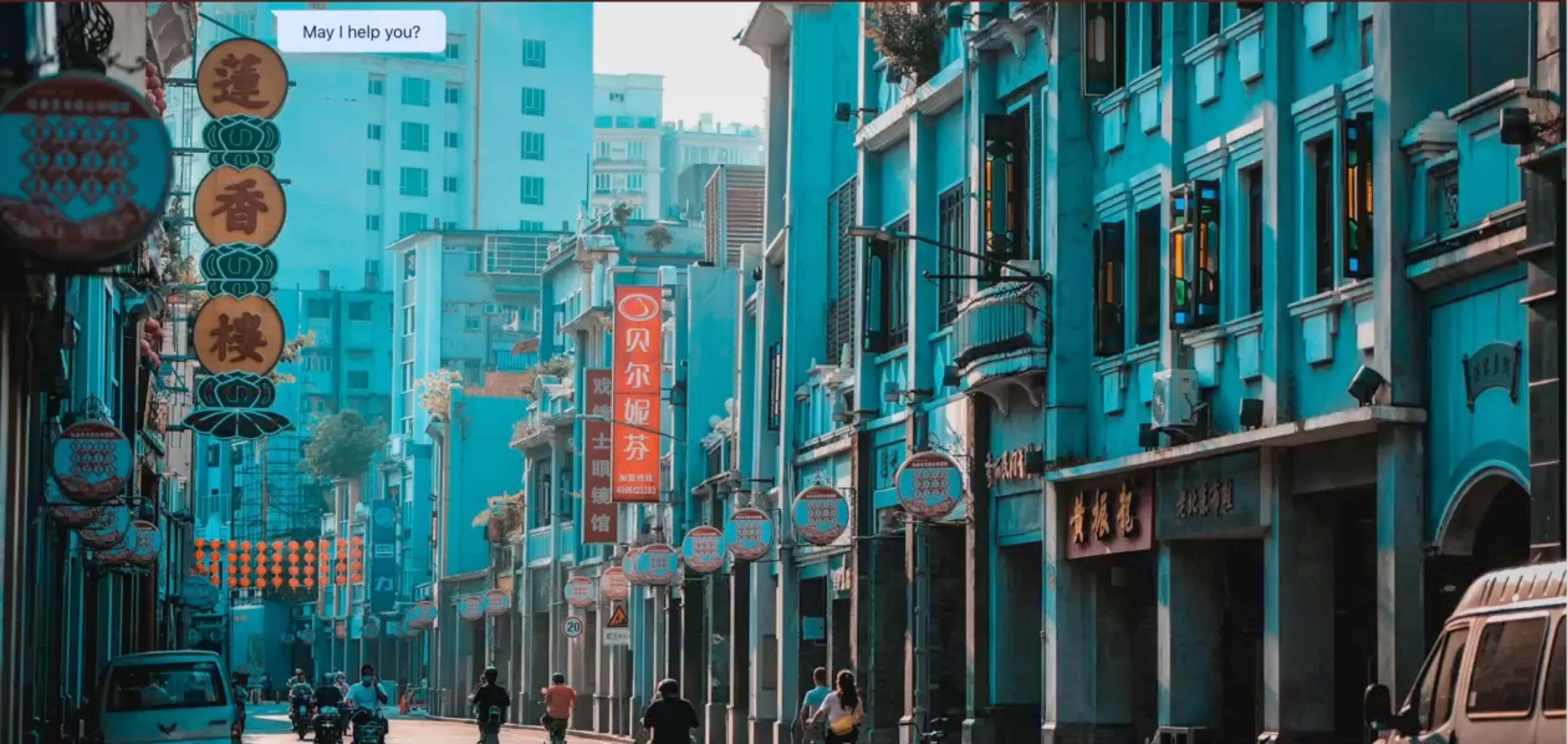
ChatGPT understands consumers better than traditional retail chatbots



Generative AI Chatbots

Item	Traditional Chatbots	AI Chatbots
Language Processing	Keywords based	NLP & machine learning
Database Size	Smaller	Massive
Scalability	Limited scalability	Scalable enough
Personalization	Limited	High

May I help you?



Common website

<https://www.simprobot.com/>

The screenshot displays the SimProBot website homepage. At the top, a blue navigation bar contains the SimProBot logo on the left and a menu of links: Home, About SimProBot, STEM, Bot Types, Use Cases, Contact, and Login. A language dropdown menu is set to 'English'. The main content area features the SimProBot logo and the tagline 'Simple setup, professional chatbots'. A prominent blue button labeled 'Request a Demo' is positioned below the tagline. To the right, four green speech bubbles illustrate chatbot interactions with user prompts such as 'What are the school admission process and deadlines?', 'Recommend some books in this library for organic chemistry, college level.', 'Help me make a travel plan for a 4 day trip to Paris.', and 'How many patients were involved in this clinical study?'. At the bottom, the text 'Create and deploy your' is visible, followed by a blue square button with an upward-pointing arrow.



Home

About SimProBot

STEM

Bot Types

Use Cases

Contact

Login

English

SimProBot

Simple setup, professional chatbots

Request a Demo

You

What are the school admission process and deadlines?

You

Recommend some books in this library for organic chemistry, college level.

You

Help me make a travel plan for a 4 day trip to Paris.

You

How many patients were involved in this clinical study?

Create and deploy your



Common website

<https://www.hubspot.com/>



+85 2 3011 4980

Get a Free Demo of HubSpot's Customer Platform

HubSpot's customer platform offers enterprise software for marketing, sales, customer service, content management, operations, and commerce.

Our platform was built from the ground up to be a powerful, integrated, and easy-to-use system that leads to better customer experiences and business growth.

First Name *

Last Name *

Email *

Phone *

Company Name *

Website *

How many employees work there? *

Subscribe to HubSpot's marketing blog

HubSpot uses the information you provide to us to contact you about our relevant content, products, and services. You may unsubscribe from these communications at any time. For more information, check out our [privacy policy](#).

Write a message



We're committed to your privacy. HubSpot uses the information you provide to us to contact you about our relevant content, products, and services. You may unsubscribe from these communications at any time. For more information, check out our [Privacy Policy](#).



Common website

<https://www.zendesk.com/>

zendesk

Demo



"There is absolutely no way we could handle the volume and variety of transactions that come through without a tool like Zendesk."

Melissa Shuter

Executive Director of Operations Support Services

200%+

Team growth since implementing Zendesk

15%

Improvement in SLA adherence

ZENDESK FOR SMALL BUSINESS

Empowering schools in the digital age

Zendesk is user-friendly support that provides seamless communication for enhanced productivity.

Work email

Start your free trial

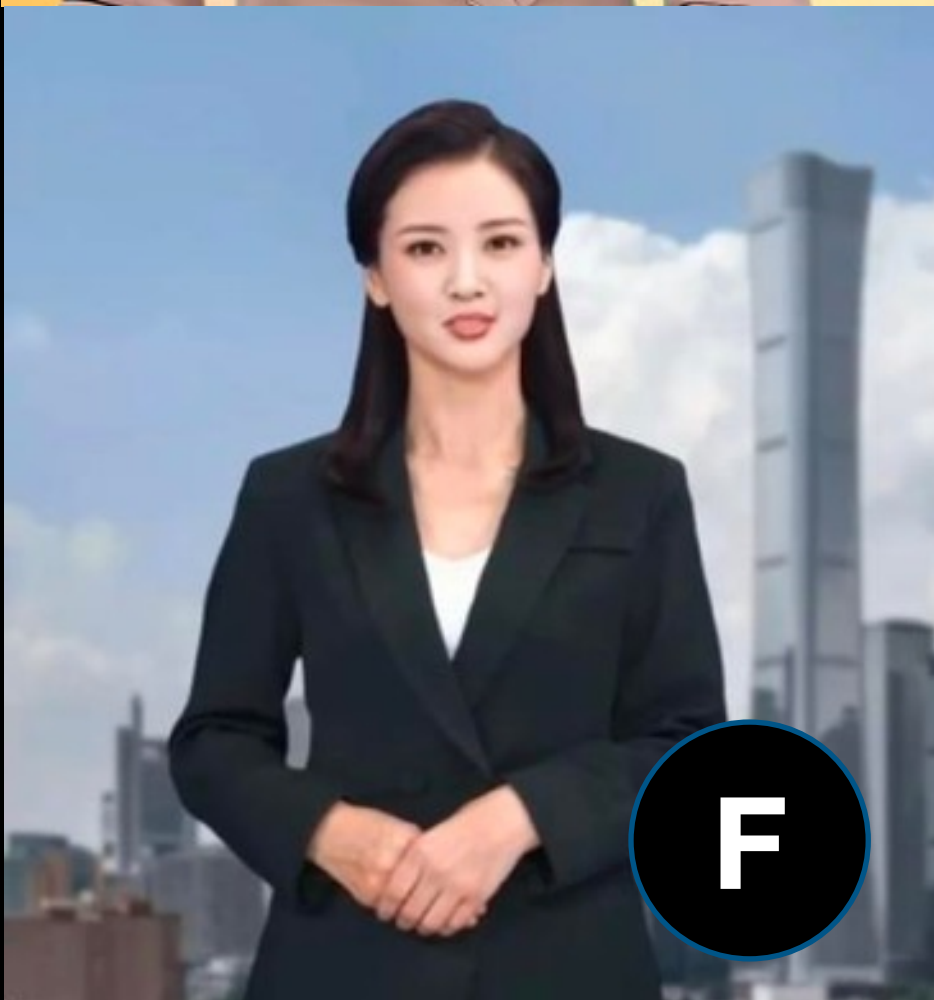
✓ No credit card required

✓ Full access to our cloud software

✓ Cancel at any time

AI Anchors

Who are AI Anchors?



AI Anchor Example



<https://www.youtube.com/watch?v=bciKYfXKE7U>

AI Anchor Example



<https://www.youtube.com/watch?v=YSGS3gvOabl>

AI Anchor Example



<https://www.youtube.com/watch?v=C6CTQgQ04jE>

Technology used in AI Anchors

- Face recognition and facial feature extraction
- Expression and motion capture
- Virtual image generation
- Speech synthesis
- Machine learning and deep learning



AI Anchors Advantage



24/7

AI virtual anchors are not subject to time limitations and do not require rest or holidays. This allows viewers to access news and entertainment content at any time.



Multiple language

AI virtual anchors can support multiple languages without the need for translation. This allows virtual anchors to serve a global audience across different countries without language barriers.



Cost-effectiveness

Compared to human anchors, the operational costs of AI virtual anchors are typically lower. Virtual anchors do not require salaries, insurance, or other benefits.

AI Anchor Example

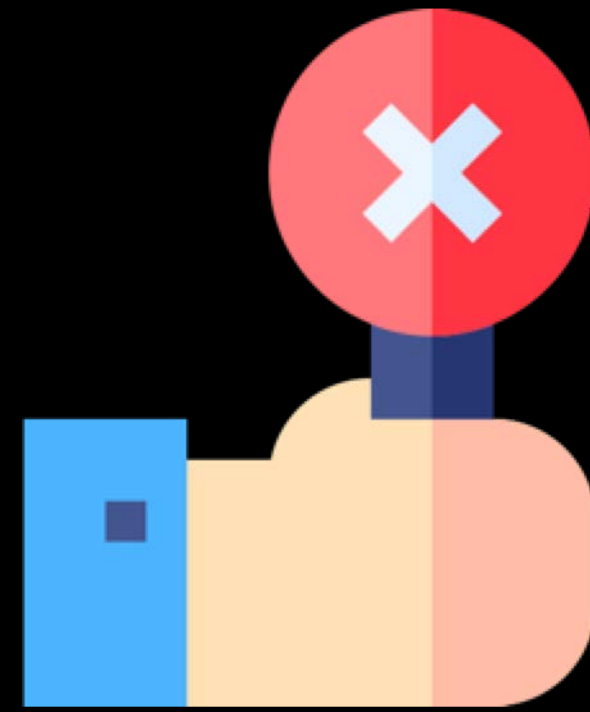


A.I. Anchor vs Live-action shooting



Simpler production

Live-action shooting requires pre-production planning, scene design, organizing actors and crew, as well as post-production and editing



Reduced mistakes

Different personnel involved in live-action shooting, including anchors, have the potential to make errors.

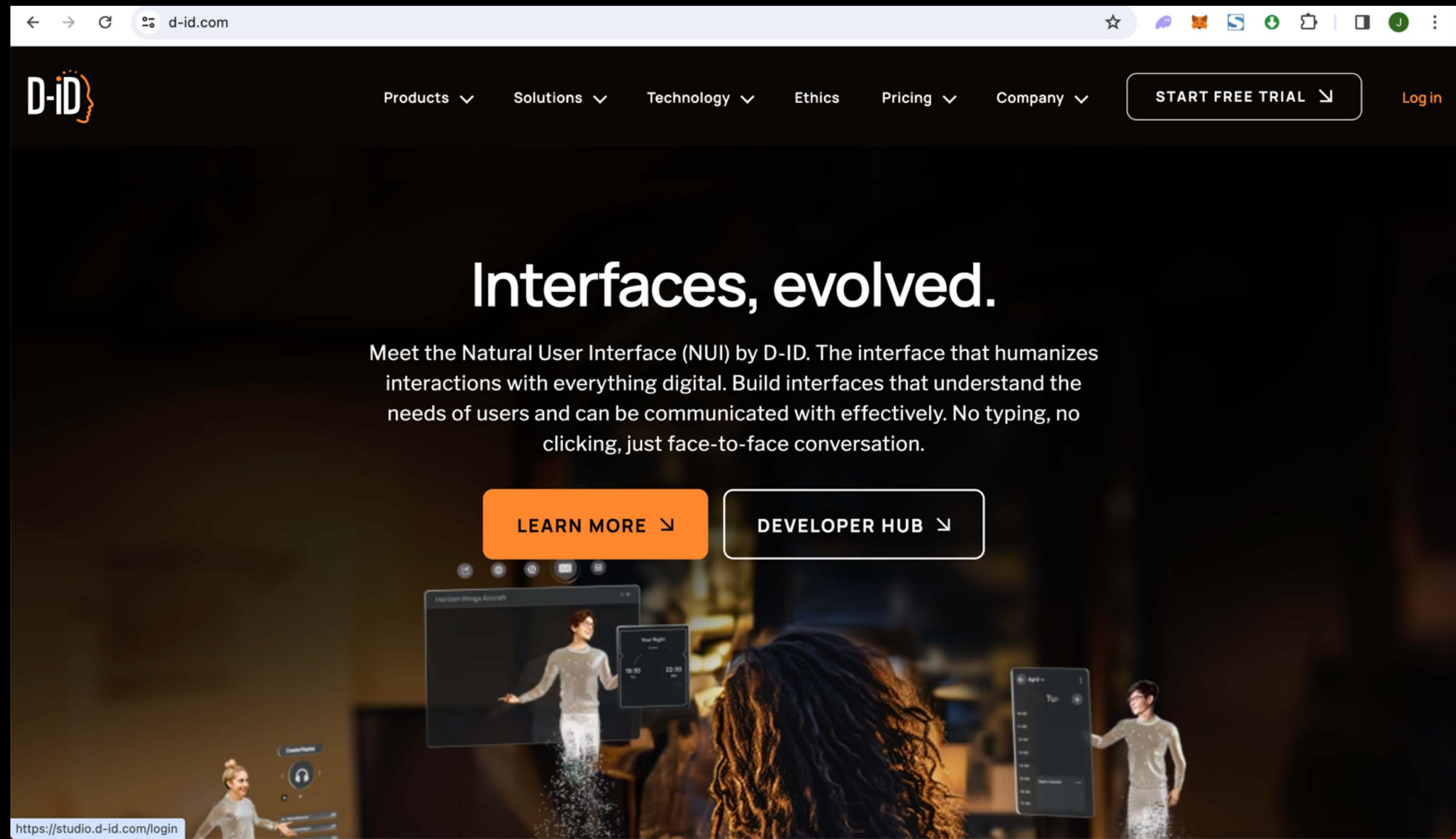


Faster production

Creating AI virtual anchor videos allows for the reuse of a large amount of pre-prepared material and the incorporation of the latest content to complete the production.

Common websites

<https://www.d-id.com/>



The screenshot shows the homepage of D-ID. At the top left is the D-ID logo. The navigation menu includes: Products, Solutions, Technology, Ethics, Pricing, and Company. On the right side of the navigation bar are a "START FREE TRIAL" button and a "Log in" link. The main content area features the headline "Interfaces, evolved." followed by a paragraph: "Meet the Natural User Interface (NUI) by D-ID. The interface that humanizes interactions with everything digital. Build interfaces that understand the needs of users and can be communicated with effectively. No typing, no clicking, just face-to-face conversation." Below this text are two buttons: "LEARN MORE" and "DEVELOPER HUB". The background image depicts a person's head in the foreground, looking at a digital interface that shows a flight schedule for "Harrison Wings Aircraft" with a "Your flight" section listing times from 18:50 to 22:50. Another person is visible in the background interacting with a similar interface.

<https://studio.d-id.com/login>

Common websites

<https://www.heygen.com/>

The image shows the homepage of the HeyGen website. At the top left is the HeyGen logo. The navigation menu includes 'Use Cases', 'Features', 'Resources', 'Company', and 'Pricing', each with a dropdown arrow. On the top right, there is a 'Contact Sales' link and a prominent blue 'Get started' button. The main heading reads 'AI-powered video creation at scale'. Below this, a sub-headline states 'Effortlessly produce studio-quality videos with AI-generated avatars and voices.' A central blue button says 'Get started for free' with a right-pointing arrow, and below it, the text 'No credit card needed' is displayed. A row of logos for partner companies follows, including Microsoft, Adobe, Amazon, Columbia University, Keller Williams, NVIDIA, Pattern, Salesforce, and Volvo. At the bottom, a video player is partially visible with a 'Demo' button.

HeyGen

Use Cases ▾ Features ▾ Resources ▾ Company ▾ Pricing

Contact Sales [Get started](#)

AI-powered video creation at scale

Effortlessly produce studio-quality videos with AI-generated avatars and voices.

[Get started for free →](#)

No credit card needed

Microsoft Adobe Amazon COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK kw KELLER WILLIAMS NVIDIA pattern salesforce VOLVO

▶ Demo



On-premise AI deployment

BREAKING

Samsung Bans ChatGPT Among Employees After Sensitive Code Leak

Siladitya Ray Forbes Staff

Covering breaking news and tech policy stories at Forbes.

Follow



May 2, 2023, 07:17am EDT

Updated May 2, 2023, 07:31am EDT

TOPLINE Samsung Electronics has banned the use of ChatGPT and other AI-powered chatbots by its employees, Bloomberg reported, becoming the latest company to crack down on the workplace use of AI services amid concerns about sensitive internal information being leaked on such platforms.



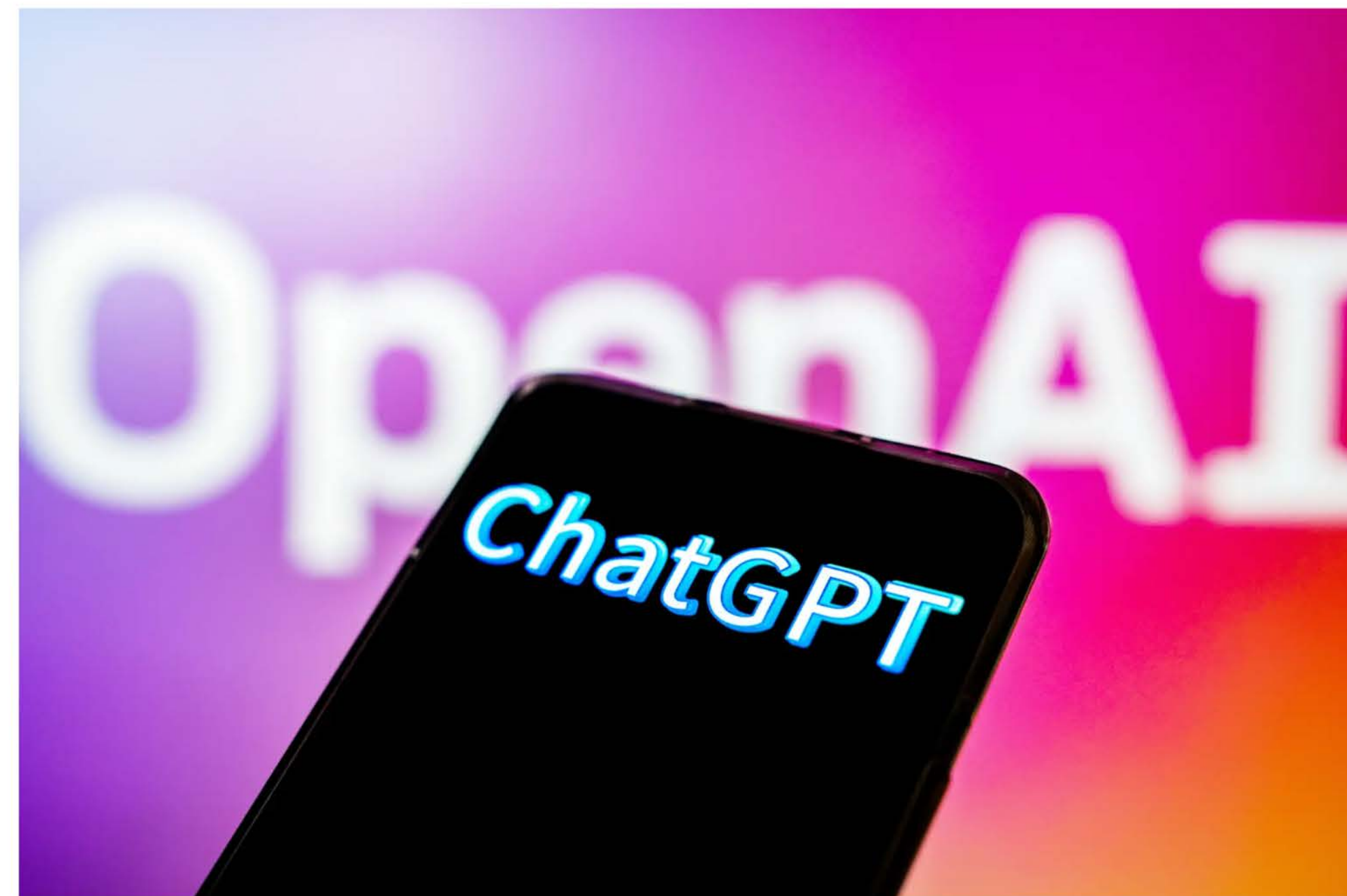
<https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-g-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/?sh=501995e76078>



Amazon, Apple, and 12 other major companies that have restricted employees from using ChatGPT

Aaron Mok Jul 11, 2023, 11:38 PM GMT+8

Share Save



Some companies have issued bans or restrictions around using the buzzy AI chatbot ChatGPT. Getty Images

<https://www.businessinsider.com/chatgpt-companies-issued-bans-restrictions-openai-amazon-apple-2023-7#accenture-doesnt-have-a-formal-ban-on-generative-ai-tools-a-spokesperson-said-14>

What is on-premise computing?

- Computing tasks performed locally
- Infrastructure located within organization's premises
- Data processing and storage on-site
- Reduced reliance on external servers or cloud services
- Provides greater control over security and privacy

Common on-premise computer and applications



PC



Words



PowerPoint

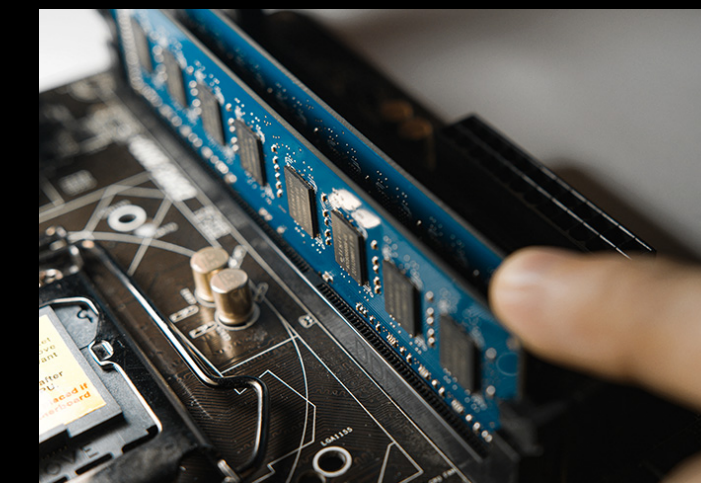
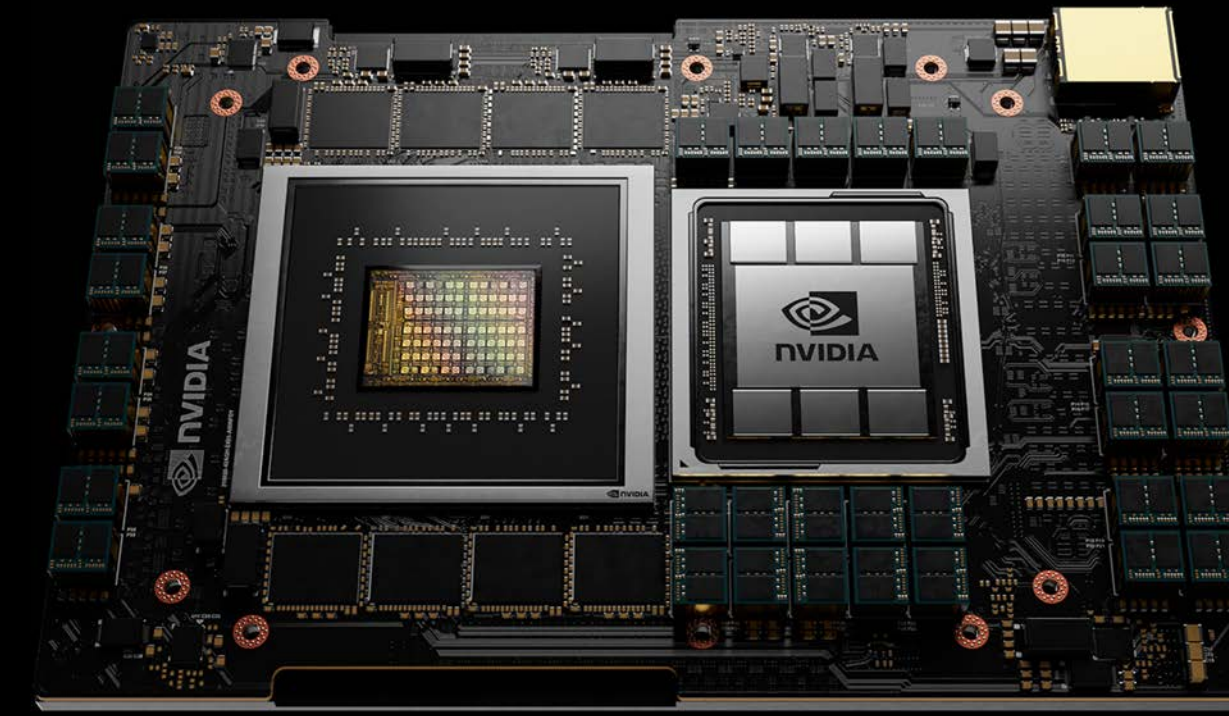


Excel

It works out of the box!

Edge AI Hardware

- Processing Units: CPU and GPU, general purpose and parallel processing to handle AI tasks
- Memory (RAM): for storing and data access during during inference
- Power Supply: Reliable power supply, including adapters
- Storage: Incorporate SSDs, or hard disk drives for storing datasets, models, and quick access to offline functionality



Benefits of using on-premise AI language models



Enhanced Efficiency

Faster response and training:
On-premises AI offers quicker conversation generation and training times.

Local Data Backup:
On-premises backup mechanisms reduce the risk of data and conversation history loss.



Data Privacy

Local Language Model Processing:
Clients have full control over what is saved on local AI servers, and data never goes through external AI vendors.

Local Training Materials Storage:
Sensitive data remains within the local device, alleviating concerns about data privacy and security.



Cost-effectiveness

Scale up without Extra Cloud Costs:
The flexibility to scale usage without incurring additional cloud costs based on consumption or resource scaling.

Efficient Resource Utilization:
By sharing a common pool of AI computation resources, clients can avoid paying for idle resources, as resource usage may vary among users.

Benefits of on-premise AI deployment



- **Data Security**
- **Low Latency**
- **Customization**
- **Regulatory Compliance**
- **Cost Efficiency**



What can generative AI do for engineer?

Generative AI for engineers

A. Engineering assistant:

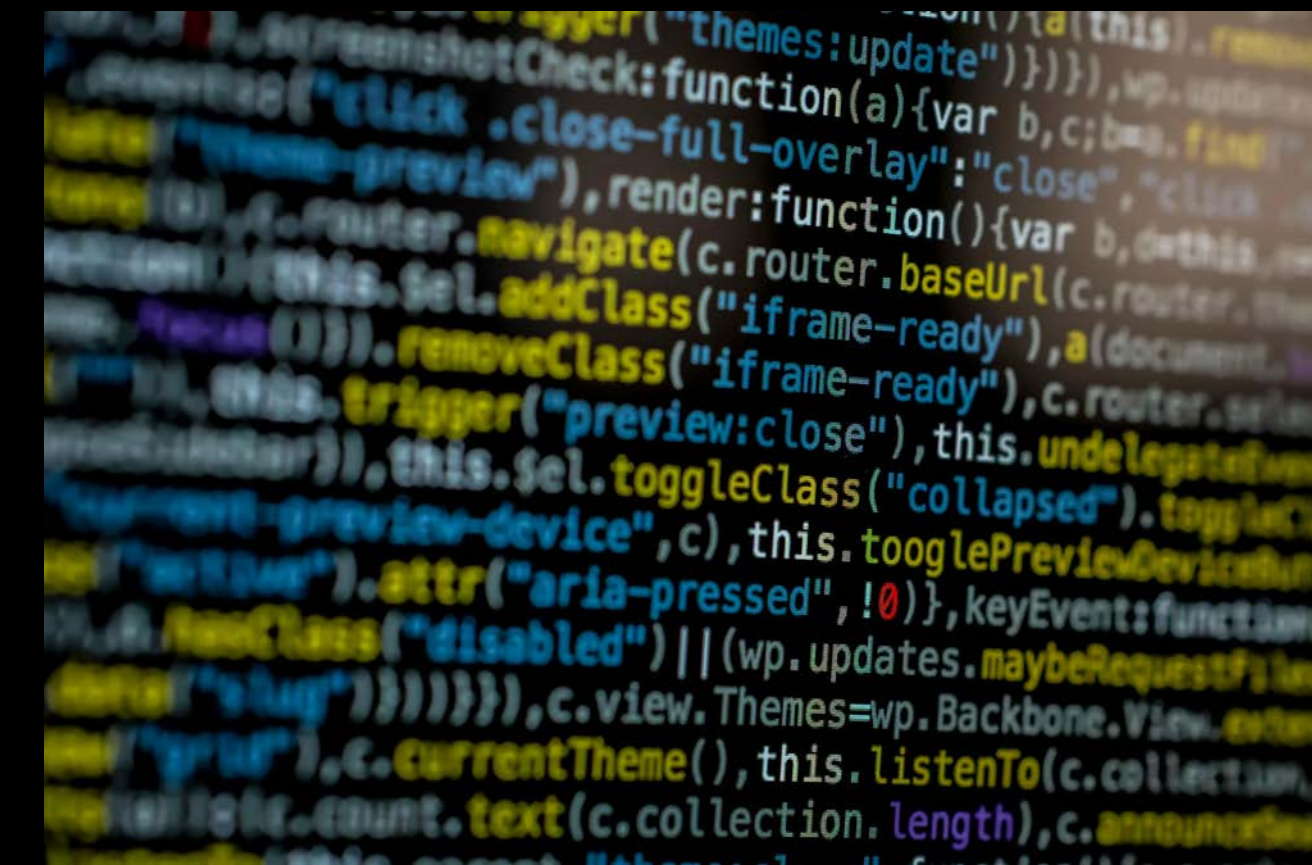
- Help confirm details of mathematical formulas, competitor announcements, and specialized material specifications faster than search engines.

B. Summarizing engineering information:

- Summarize extensive collections of documents.

C. Developing software:

- Produce first-draft software applications quickly.
- Eliminates repetitive coding tasks like input-handling, data validation, and read-and-write operations.
- Rapid software development.



Generative AI for engineers

D. Creating written content:

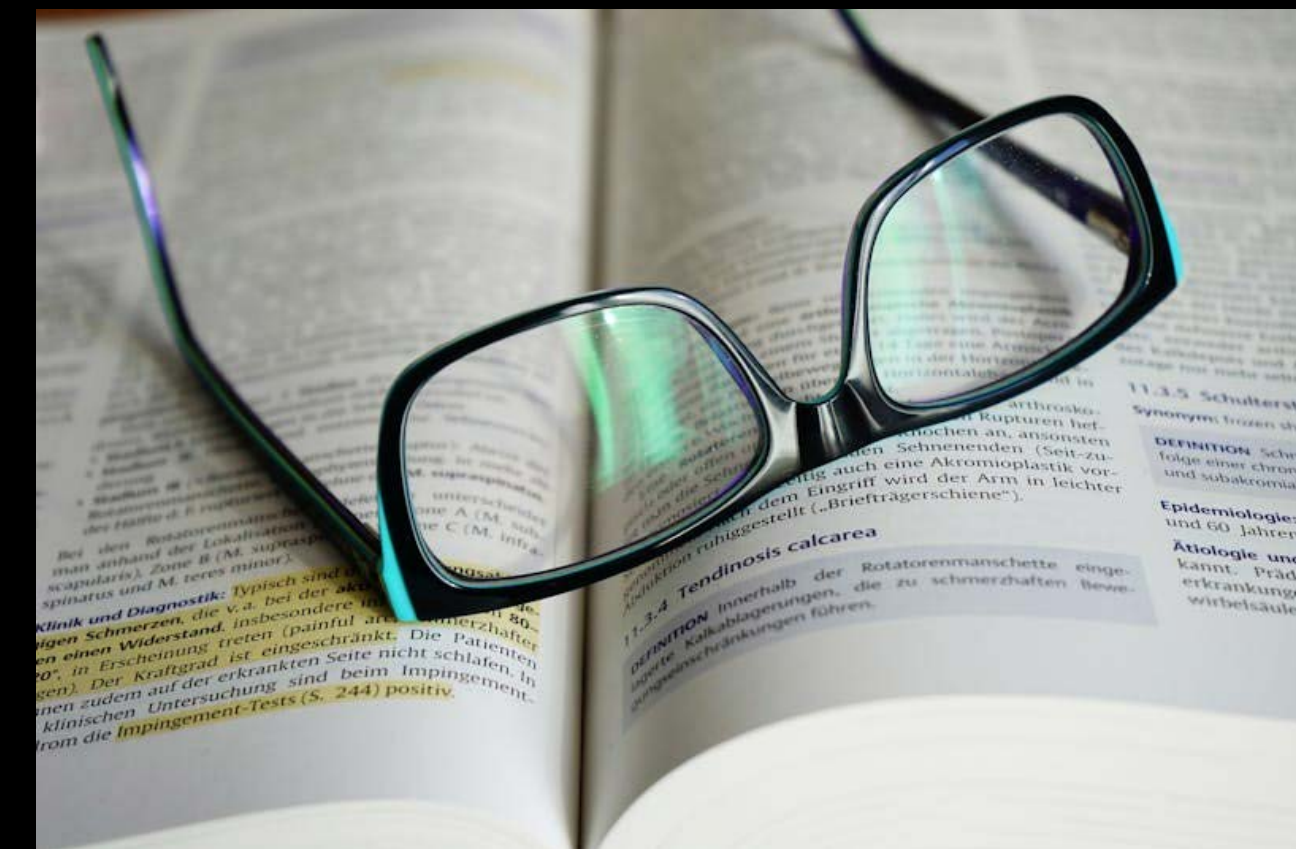
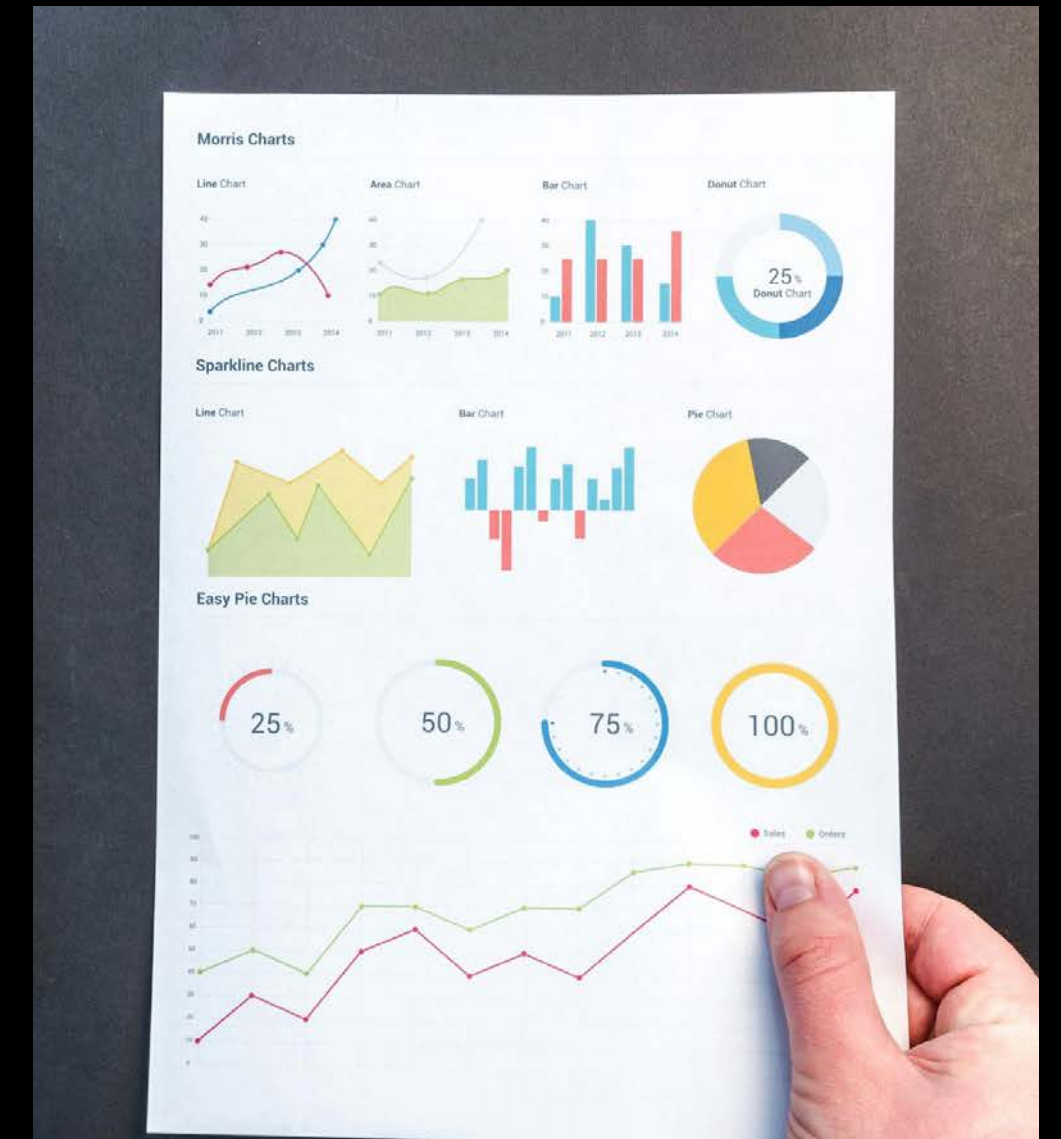
- Generate first drafts of reports, for engineers to polish and finalize.
- Produce reports from lab notes, web pages, emails, and engineering case studies.

E. Comparing engineering regulations:

- Compare regulatory documents from multiple jurisdictions.
- Highlight differences, gaps, and incompatibilities.
- Handle variations in terminologies and synonyms across documents.

F. Drilling into information:

- Process text and can summarize lengthy reports in minutes.
- Drill down into specific issues of interest to retrieve accurate and targeted information.



What's next?

1. Further development
2. Can generative AI do more?
3. What kind of jobs will be replaced?

The applications of Generative-AI in different industry

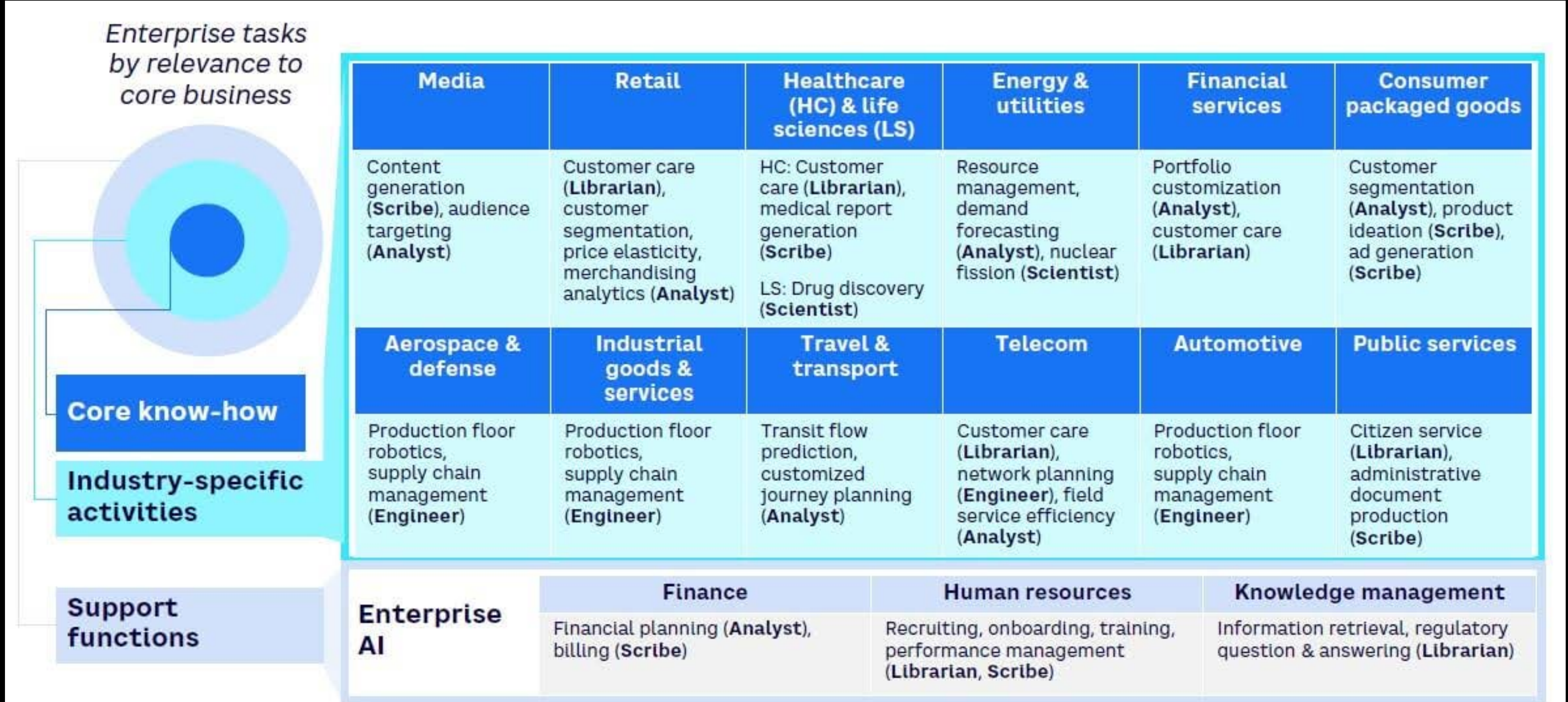
Generative AI use cases, nonexhaustive

Modality	Application	Example use cases
Text	Content writing	Marketing: creating personalized emails and posts Talent: drafting interview questions, job descriptions
	Chatbots or assistants	Customer service: using chatbots to boost conversion on websites
	Search	Making more natural web search Corporate knowledge: enhancing internal search tools
	Analysis and synthesis	Sales: analyzing customer interactions to extract insights Risk and legal: summarizing regulatory documents
Code	Code generation	IT: accelerating application development and quality with automatic code recommendations
	Application prototype and design	IT: quickly generating user interface designs
	Data set generation	Generating synthetic data sets to improve AI models' quality
Image	Stock image generator	Marketing and sales: generating unique media
	Image editor	Marketing and sales: personalizing content quickly
Audio	Text to voice generation	Trainings: creating educational voiceover
	Sound creation	Entertainment: making custom sounds without copyright violations
	Audio editing	Entertainment: editing podcast in post without having to rerecord

The applications of Generative-AI in different industry

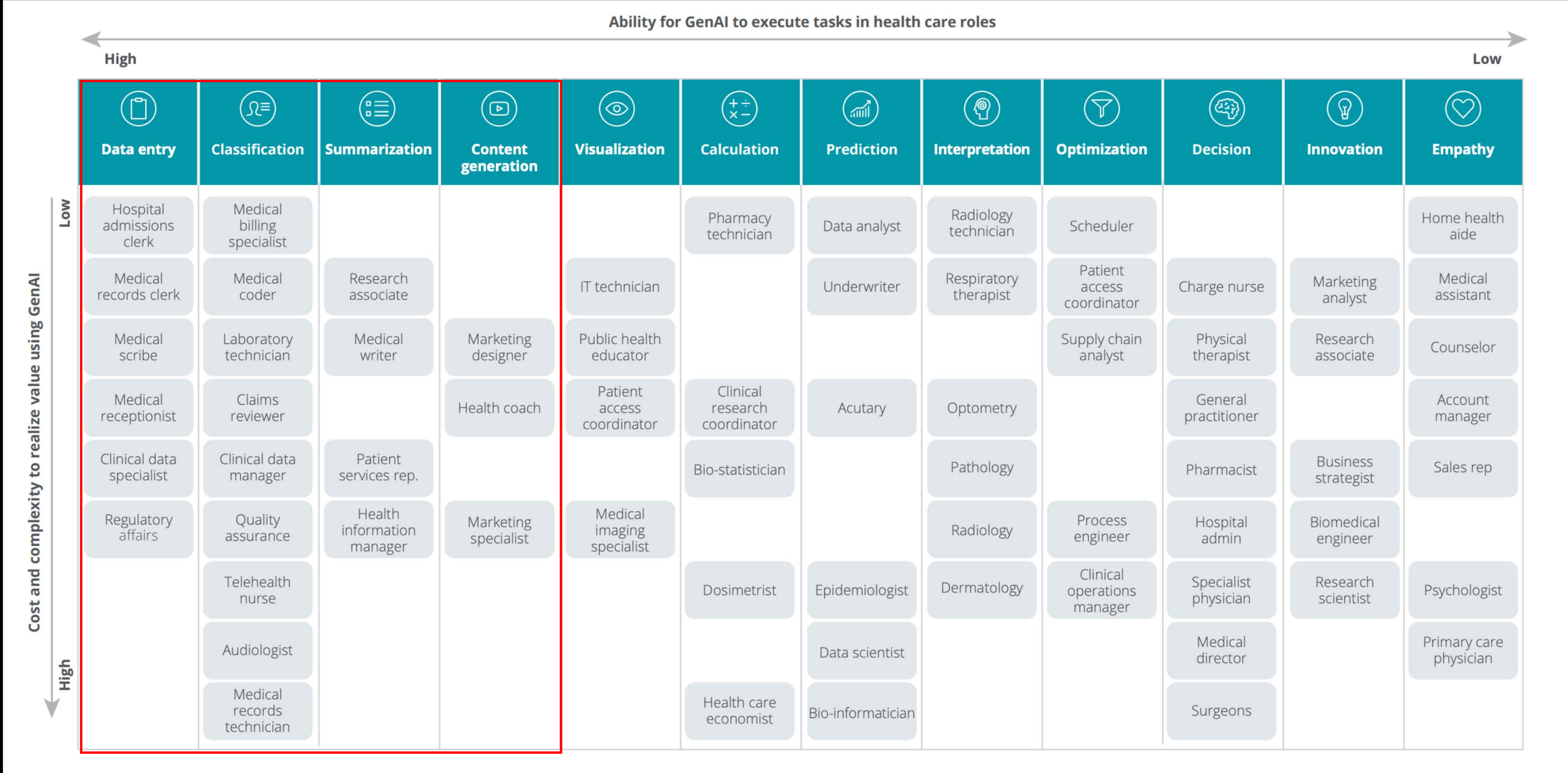
3-D or other	3-D object generation	Video games: writing scenes, characters Digital representation: creating interior-design mockups and virtual staging for architecture design
	Product design and discovery	Manufacturing: optimizing material design Drug discovery: accelerating R&D process
Video	Video creation	Entertainment: generating short-form videos for TikTok Training or learning: creating video lessons or corporate presentations using AI avatars
	Video editing	Entertainment: shortening videos for social media E-commerce: adding personalization to generic videos Entertainment: removing background images and background noise in post
	Voice translation and adjustments	Video dubbing: translating into new languages using AI-generated or original-speaker voices Live translation: for corporate meetings, video conferencing Voice cloning: replicating actor voice or changing for studio effect such as aging
	Face swaps and adjustments	Virtual effects: enabling rapid high-end aging; de-aging; cosmetic, wig, and prosthetic fixes Lip syncing or “visual” dubbing in postproduction: editing footage to achieve release in multiple ratings or languages Face swapping and deep-fake visual effects Video conferencing: real-time gaze correction

The impact of Generative-AI in different industry



Source: Arthur D. Little

The impact of Generative-AI in healthcare industry



"From code to cure, how Generative AI can reshape the health frontier" (Deloitte)

Thank you

johnson.shum@ailog.com

