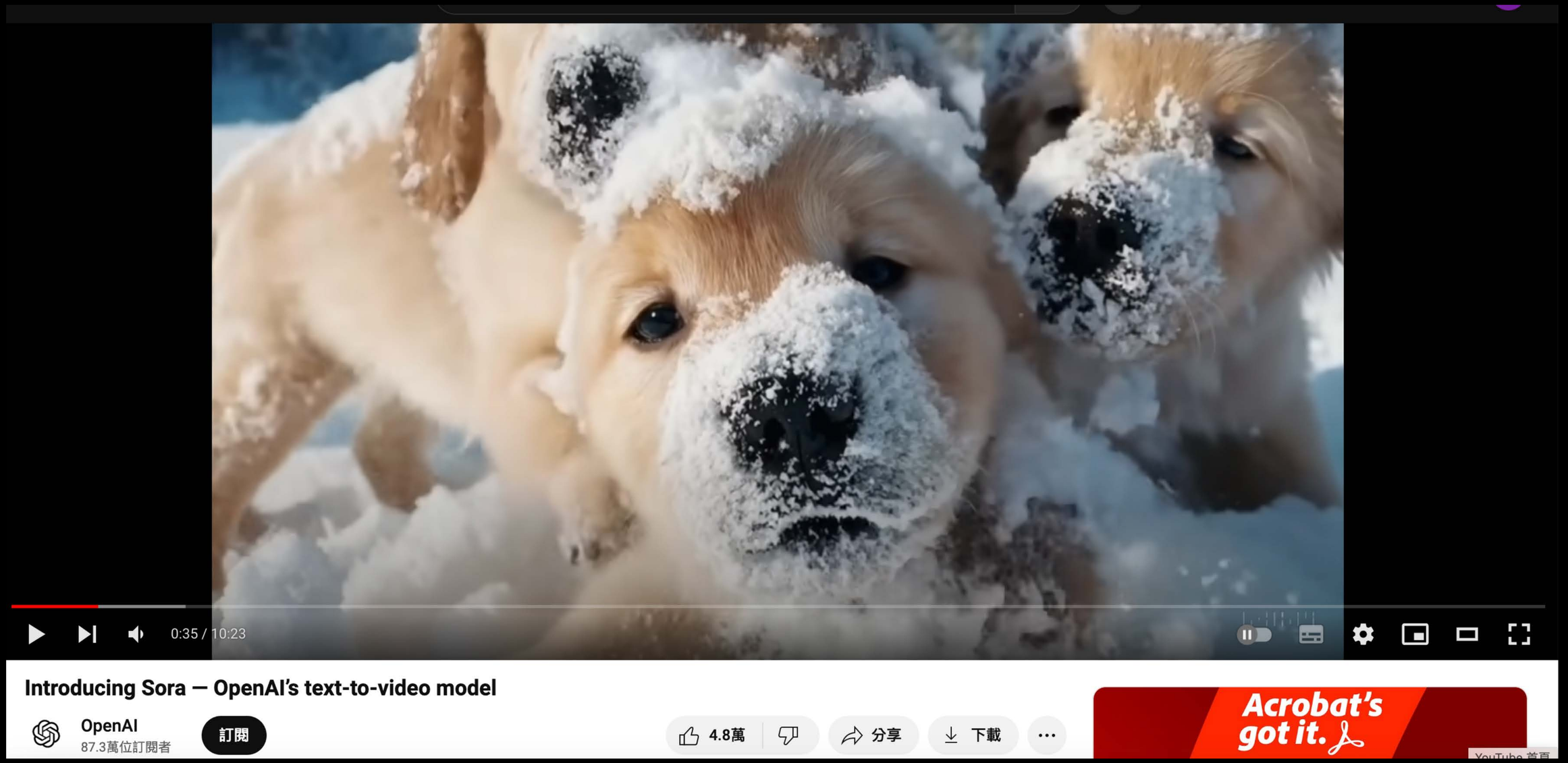
Practical Guide to Generative A.I.

Johnson Shum

Aiilog Limited

For education purpose only



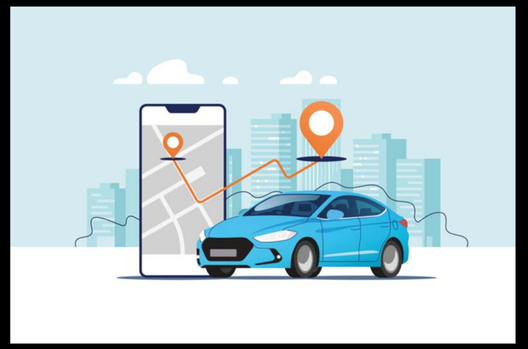
https://www.youtube.com/watch?v=HK6y8DAPN_0

What is AI

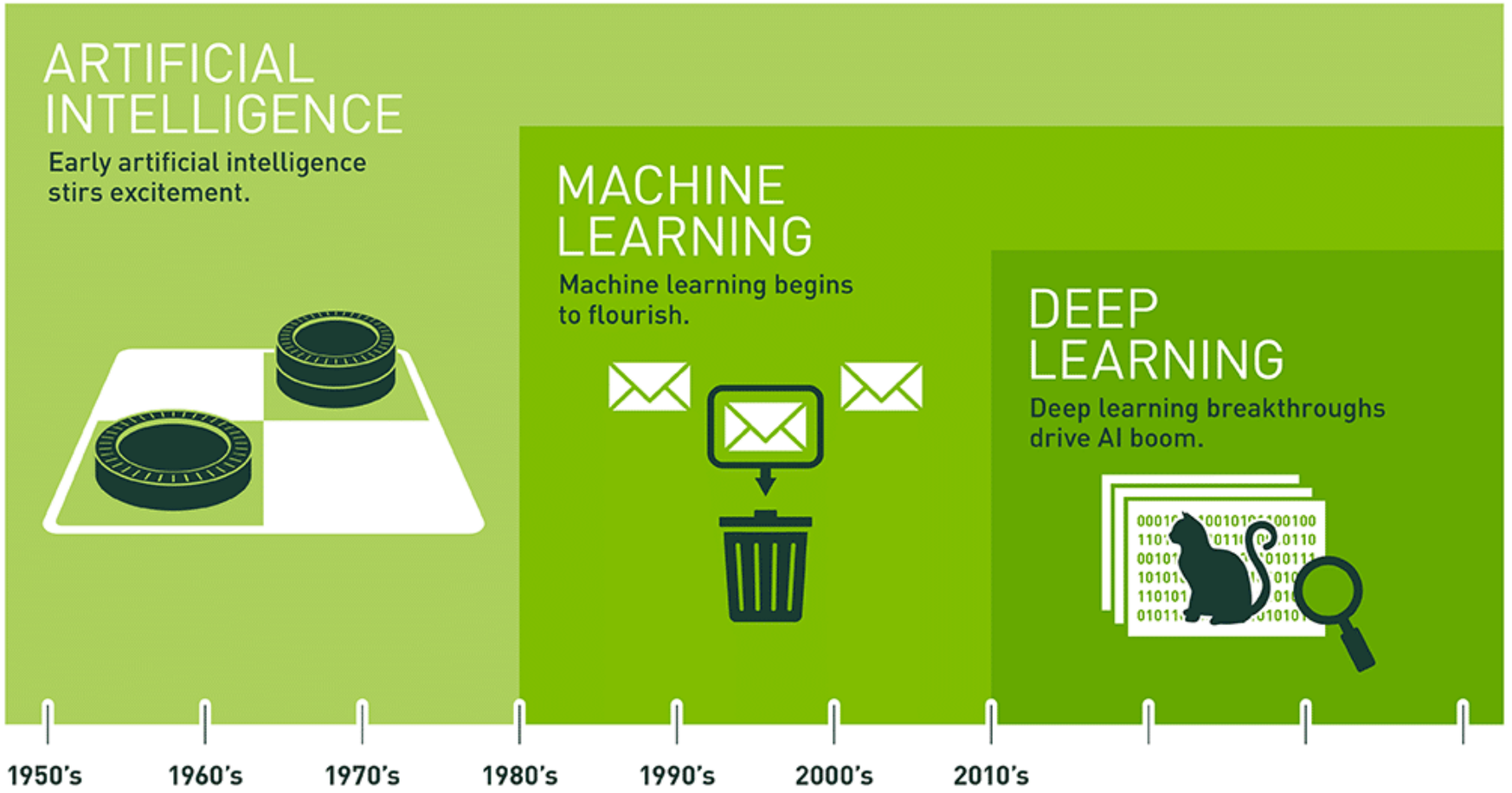
What is AI

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.

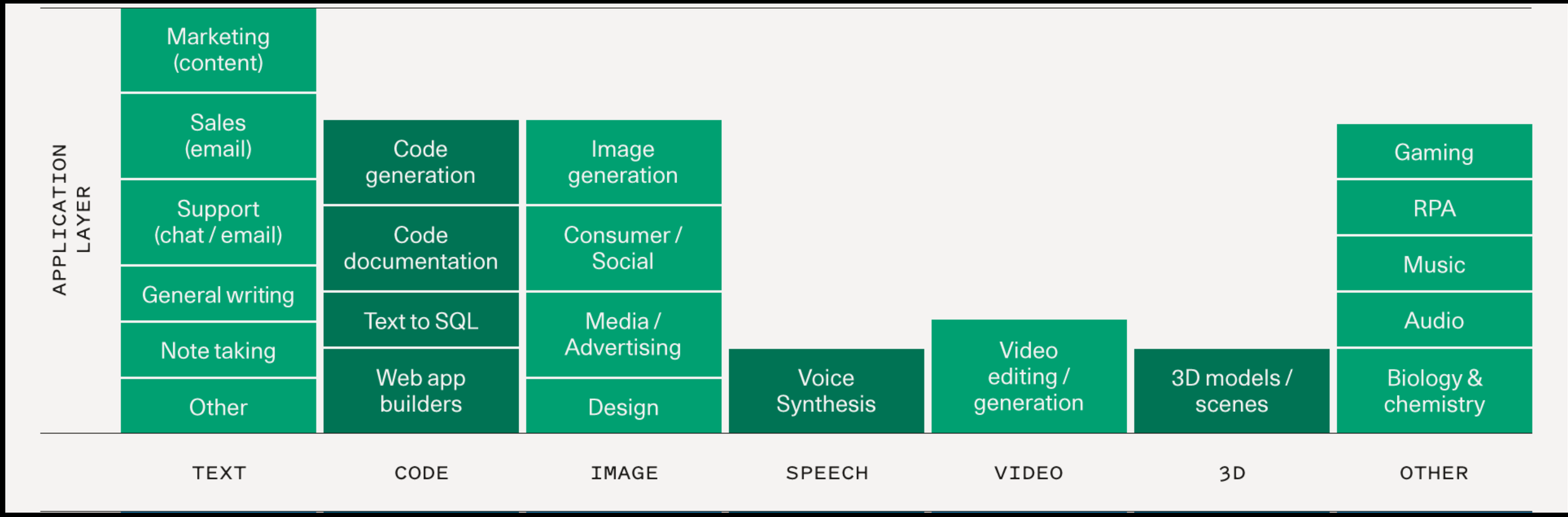
https://blogs.nvidia.com/blog/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

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



https://www.sequoiacap.com/article/generative-ai-a-creative-new-world/?trk=cndc-detail

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

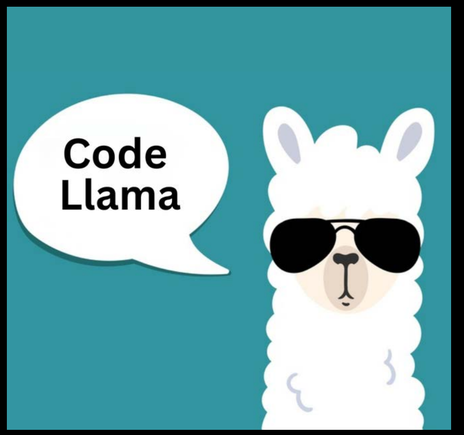
https://www.youtube.com/watch?v=3JDuR-elYWA

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.

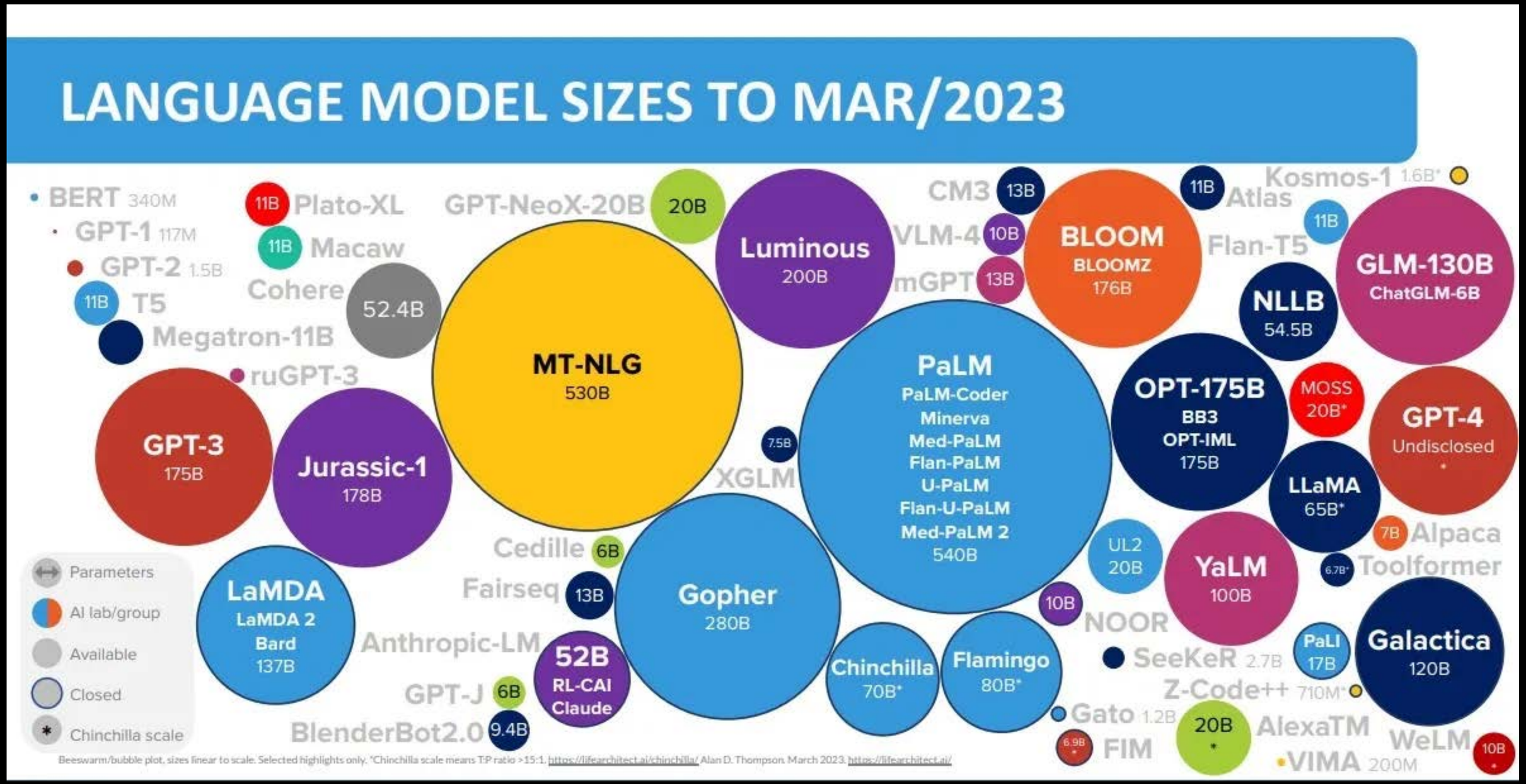
What is Large Language Models (LLM)

- 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.



Bard

Generative AI Model Size



https://jair-neto.medium.com/demystifying-large-language-models-for-everyone-fine-tuning-your-own-llm-part-1-3-12d4b0647541

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

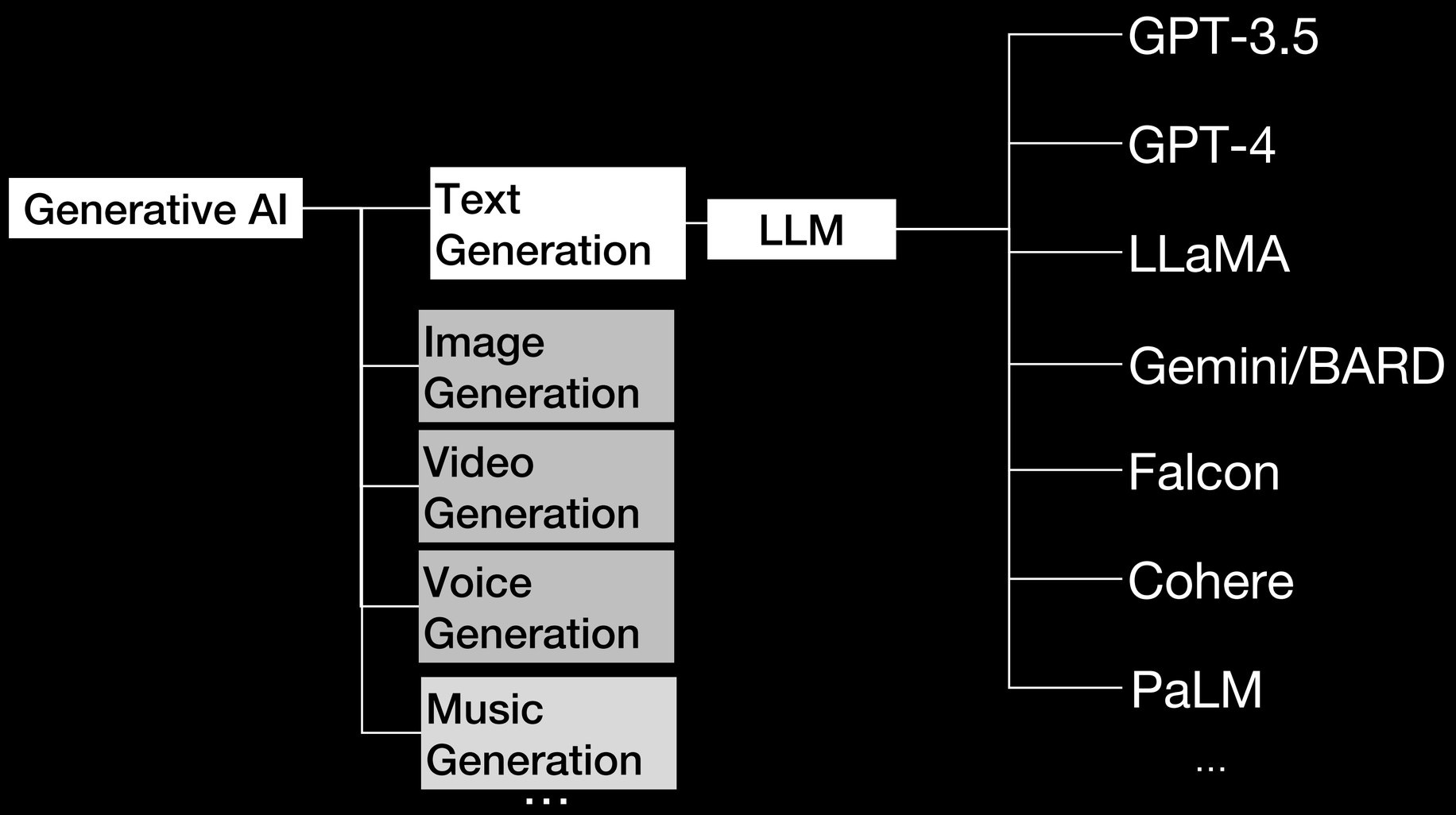
The LLM is pretrained with massive amounts of text

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.

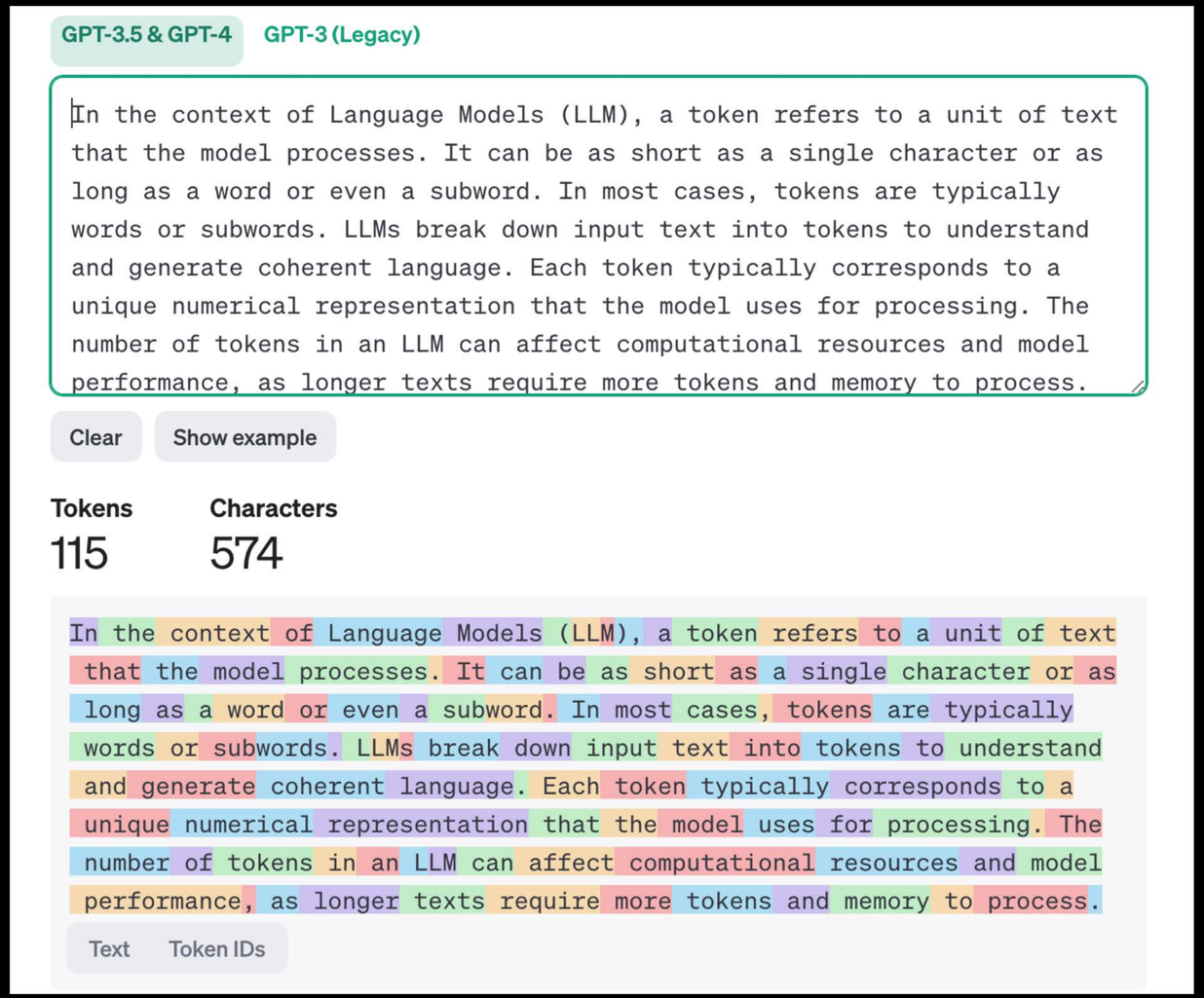
Token

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



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 ~= 4 chars in English

1 token ~= ¾ words

100 tokens ~= 75 words

Or

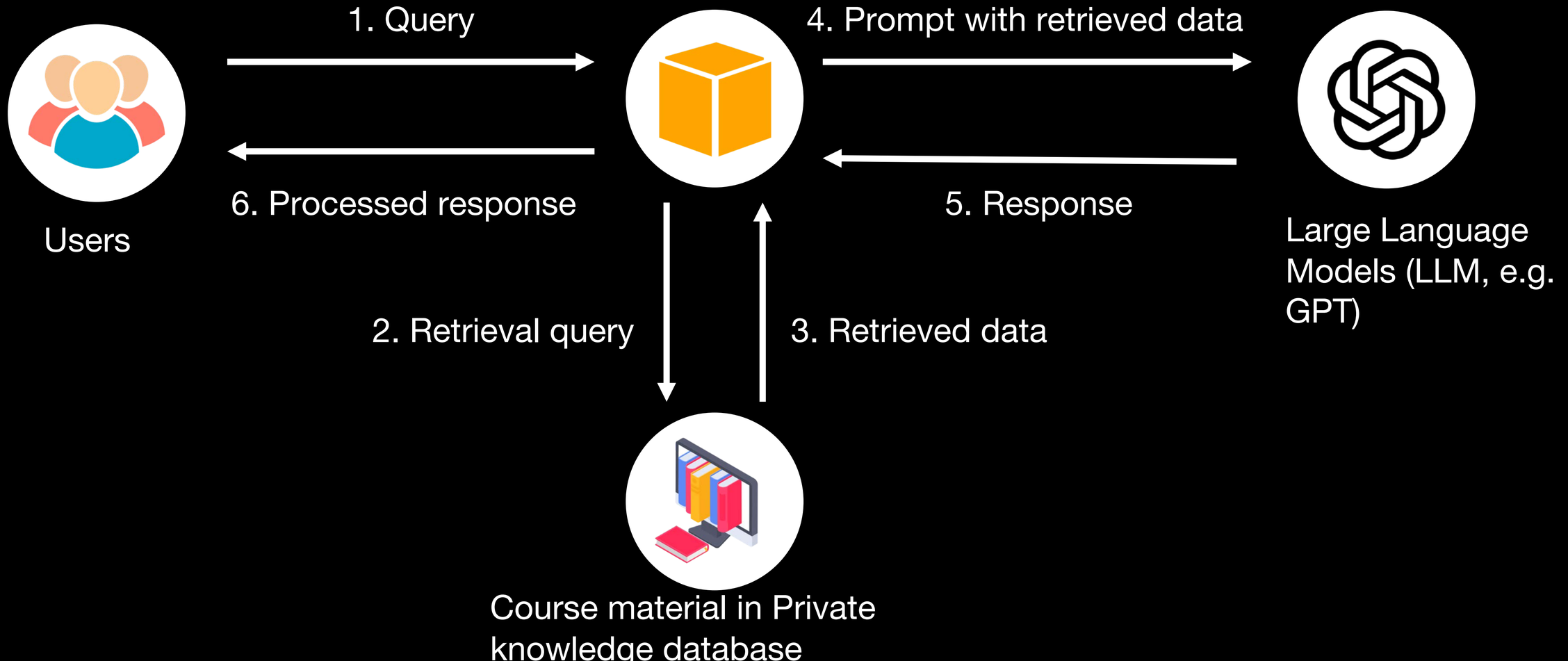
1–2 sentences ~= 30 tokens

1 paragraph ~= 100 tokens

1,500 words ~= 2048 tokens

Retrieval-Augmented Generation (RAG)

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

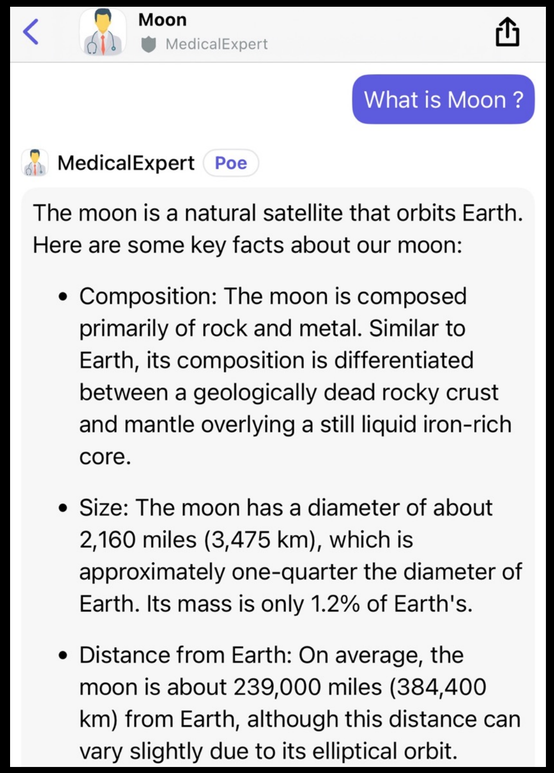


AI common problems

GPT 4

France never gifted the Vilnius TV tower to Lithuania

POE Bot creation (GPT 3.5)



Medical knowledge bots answer unrelated things

AI hallucination when an LLM presents false info
Source: https://www.tidio.com/blog/ai-hallucinations/

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

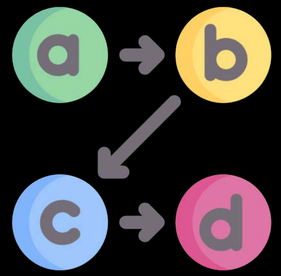
Conversation
Engage in dialogue and provide relevant responses

Coding
Interpret and generate code

LLM benchmarks

Translation
Accurately translate text

Math
Math problems, from arithmetic to calculus



Logic
Logical reasoning like inductive reasoning

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

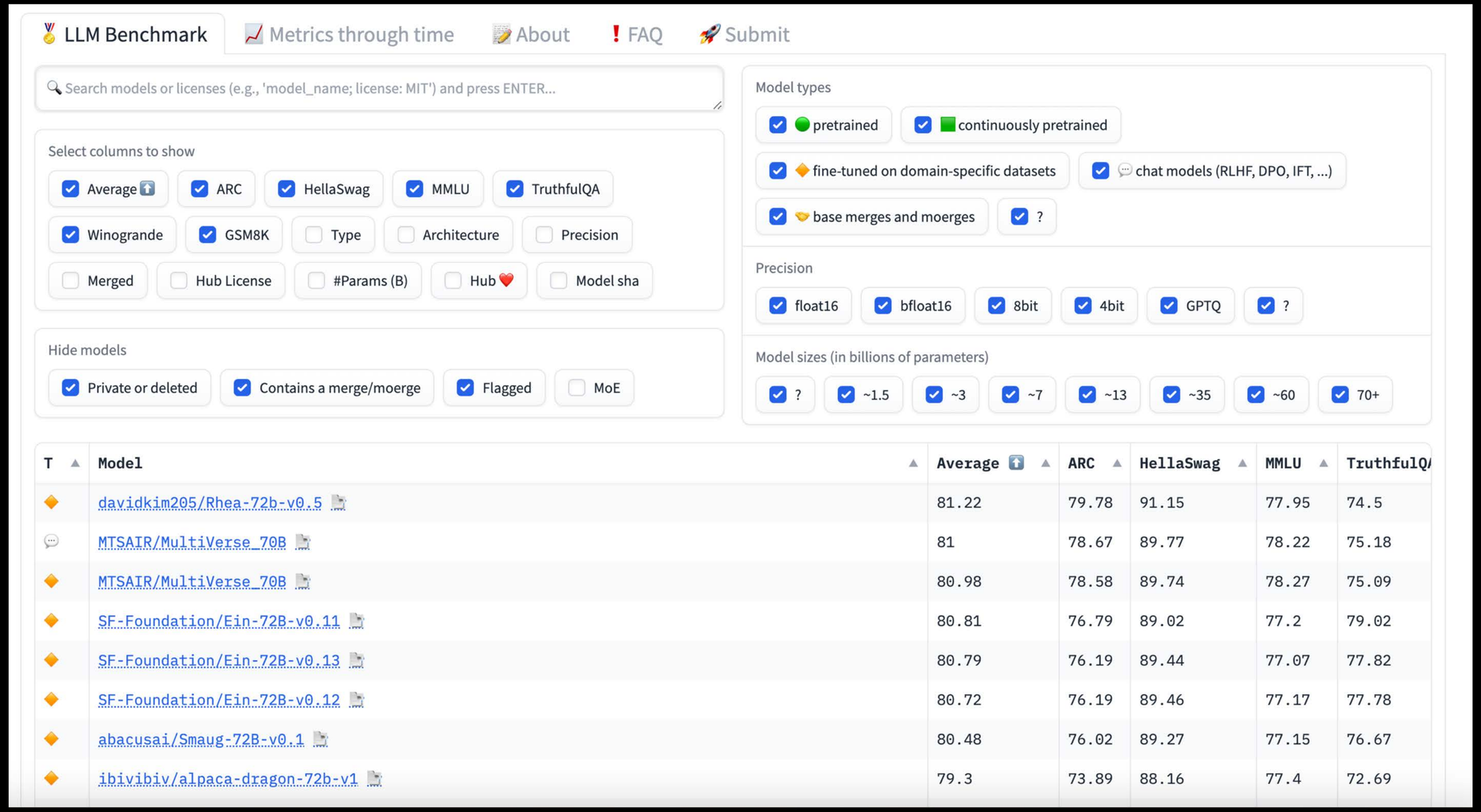
Volume of Questions

- 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

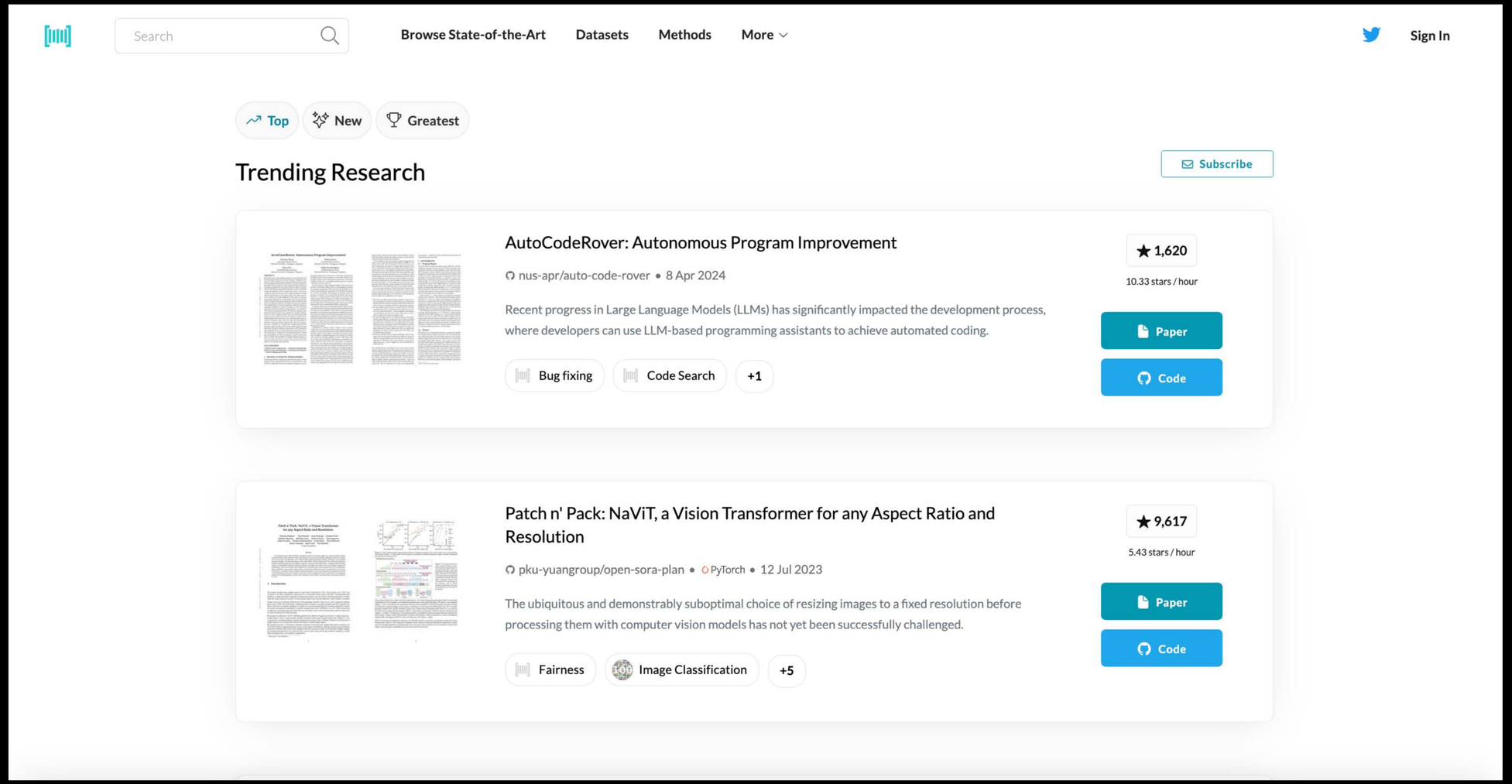
LLM benchmarks



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



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-lm-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

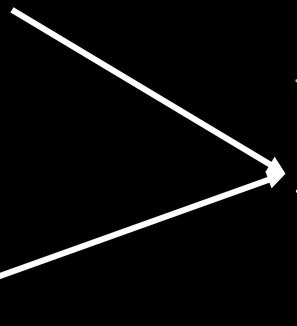
Source: https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

HellaSwag

HellaSwag

ActivityNet

WikiHow



70,000 multiple-choice questions

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.

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.

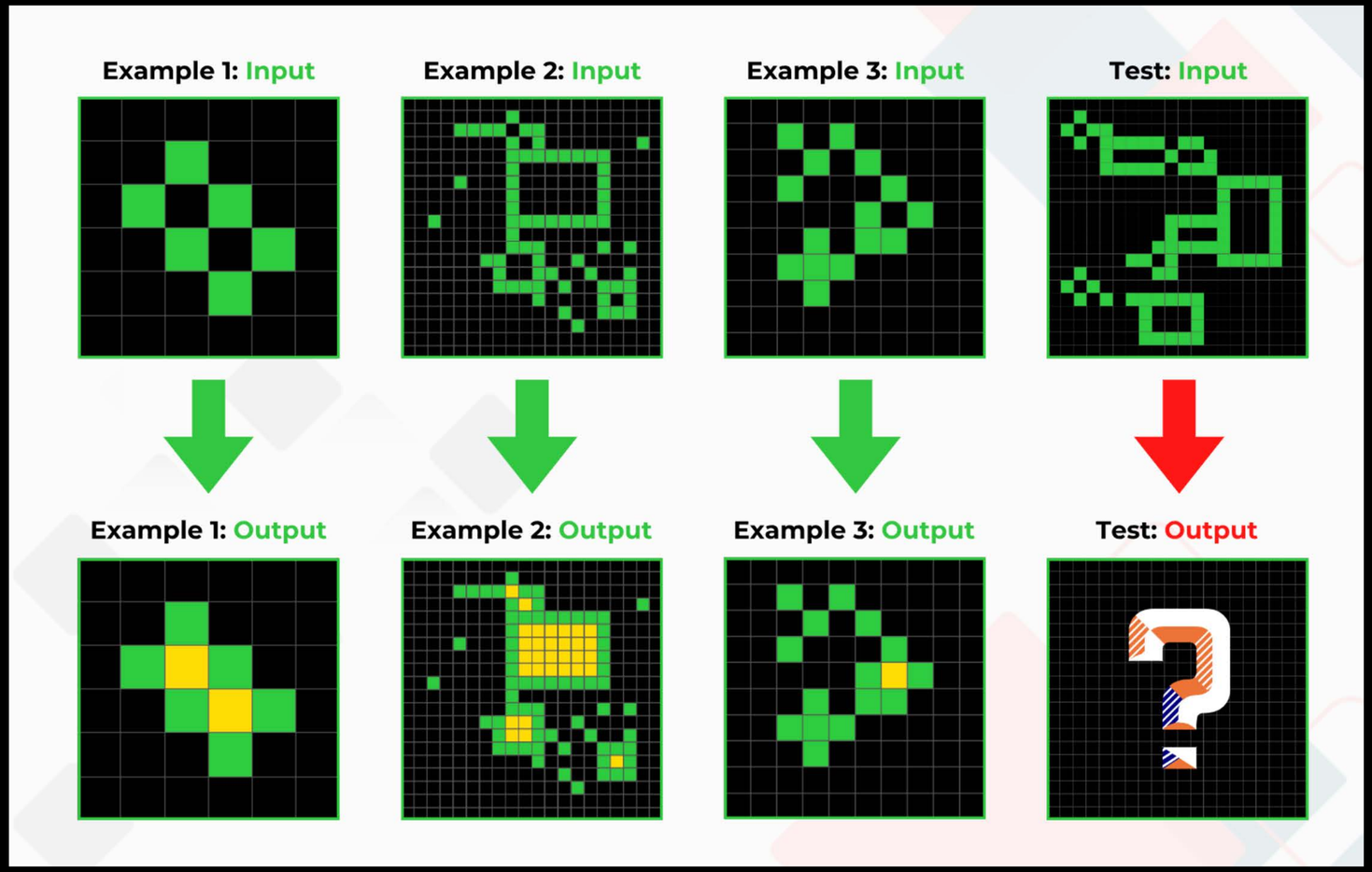
	Twin sentences	Options (**answer**)
✗	The monkey loved to play with the balls but ignored the blocks because he found **them** *exciting*.	**balls** / blocks
	The monkey loved to play with the balls but ignored the blocks because he found **them** *dull*.	balls / **blocks**
✗	William could only climb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

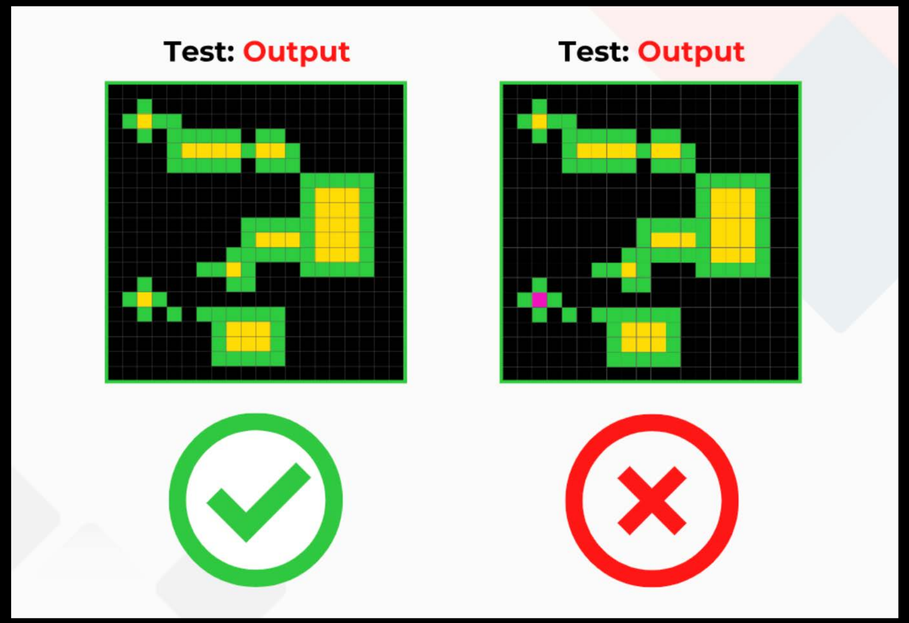
https://winogrande.allenai.org/

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
13	**PaLM 540B** (0-shot)	81.1	PaLM: Scaling Language Modeling with Pathways			2022
14	**Camelidae-8×34B**	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)





https://lab42.global/arc/

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
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-8×34B**	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

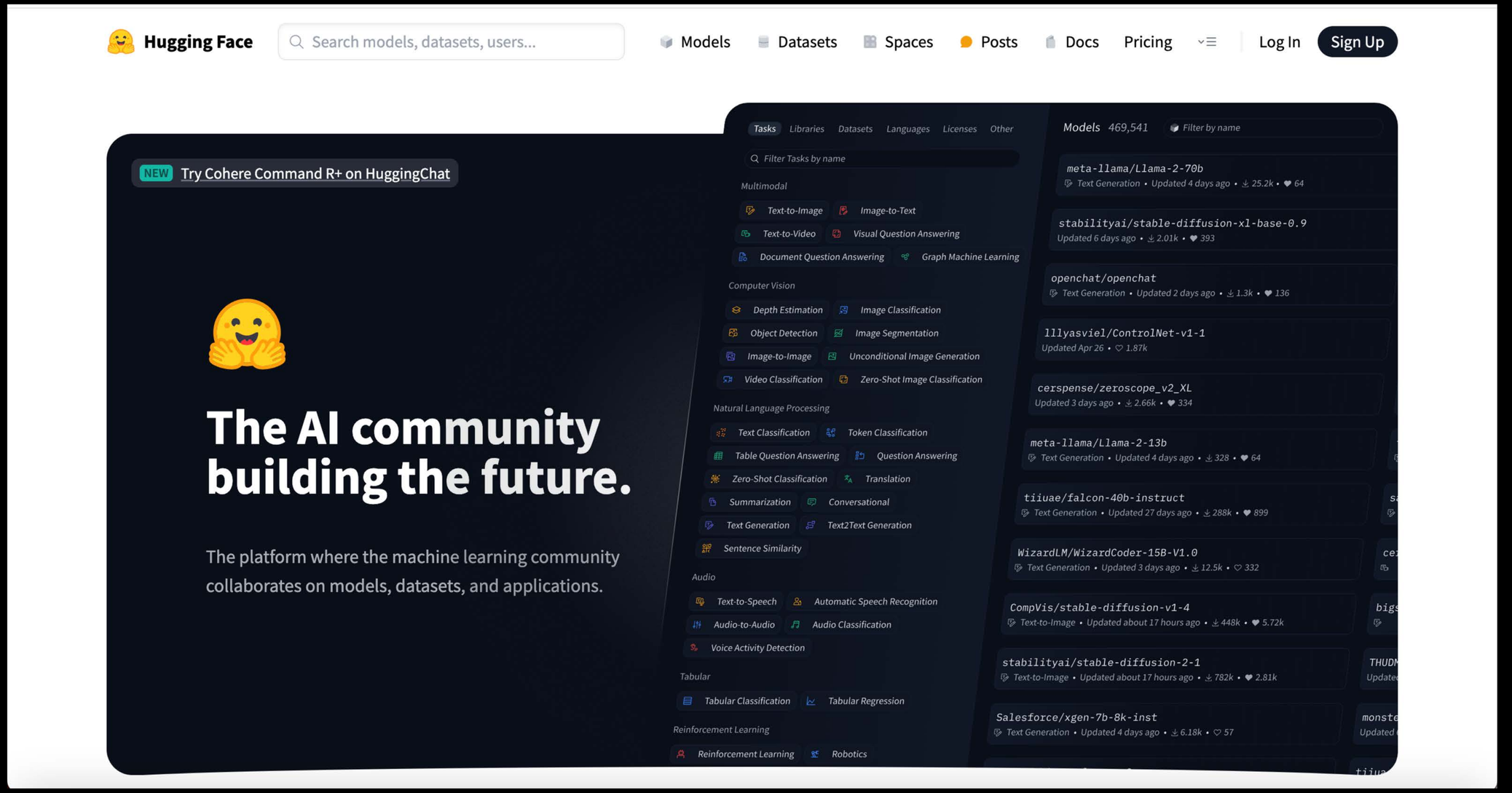
Comparison

	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

https://www.ithome.com.tw/news/160291

Common websites

https://huggingface.co/



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

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

BookQ: book-based question answering.

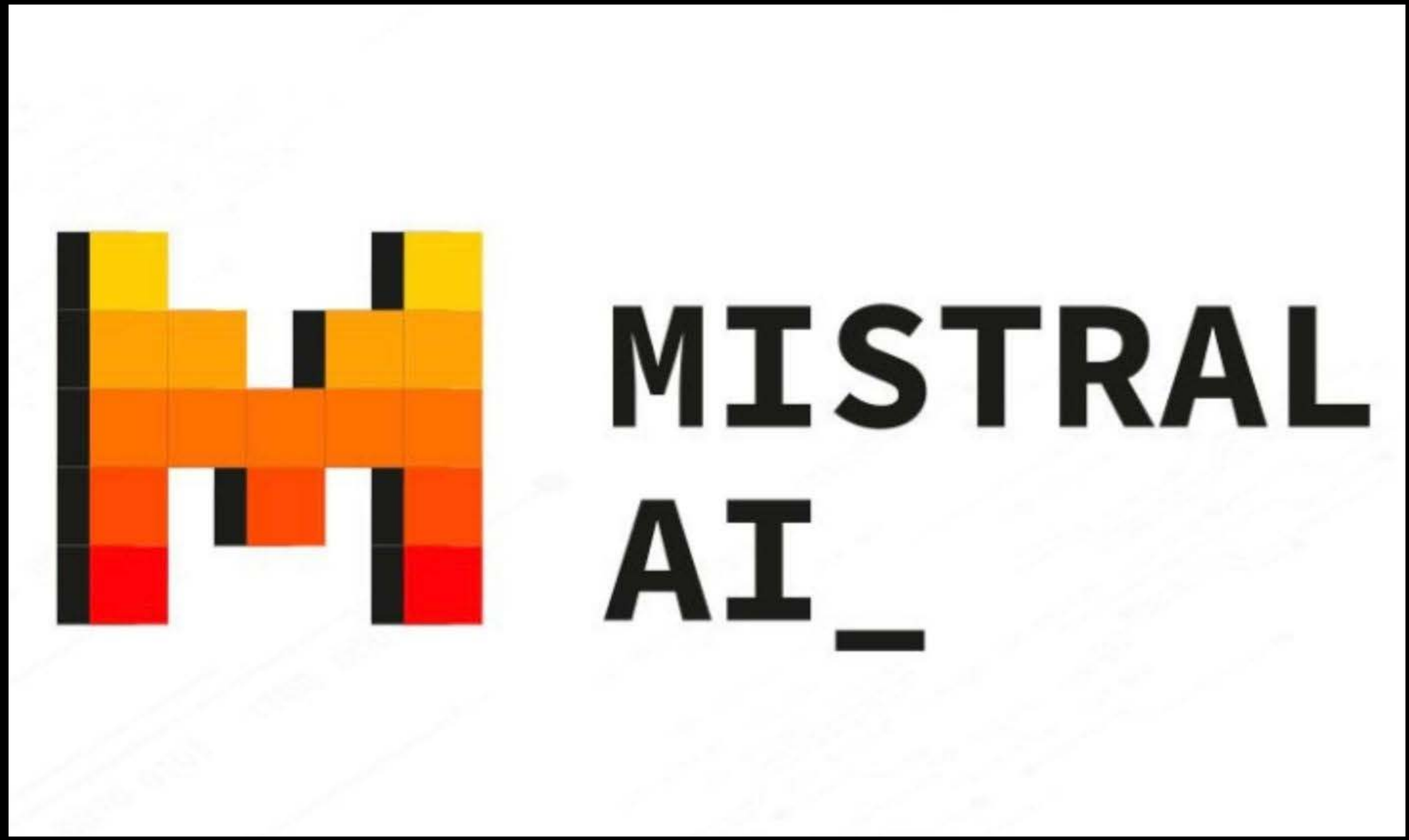
HellaSwag: Commonsense reasoning.

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

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

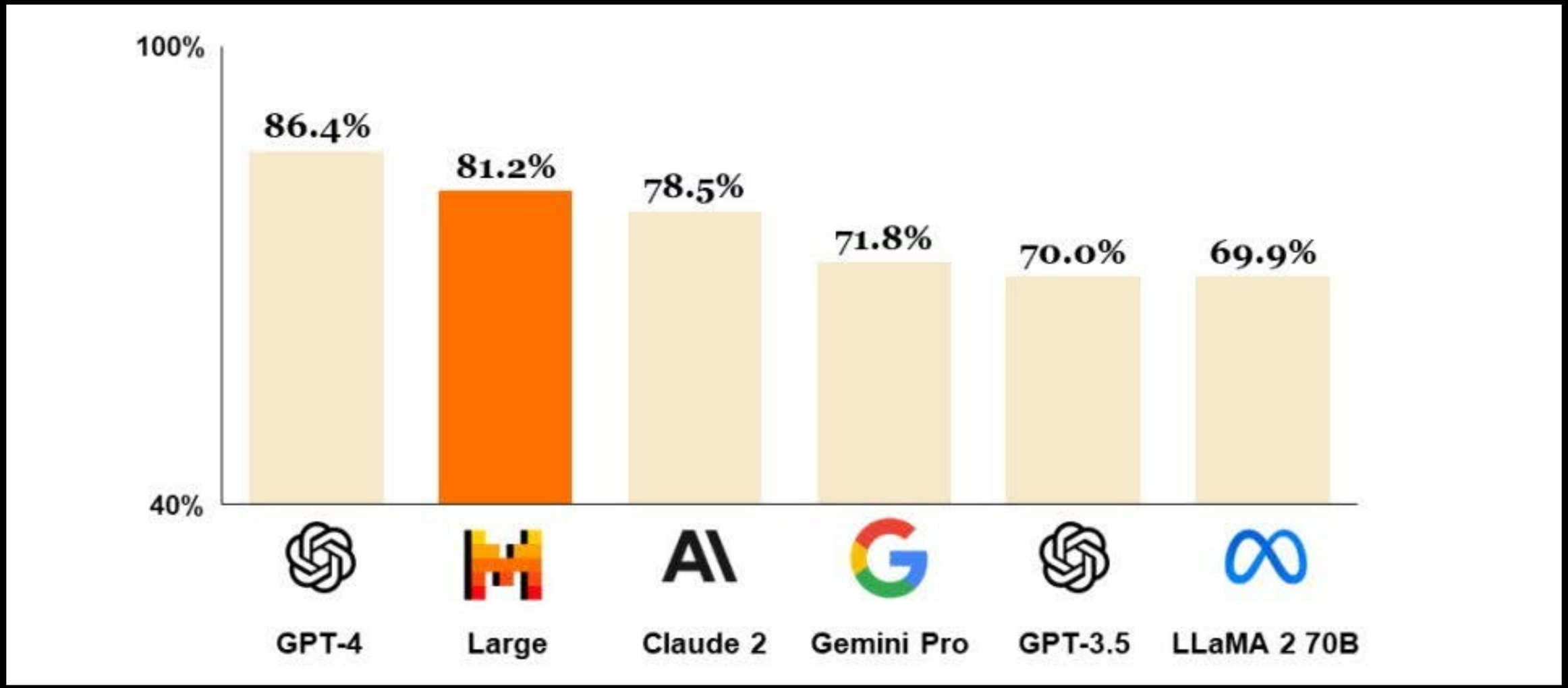
- Mistral 7B
- Mixtral 8x7B
- Mistral Medium
- Mistral Large
- Mixtral 8x22B

Mistral AI

8x7B model

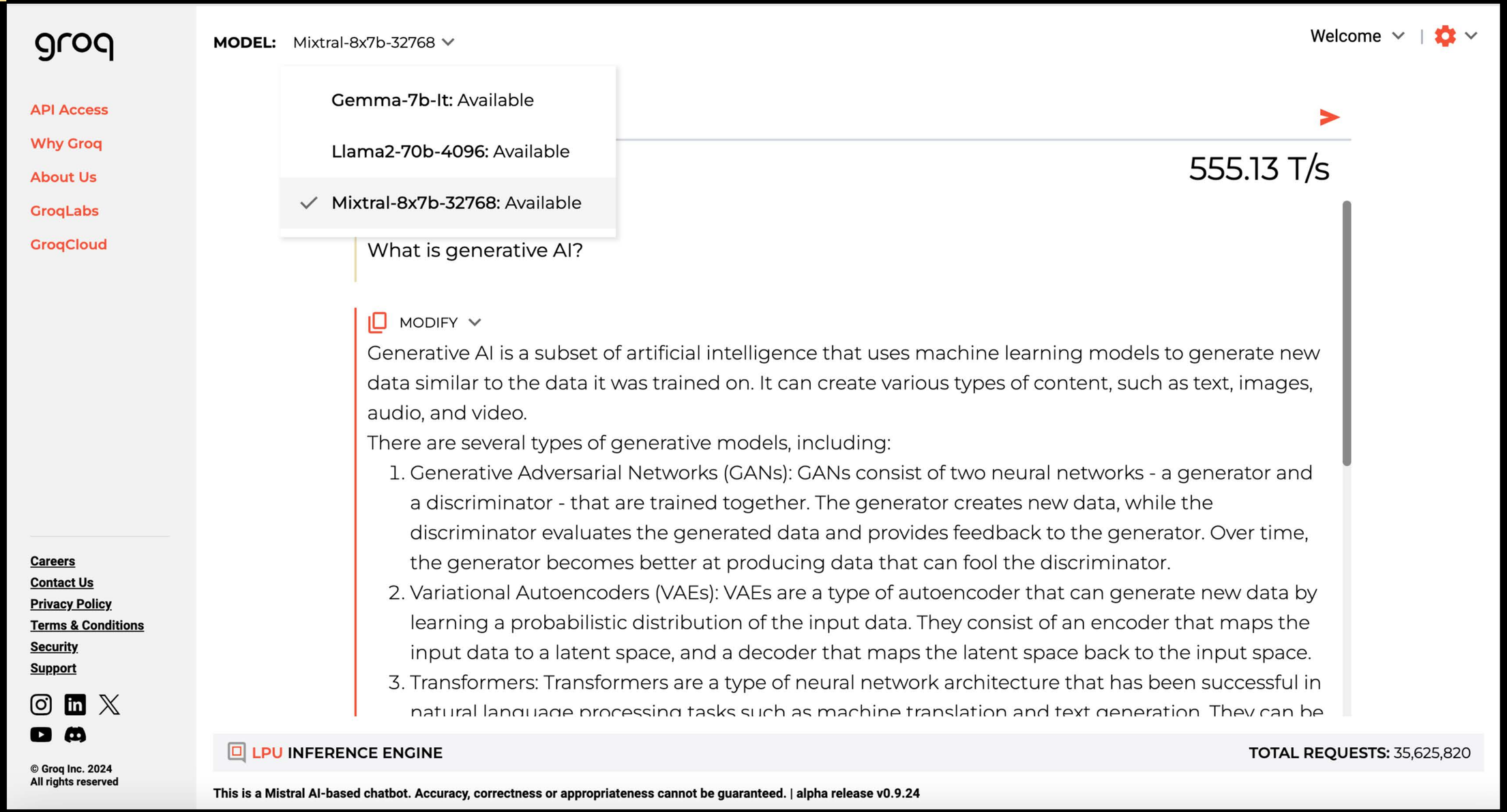
- Handle contexts with up to 32,000 tokens
- The Mixtral 8x7B model has a total of 46.7 billion parameters
- The model's computational cost is comparable to a model with only 12.9 billion parameters.
- The model supports English, French, Italian, German, and Spanish
- Mixtral 8x7B Instruct, scores 8.3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

Mistral Large



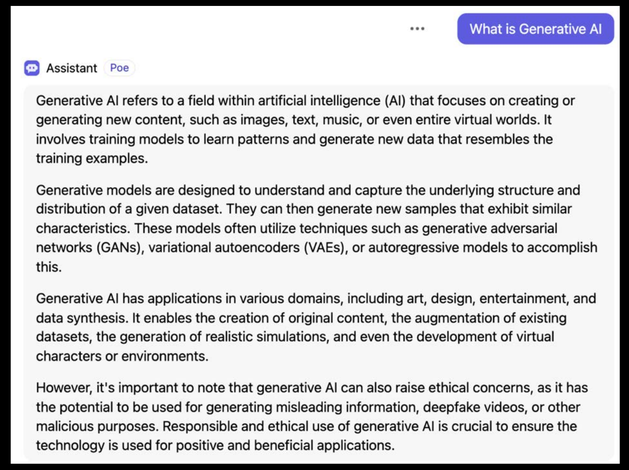
MMLU test data by Mistral AI

Open source LLM

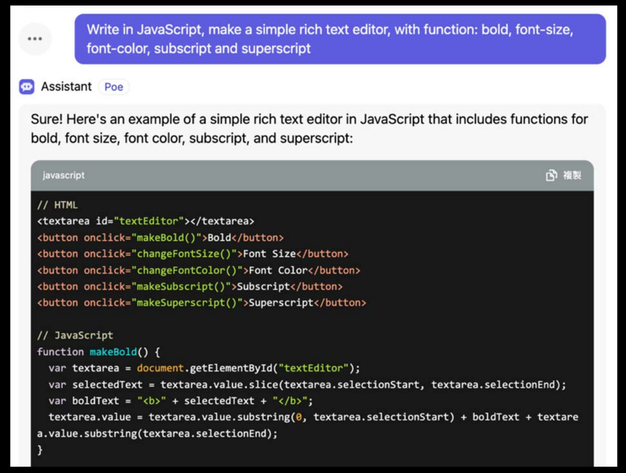


You may try them on https://groq.com

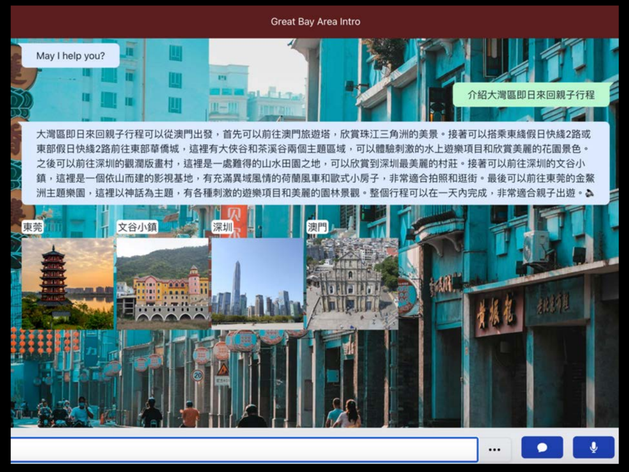
The Power of Generative AI



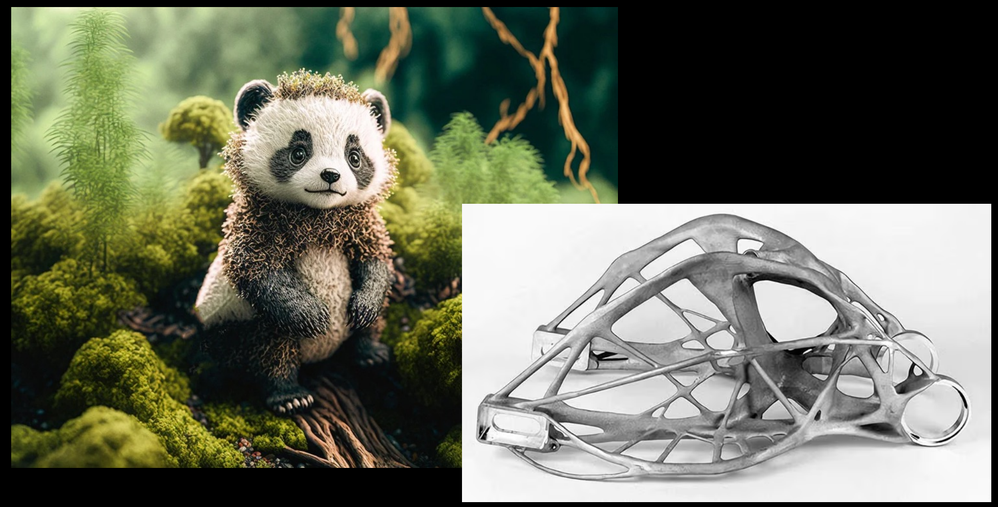
Text generation and summarization



Computer programming, writing, and debugging code



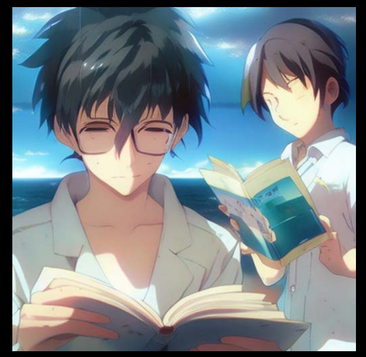
Chatbots that can understand natural language and have personalities



Generative design and artwork

AI image generation

Image generation

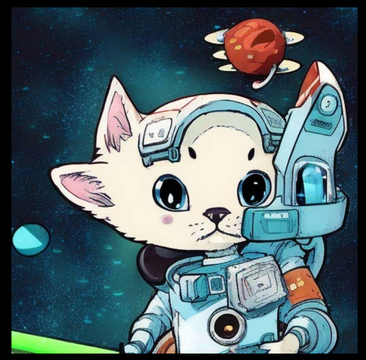









Text-to-Image



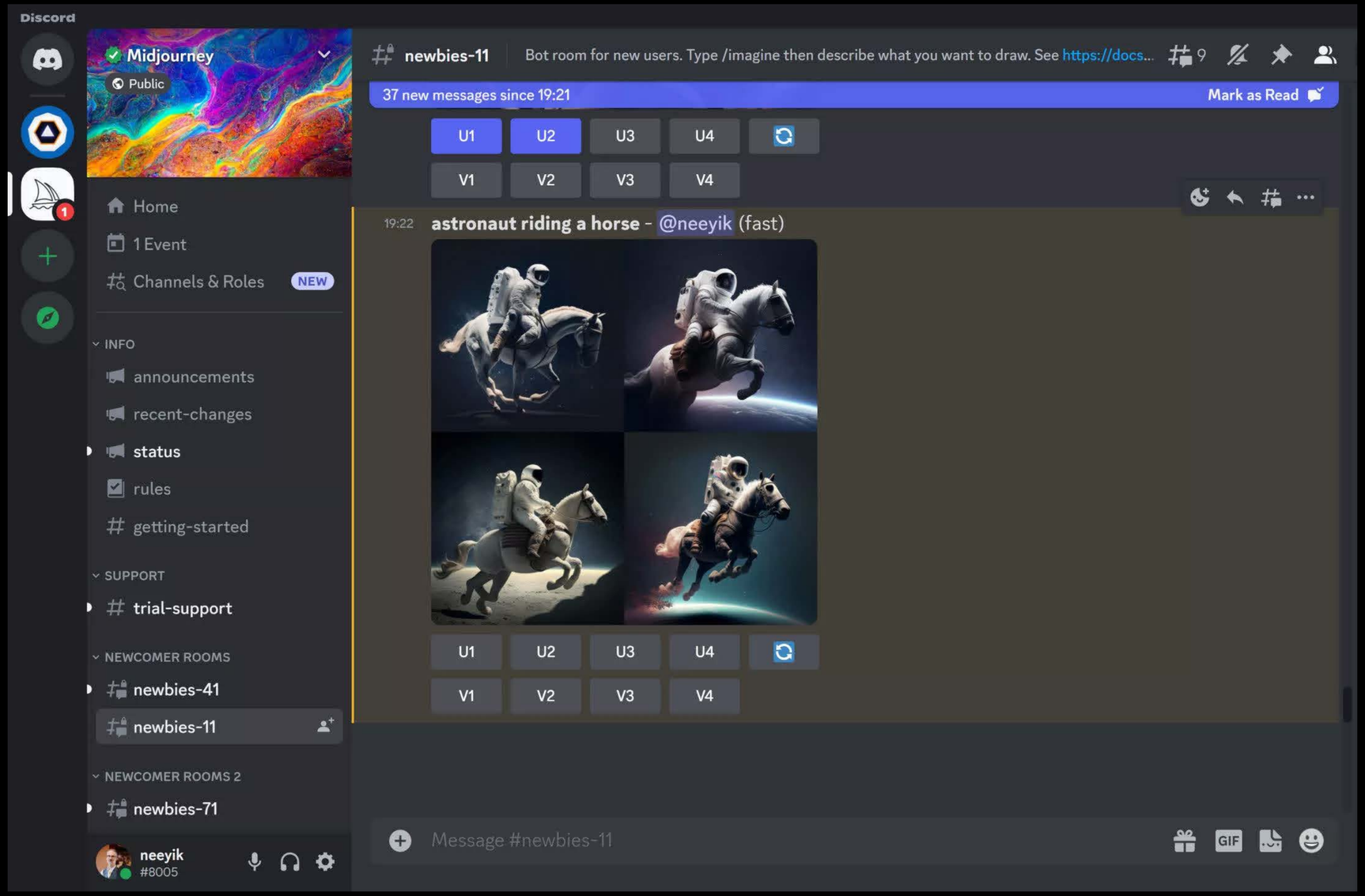
A drawing cat astronaut push random button on a spaceship, professional, UHD, HD, highly detailed, Eiichiro Oda styl

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

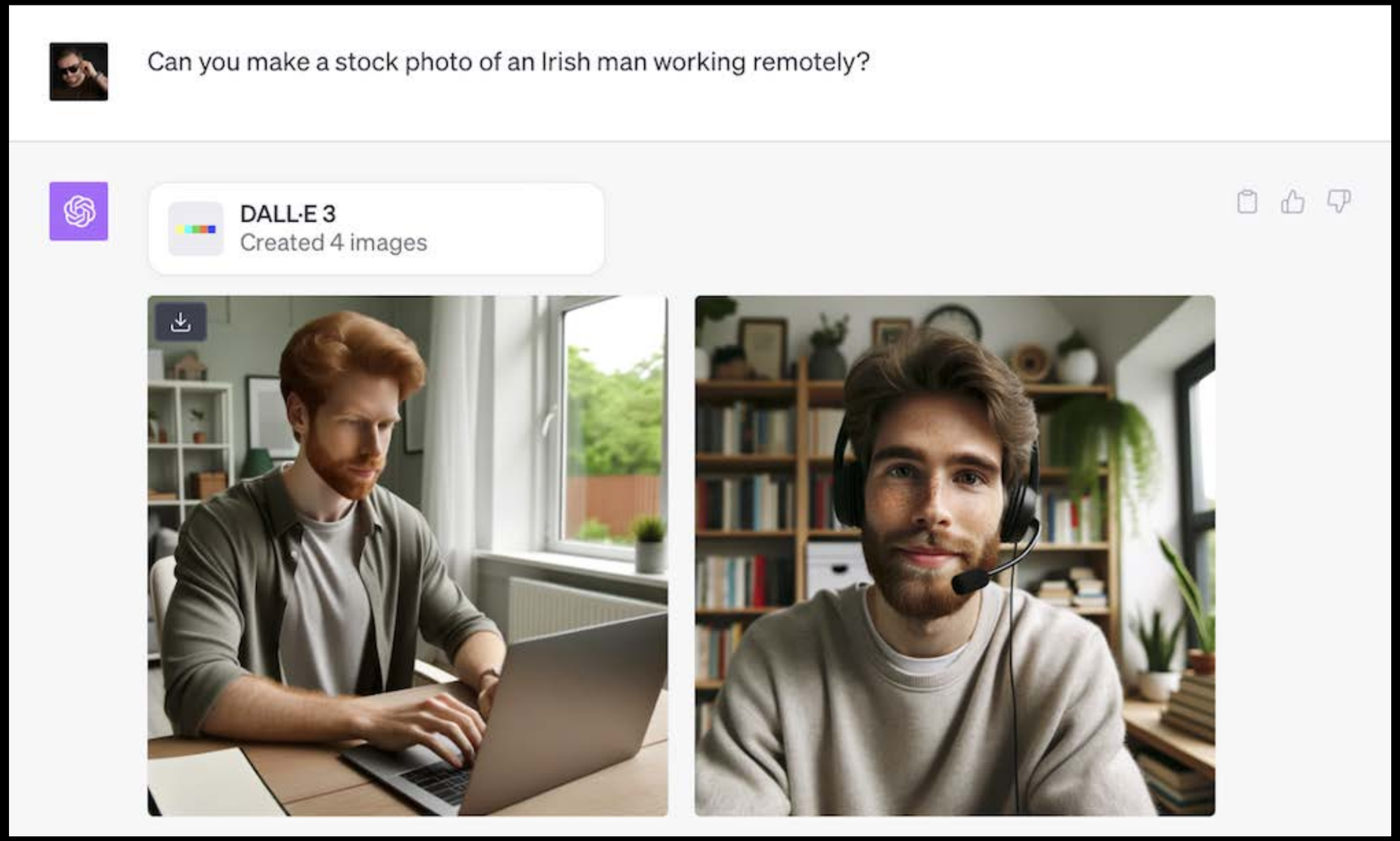
Midjourney



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

AI image generation

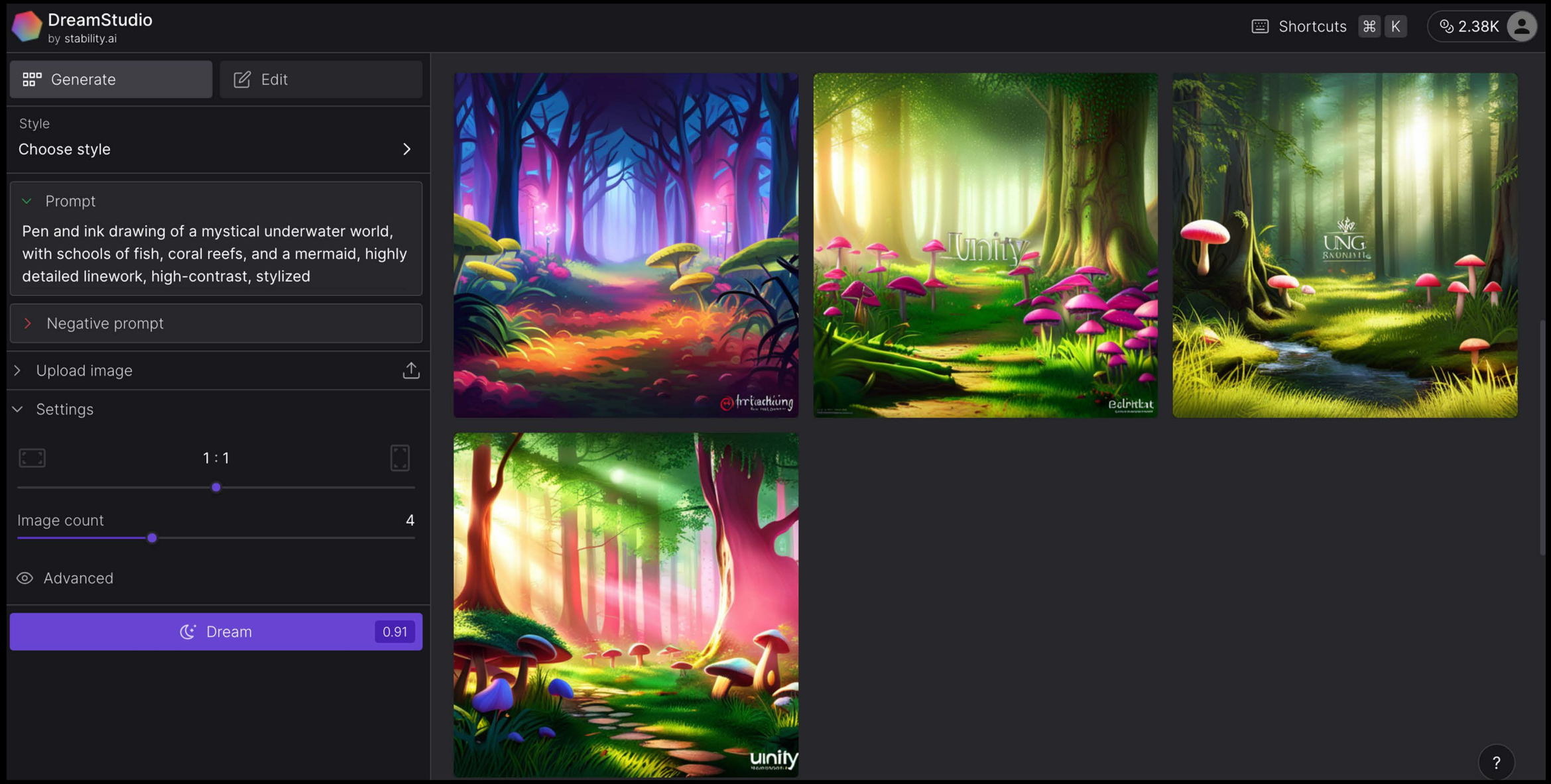
DALL·E



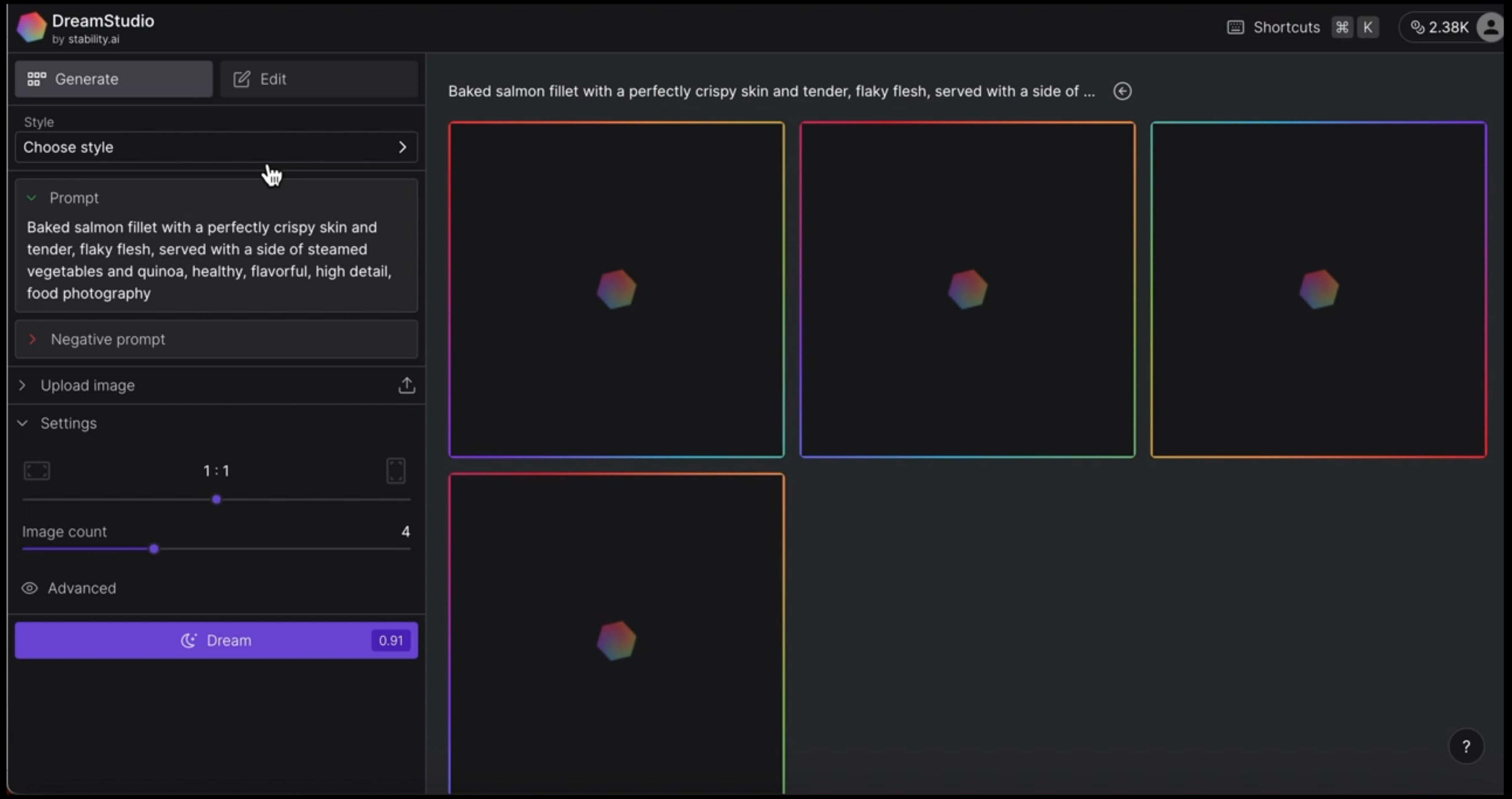
- Create by OpenAI
- Text to image
- Run on ChatGPT
- Cloud-base

AI image generation

Dreamstudio.ai

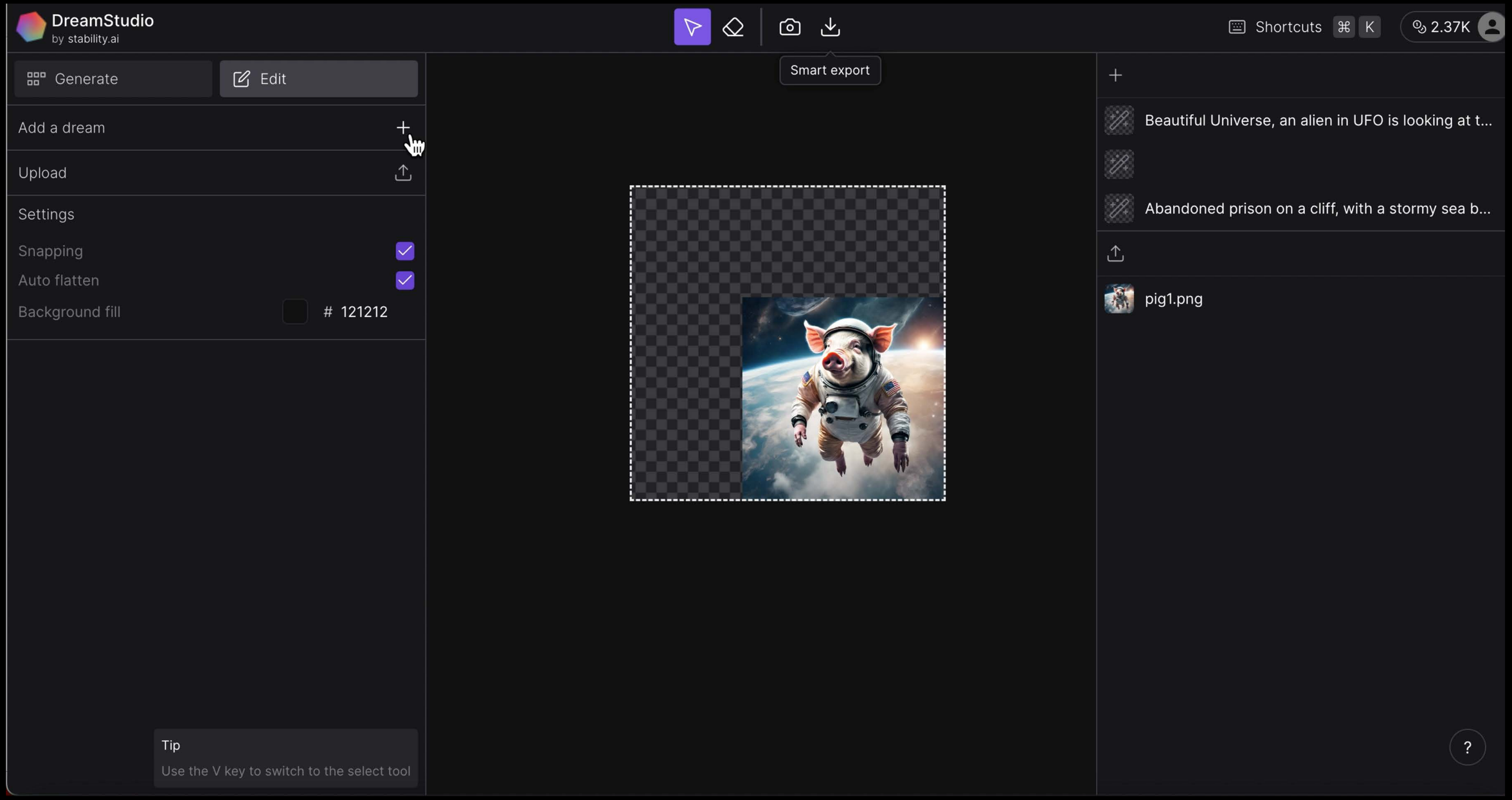


Text-to-image



AI image generation

Image-to-image Outpainting

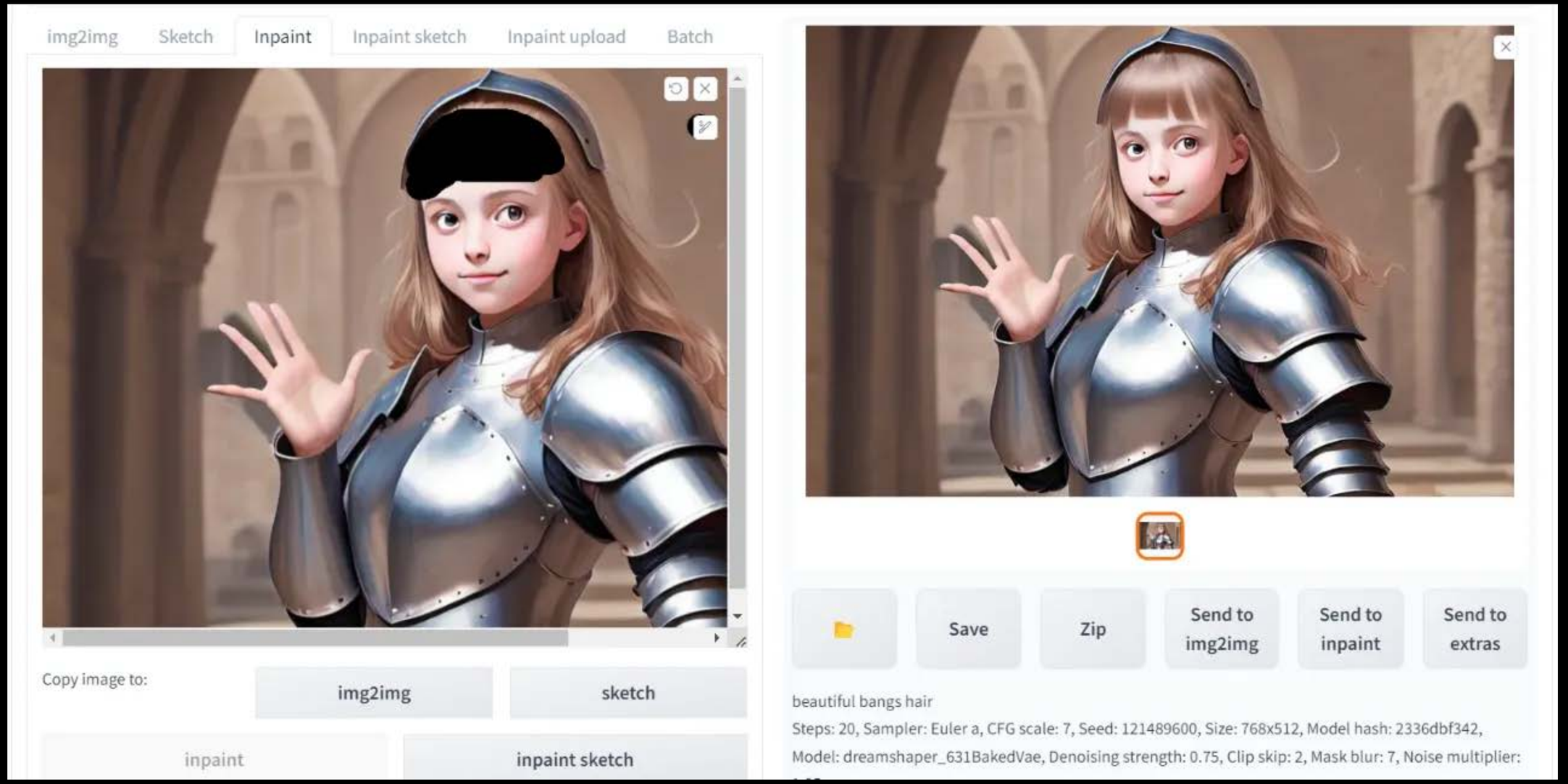


Inpaint

Inpaint

https://aituts.com/inpainting/

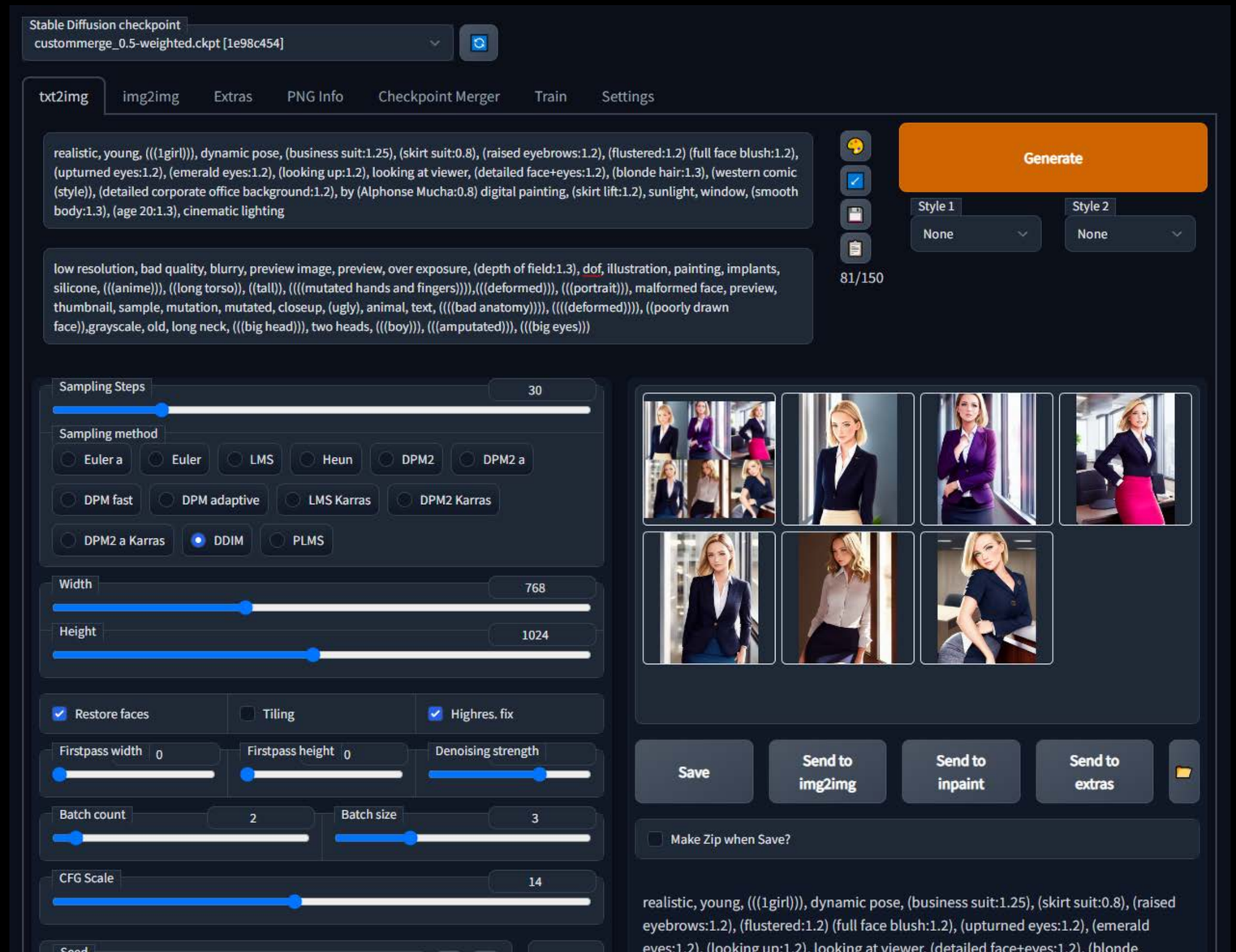
Inpaint



https://aituts.com/inpainting/

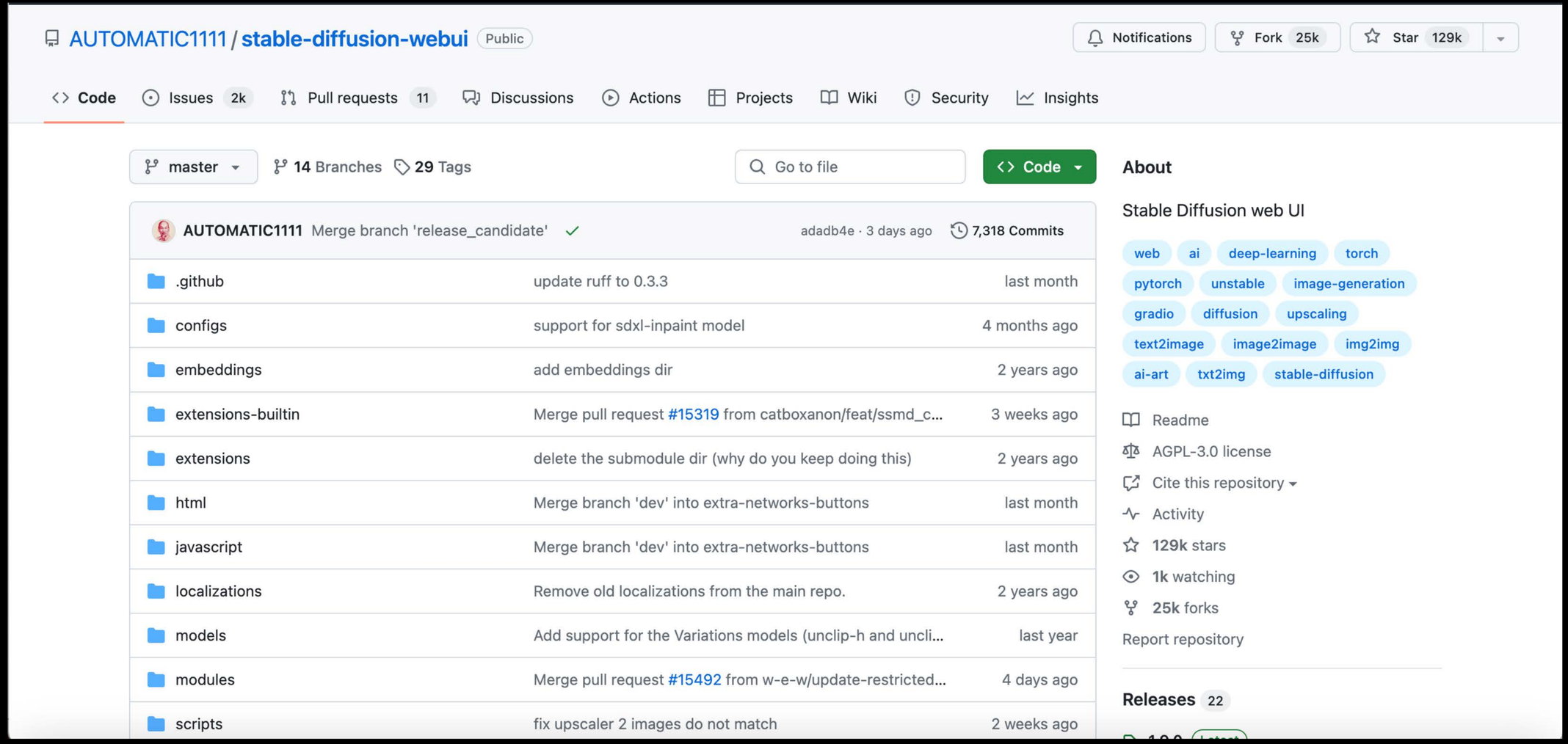
AI image generation

Stable Diffusion



AI image generation

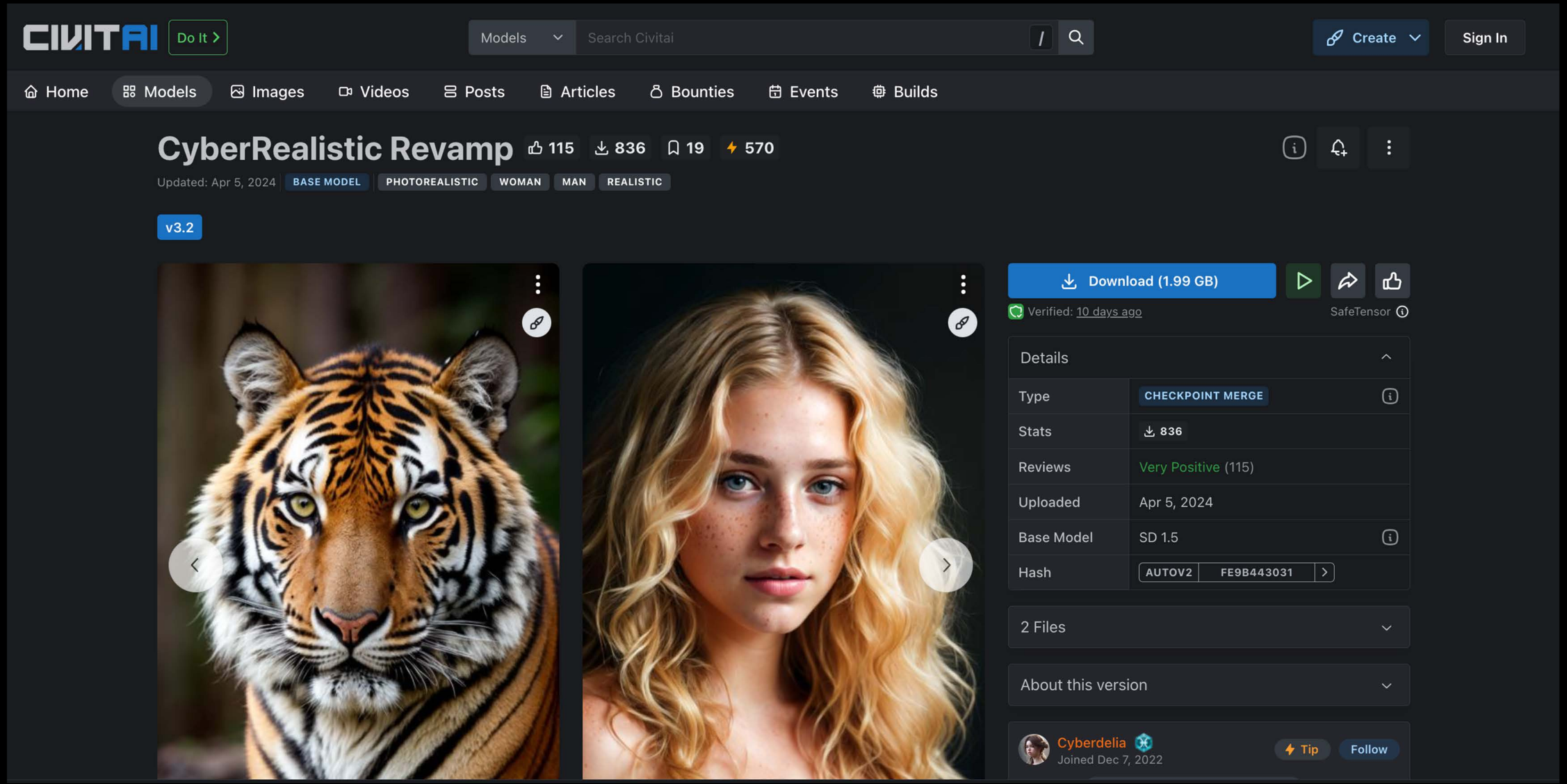
One of the github repository for stable Diffusion



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

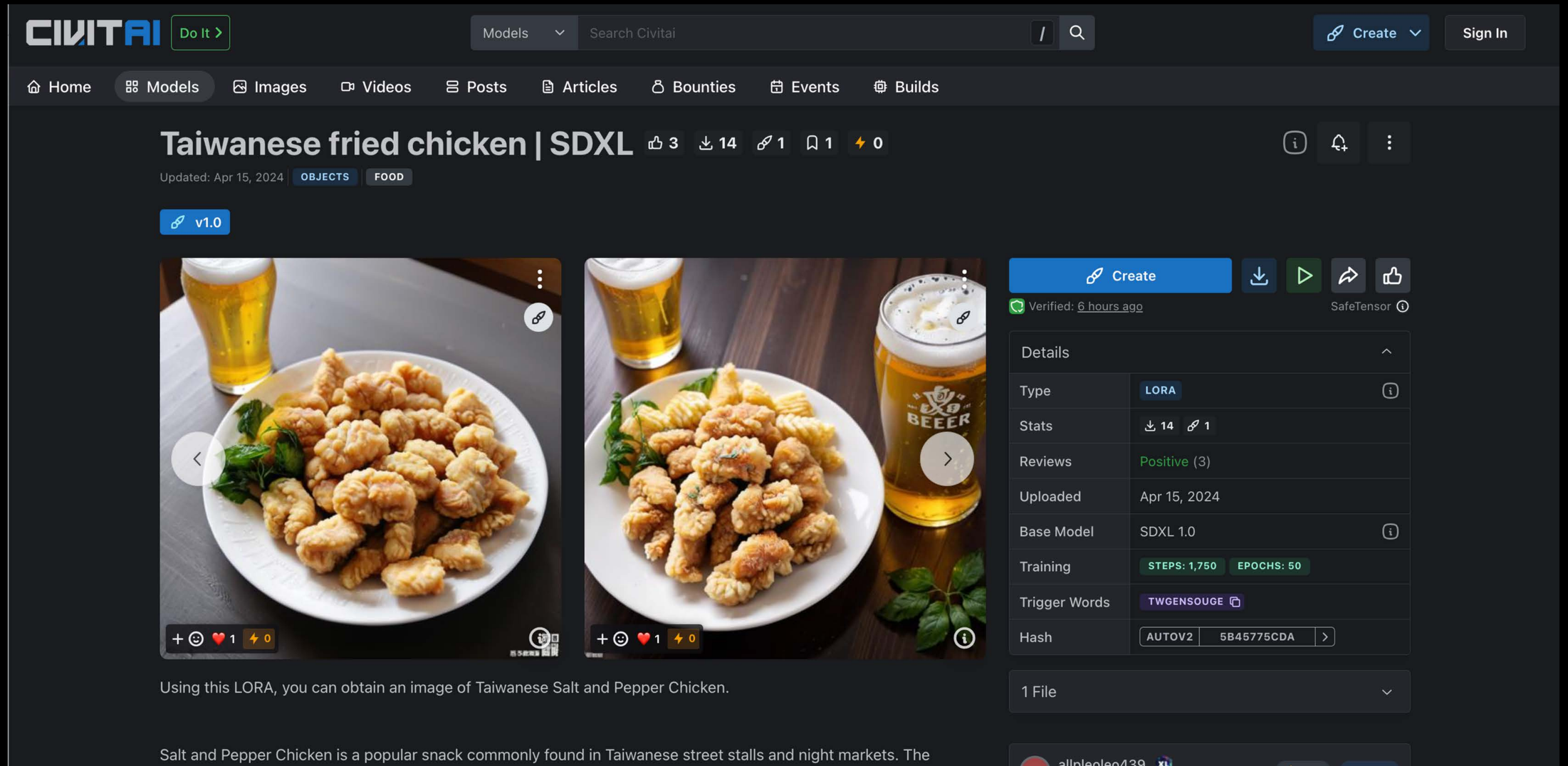
AI image generation

Base Model



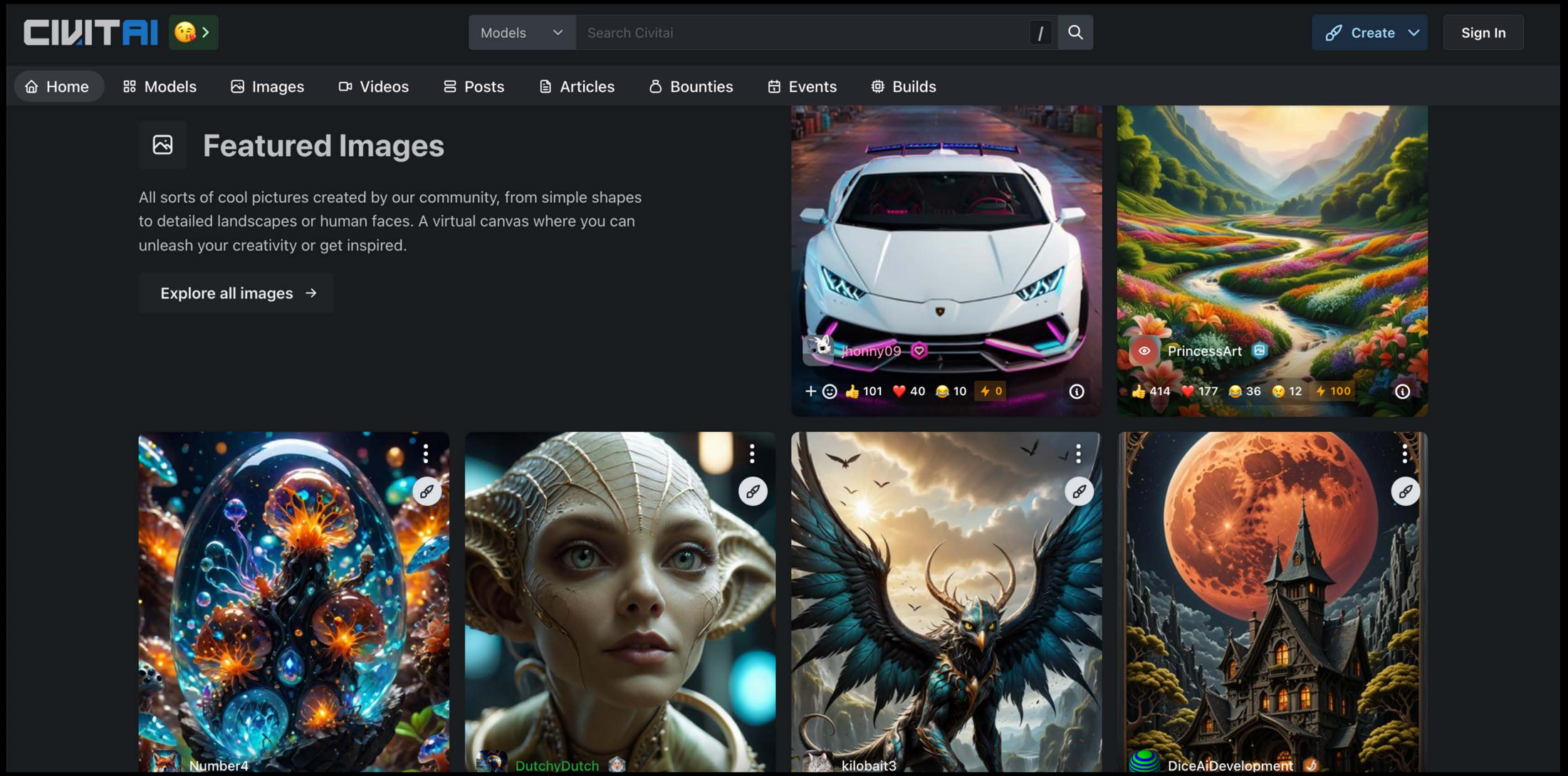
AI image generation

LORA



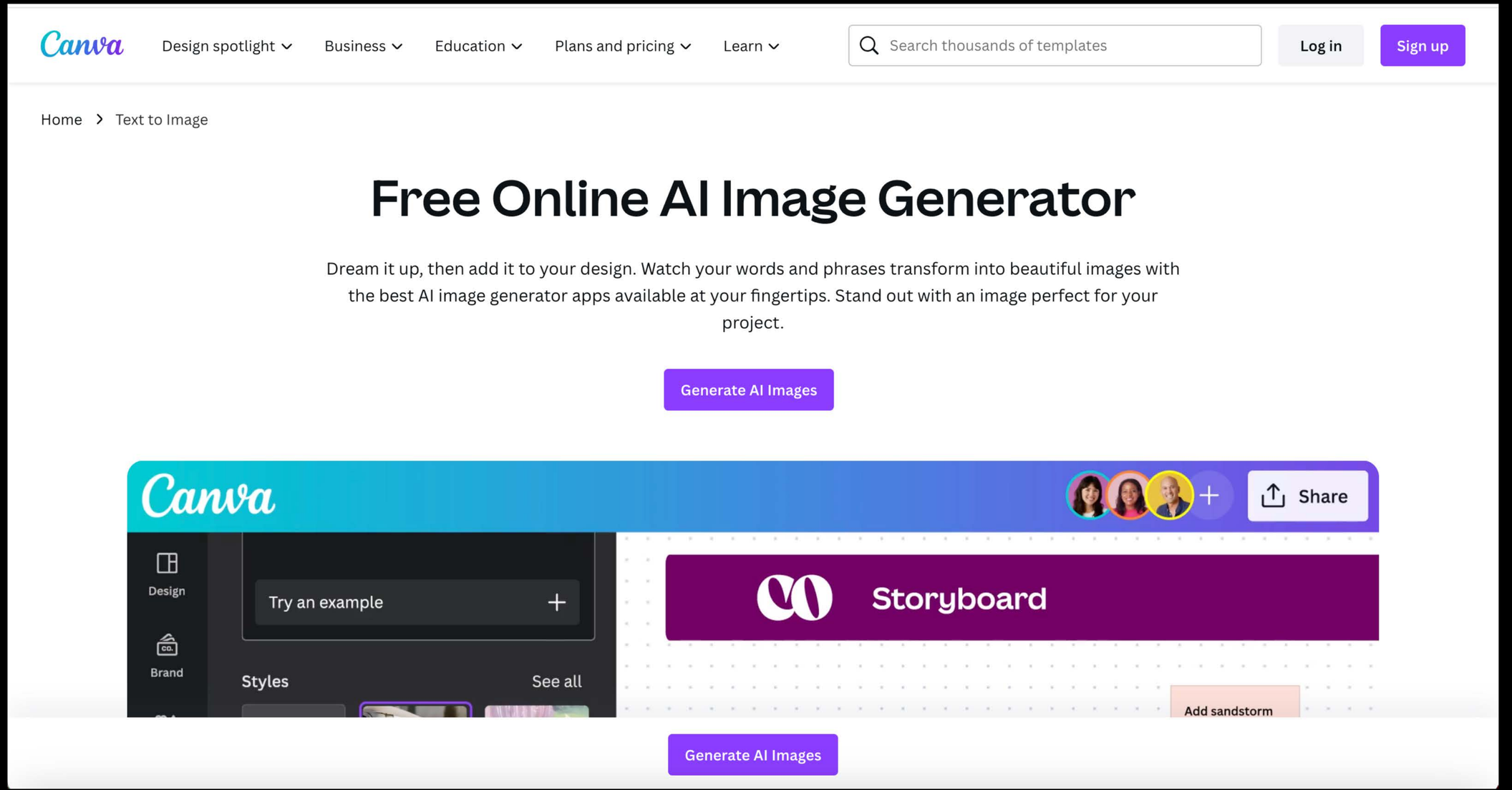
Common websites

https://civitai.com/



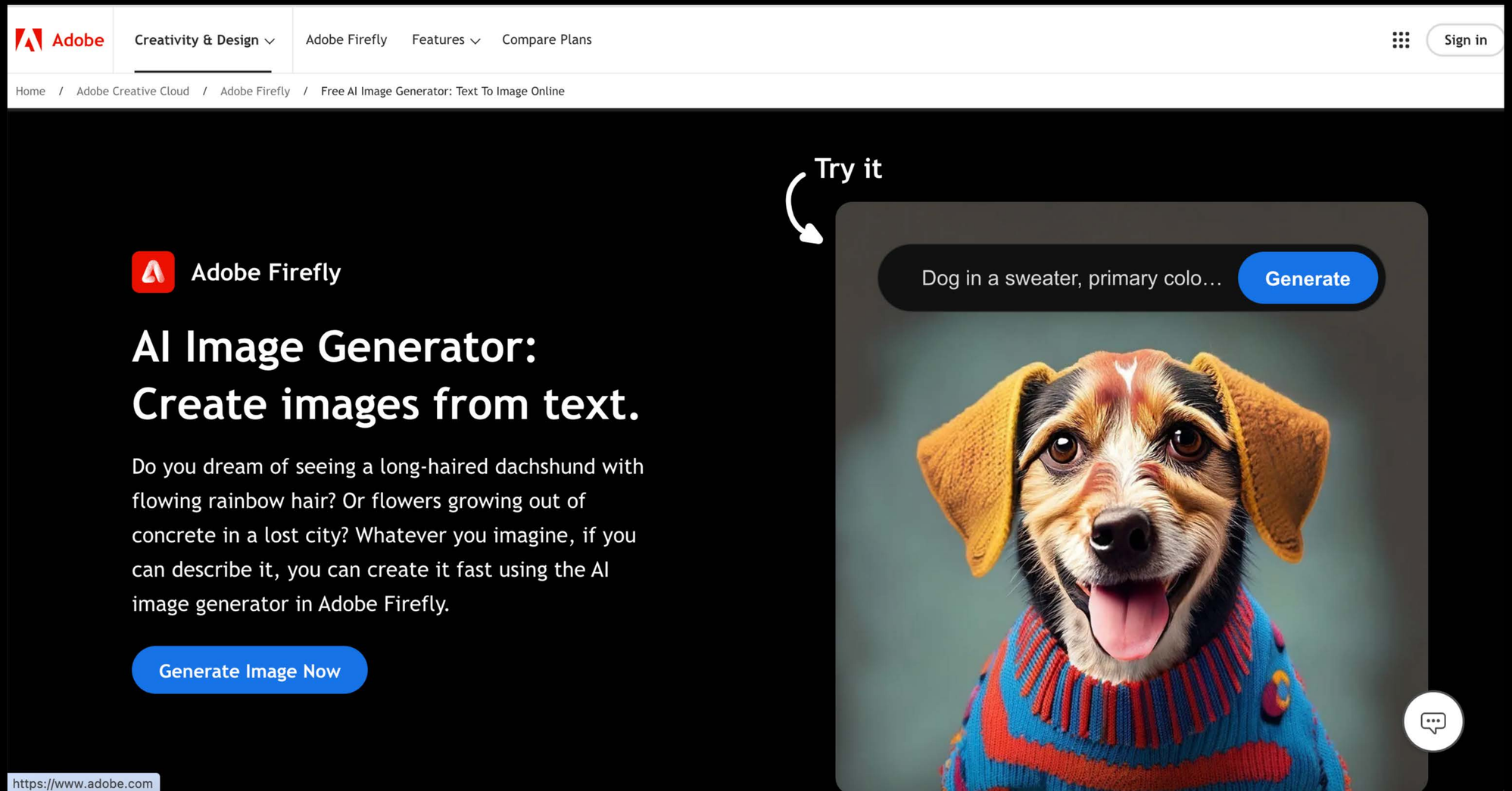
Common websites

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



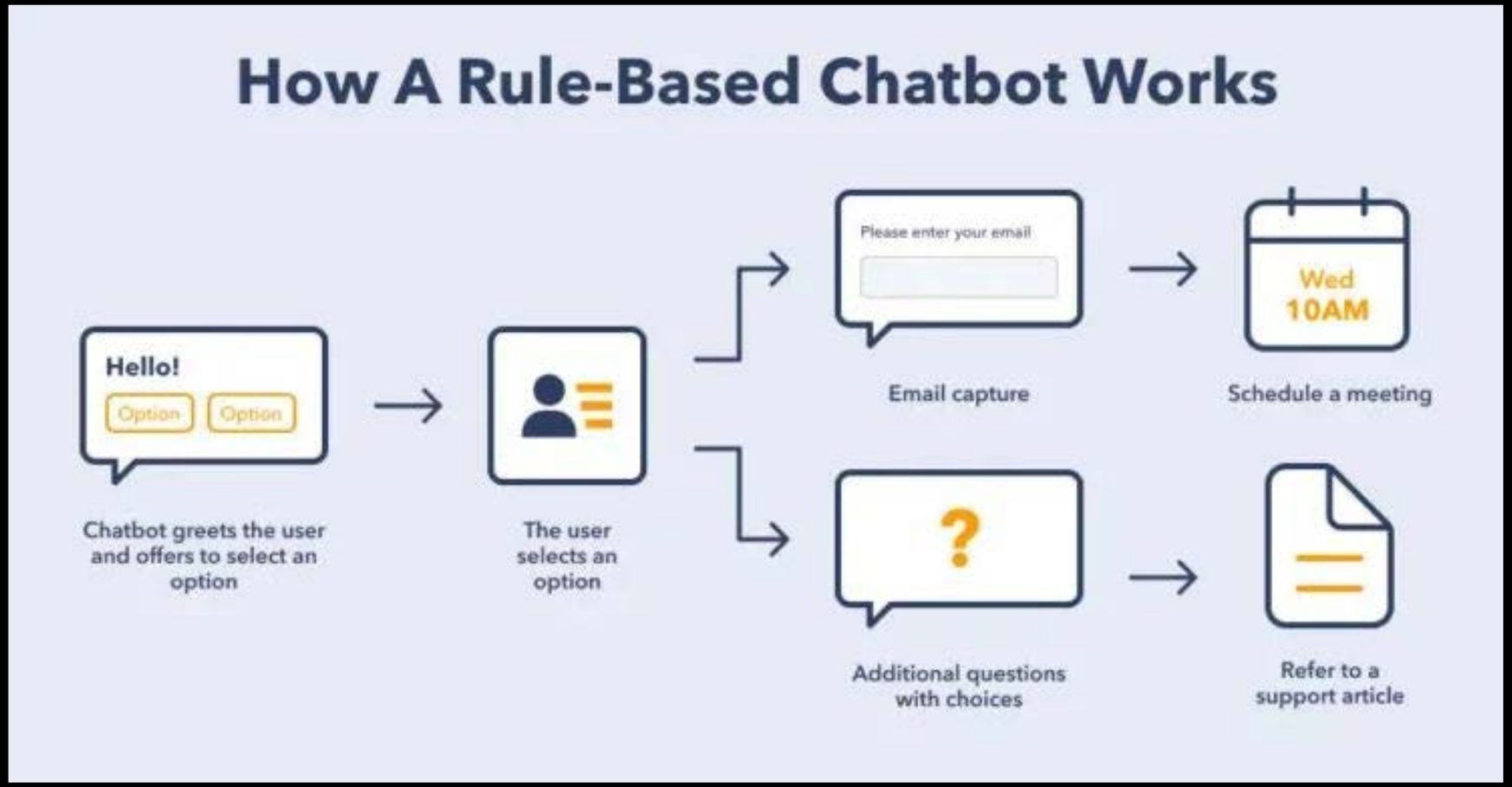
Common websites

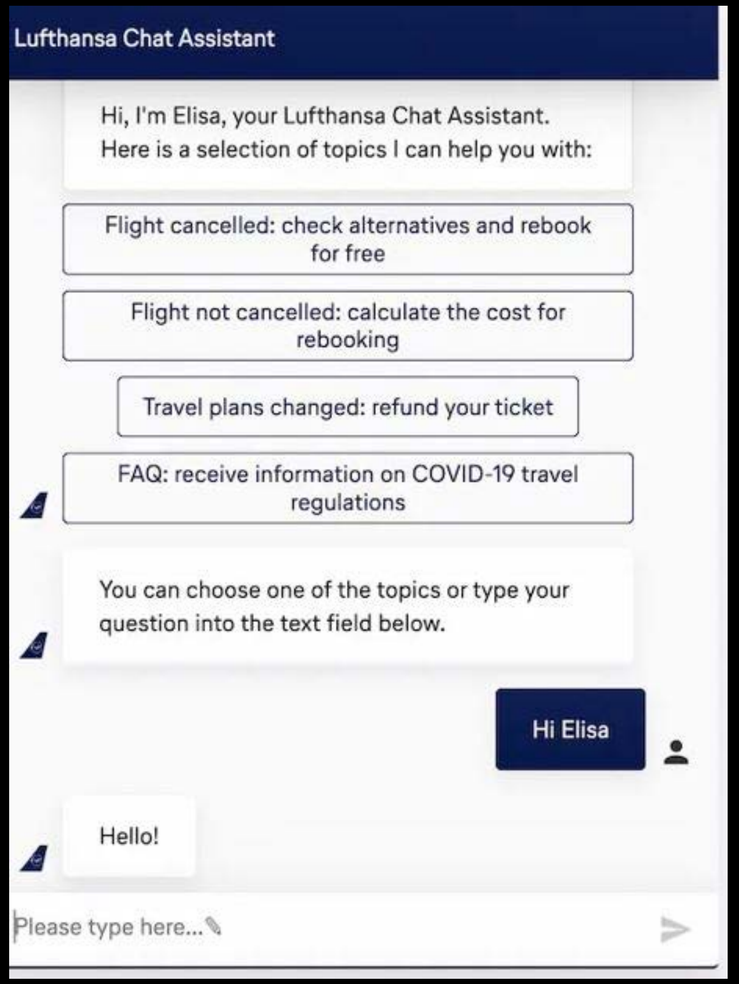
https://www.adobe.com/products/firefly/features/text-to-image.html



AI ChatBots

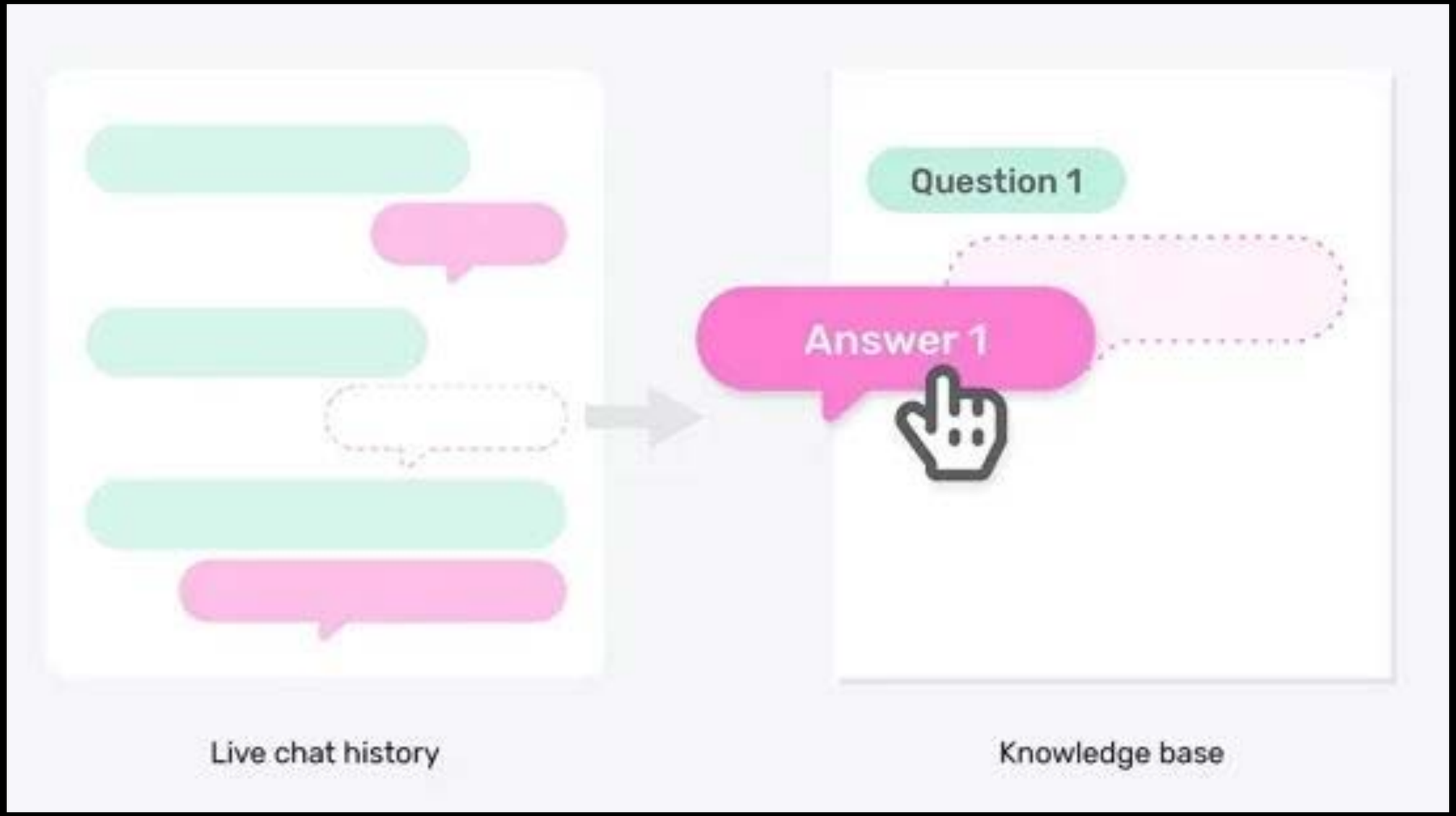
Traditional Chatbots





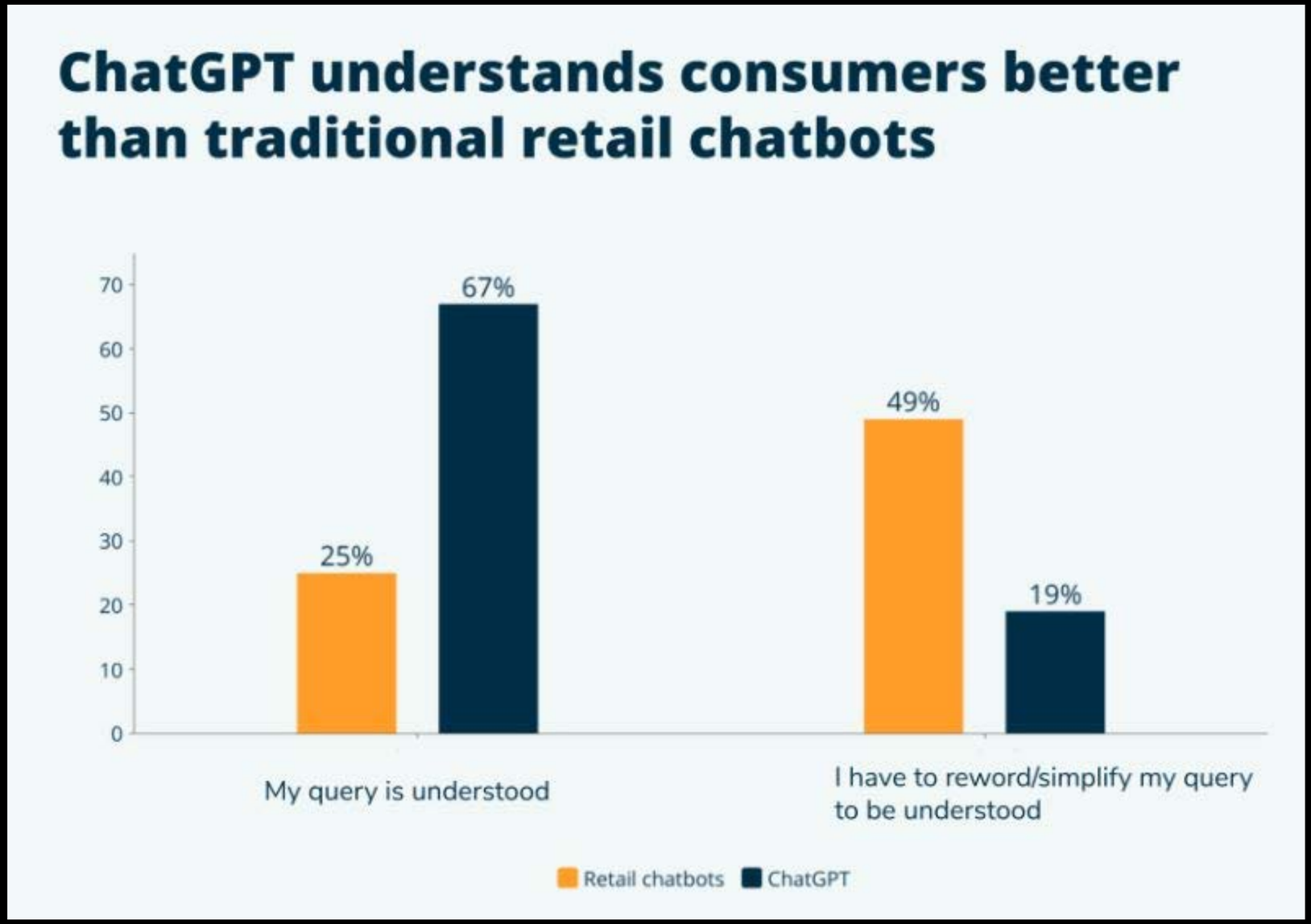
https://www.signitysolutions.com/blog/chatgpt-vs-traditional-chatbots

Generative AI Chatbots



https://www.signitysolutions.com/blog/chatgpt-vs-traditional-chatbots

Generative AI Chatbots



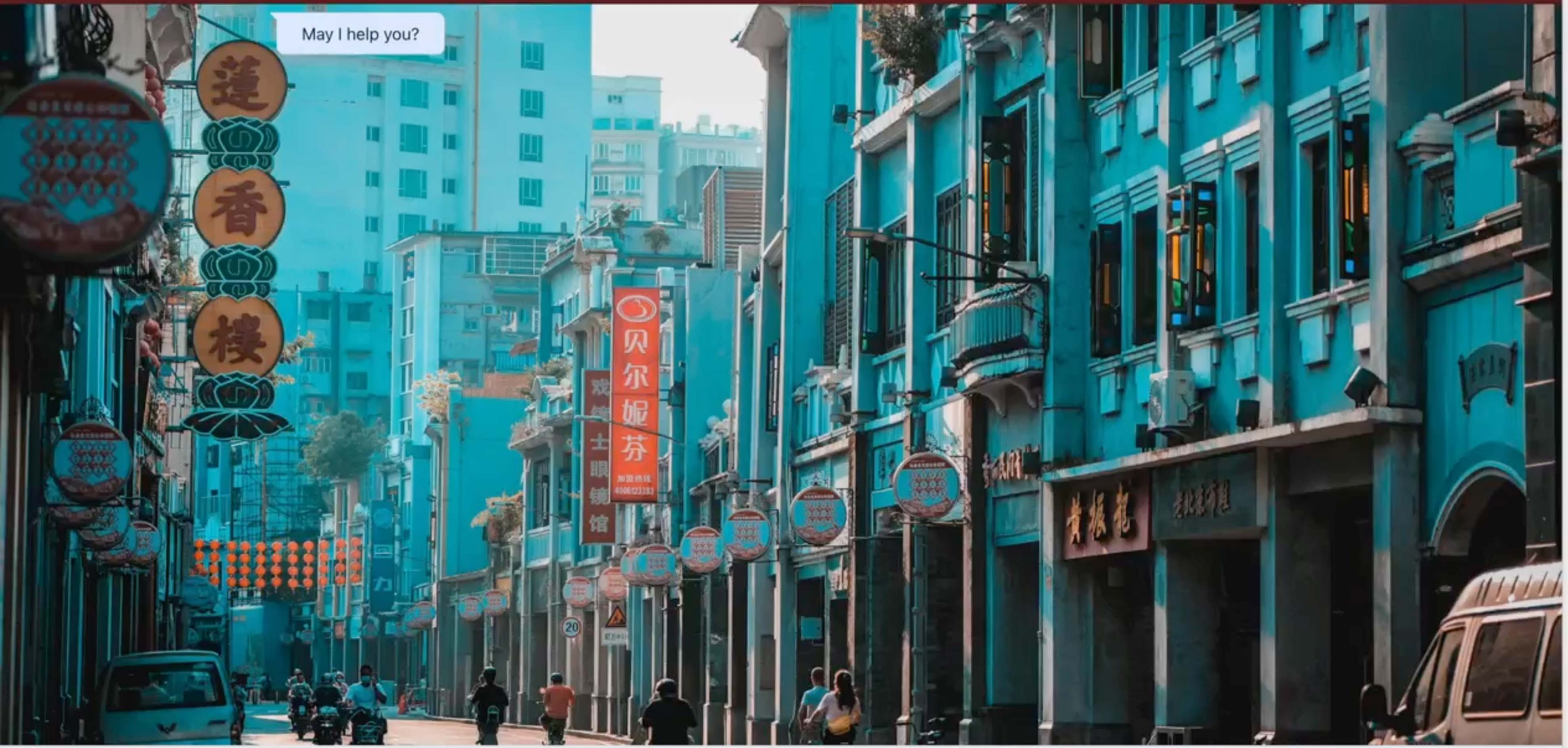
https://www.signitysolutions.com/blog/chatgpt-vs-traditional-chatbots

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

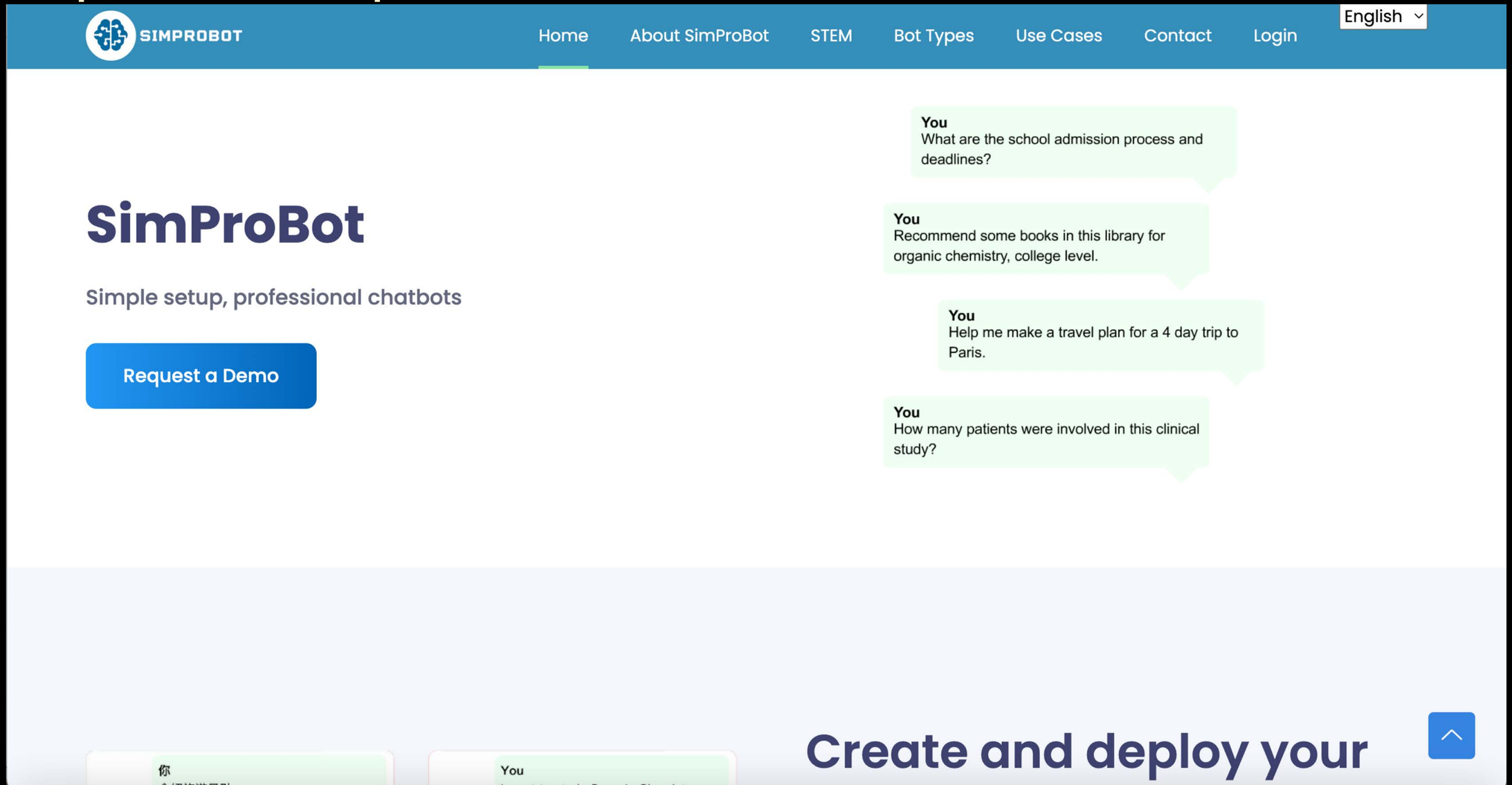
https://www.signitysolutions.com/blog/chatgpt-vs-traditional-chatbots

Great Bay Area Intro



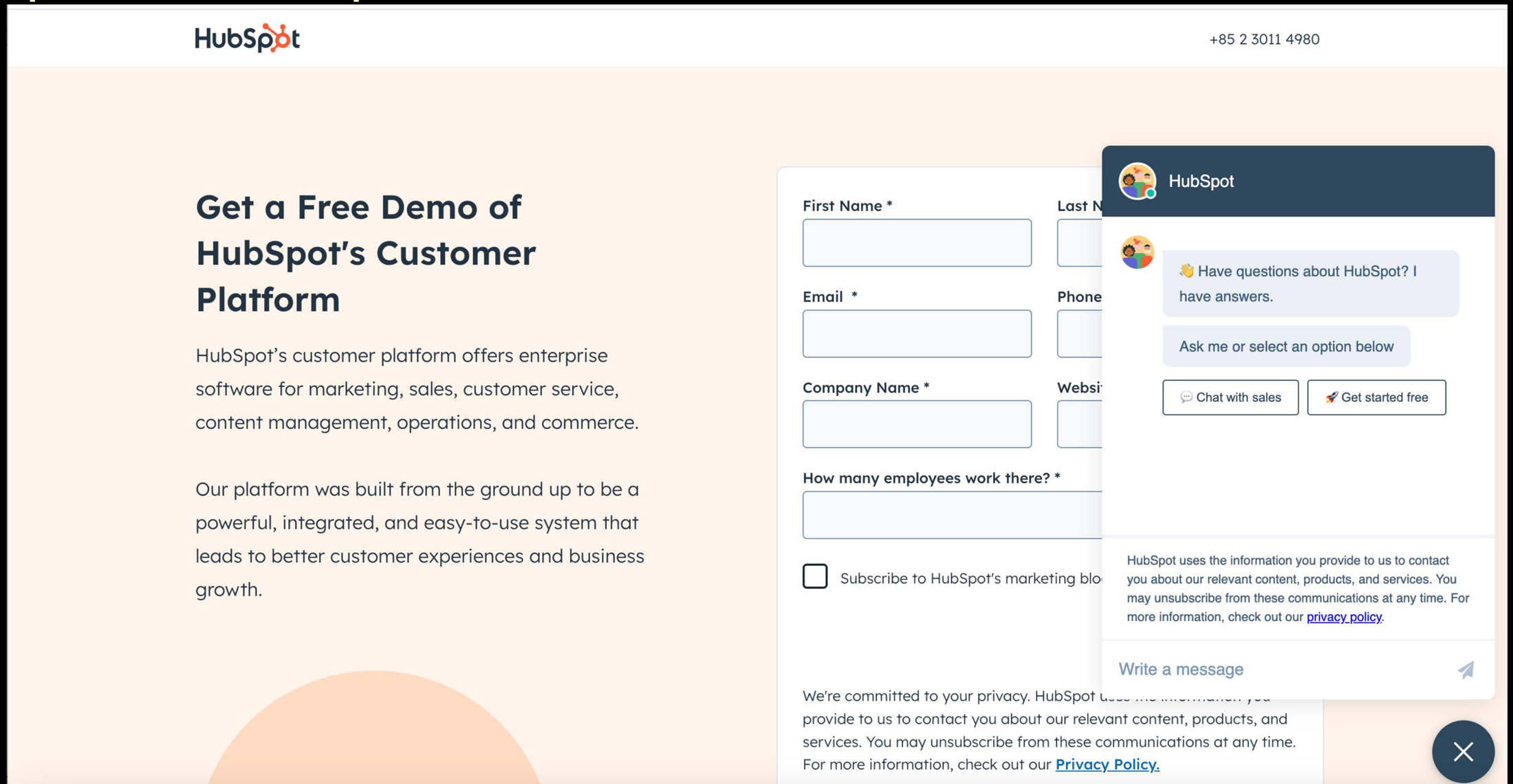
Common website

https://www.simprobot.com/



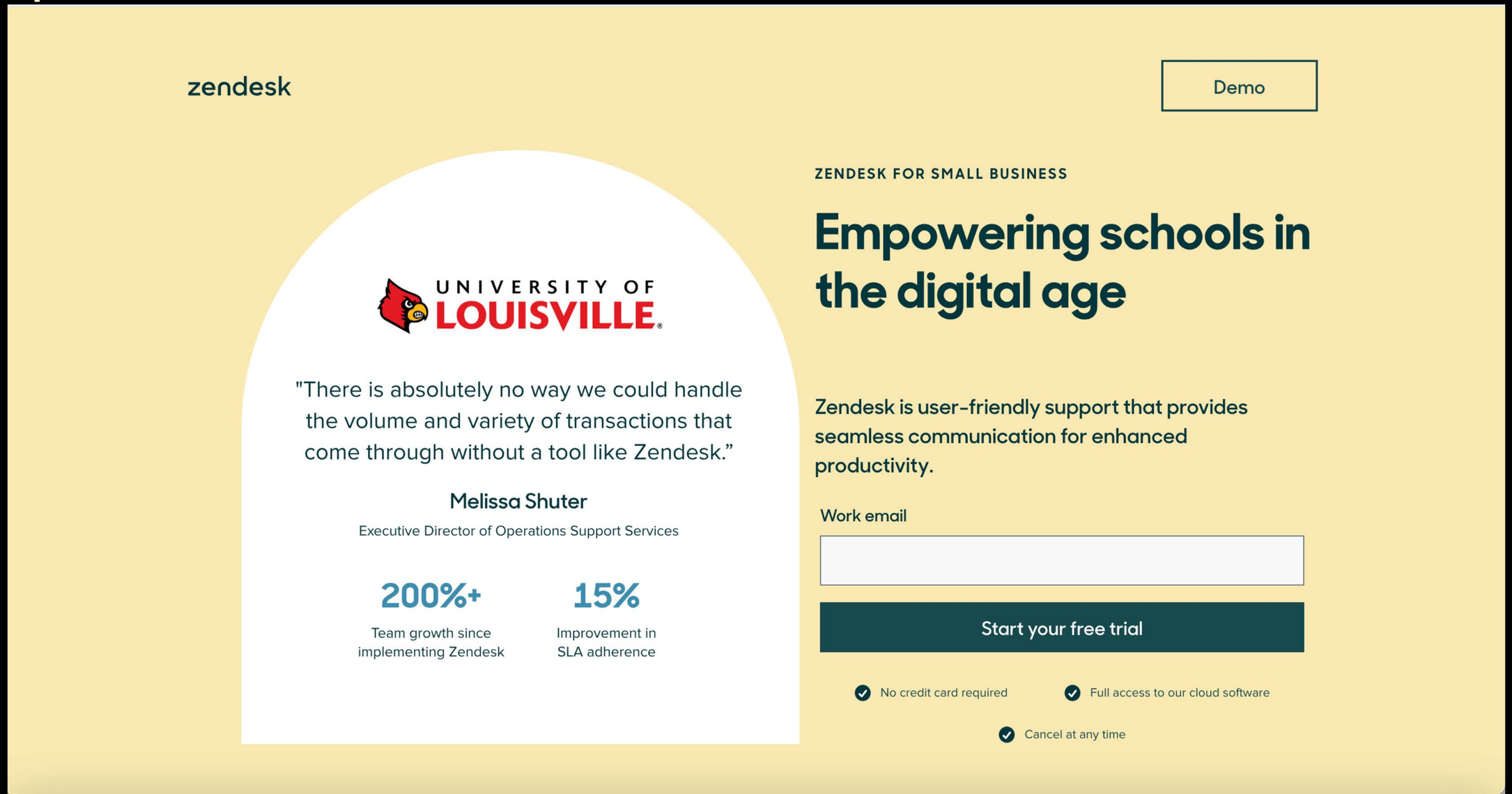
Common website

https://www.hubspot.com/



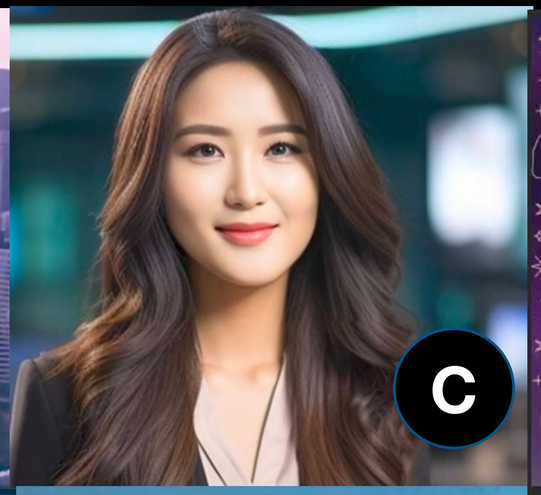
Common website

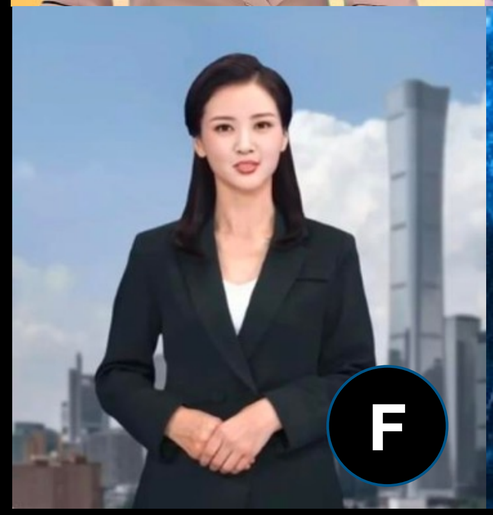
https://www.zendesk.com/

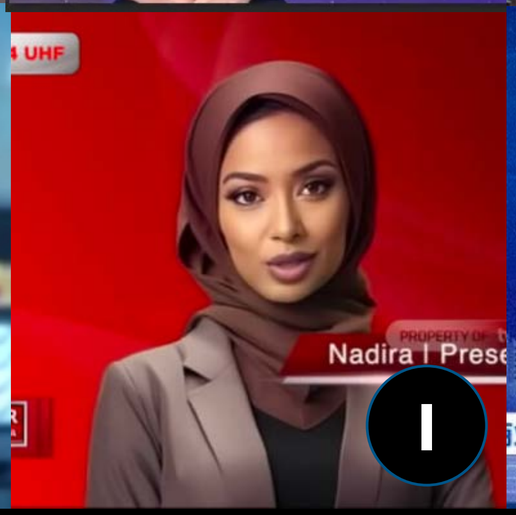


AI Anchors

Who are AI Anchors?







AI Anchor Example

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

AI Anchor Example

https://www.youtube.com/watch?v=YSGS3gvOabI

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.

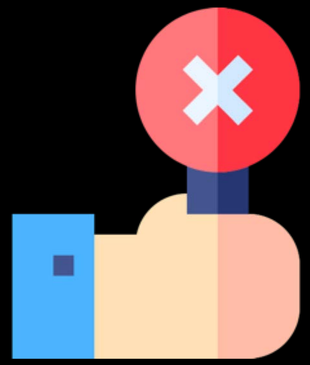
AI Anchor Example

Source: iflytek

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

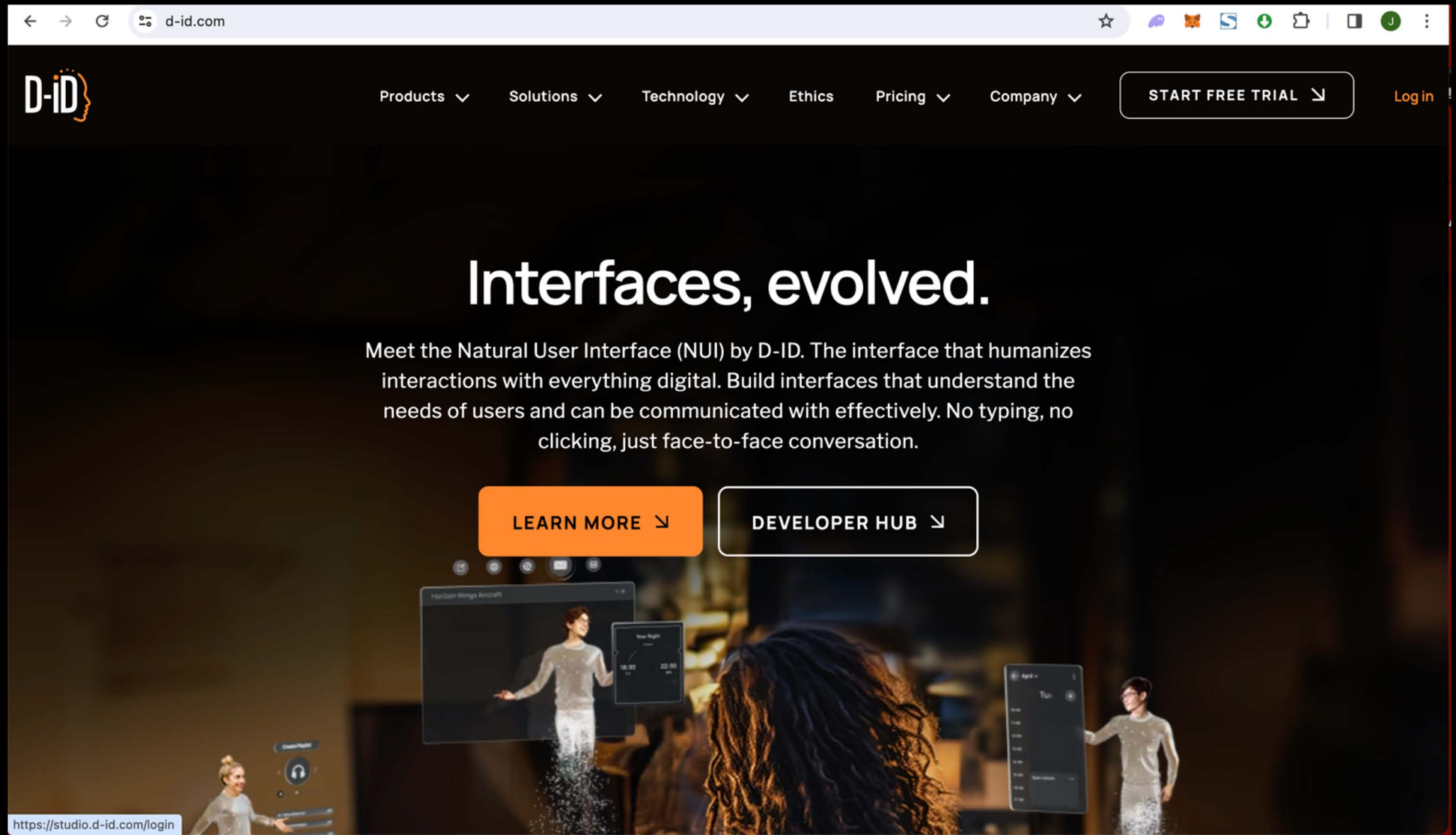
Different personnel involved in live-action shooting, including anchors, have the potential to make errors.

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.

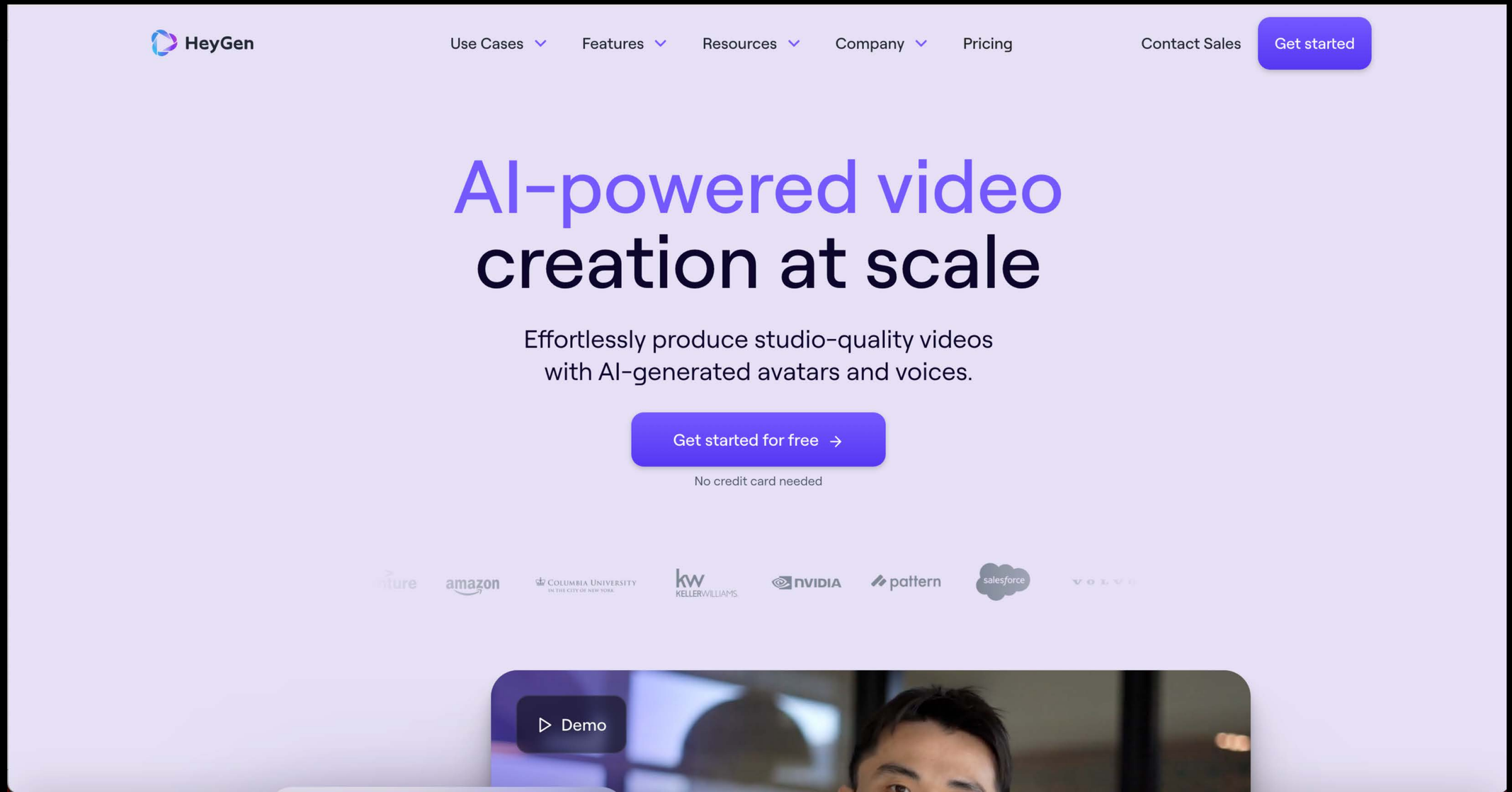
Common websites

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



Common websites

https://www.heygen.com/




On-premise AI deployment

Forbes

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.

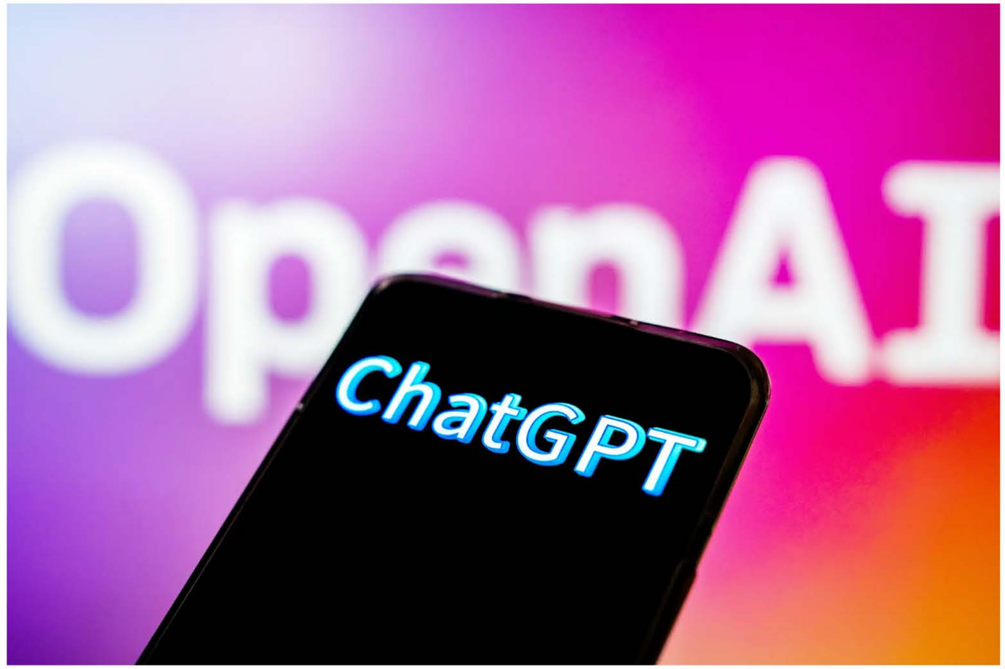
https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/?sh=501995e76078

BUSINESS INSIDER

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-ai-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

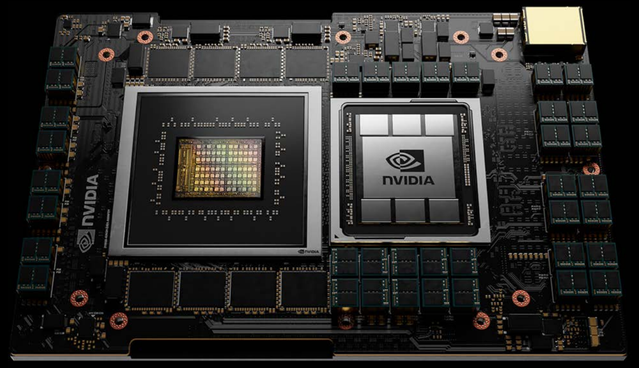
PowerPoint

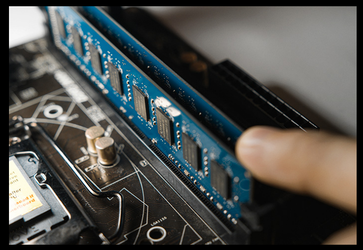
Excel

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

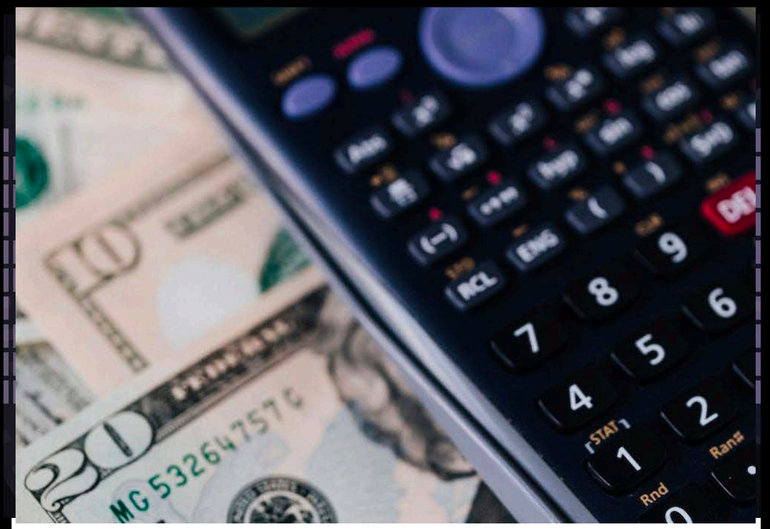
Faster response and training:
On-premises AI offers quicker conversation generation and training times.

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

ChatGPT

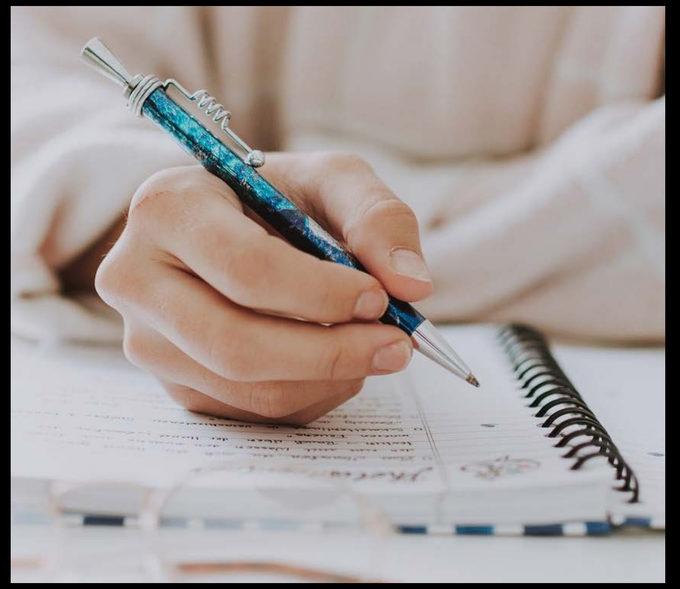
- Data Security
- Low Latency
- Customization
- Regulatory Compliance
- Cost Efficiency

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.

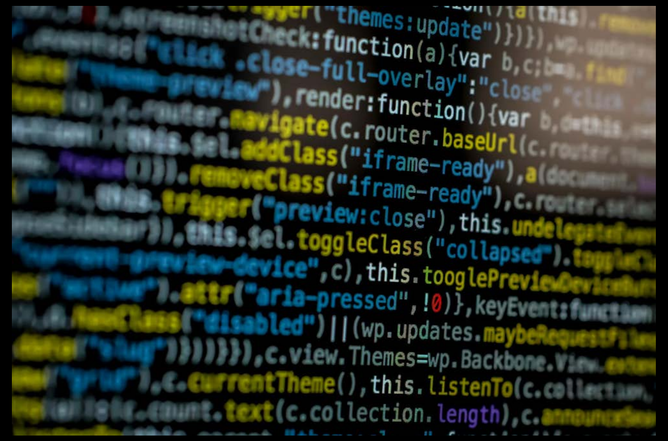


B. Summarizing engineering information:

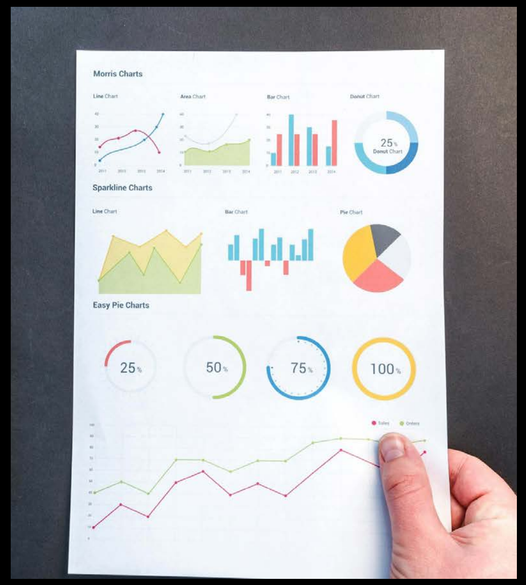
- Summarize extensive collections of documents.

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

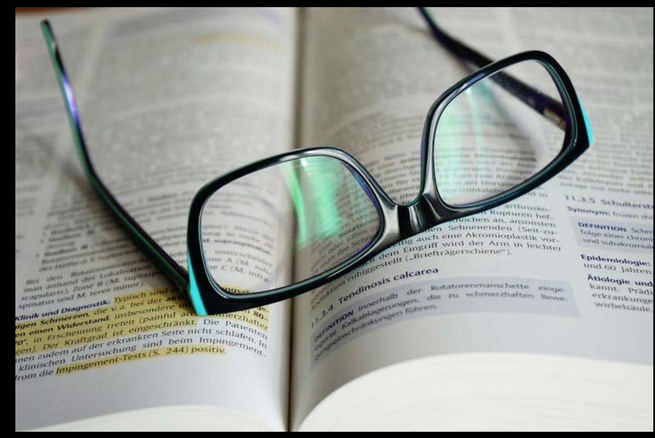


D. Creating written content:

- 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

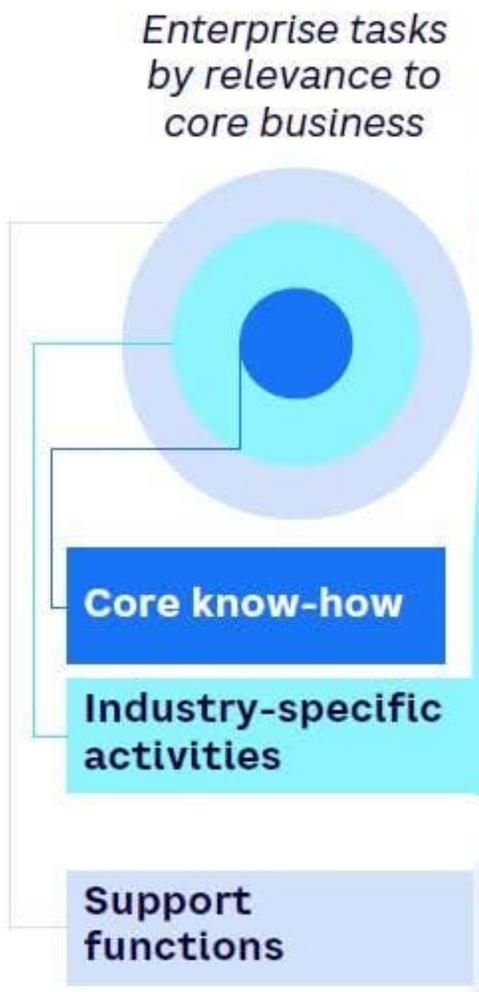
https://www.mckinsey.com/~/media/mckinsey/featured%20insights/mckinsey%20explainers/whats%20the%20future%20of%20generative%20ai%20an%20early%20view%20in%2015%20charts/svgz-futuregenaicollection-ex4.svgz?cq=50&cpy=Center

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

https://www.mckinsey.com/~/media/mckinsey/featured%20insights/mckinsey%20explainers/whats%20the%20future%20of%20generative%20ai%20an%20early%20view%20in%2015%20charts/svgz-futuregenaicollection-ex4.svgz?cq=50&cpy=Center

The impact of Generative-AI in different industry



Media	Retail	Healthcare (HC) & life sciences (LS)	Energy & utilities	Financial services	Consumer packaged goods
Content generation (**Scribe**), audience targeting (**Analyst**)	Customer care (**Librarian**), customer segmentation, price elasticity, merchandising analytics (**Analyst**)	HC: Customer care (**Librarian**), medical report generation (**Scribe**) LS: Drug discovery (**Scientist**)	Resource management, demand forecasting (**Analyst**), nuclear fission (**Scientist**)	Portfolio customization (**Analyst**), customer care (**Librarian**)	Customer segmentation (**Analyst**), product ideation (**Scribe**), ad generation (**Scribe**)

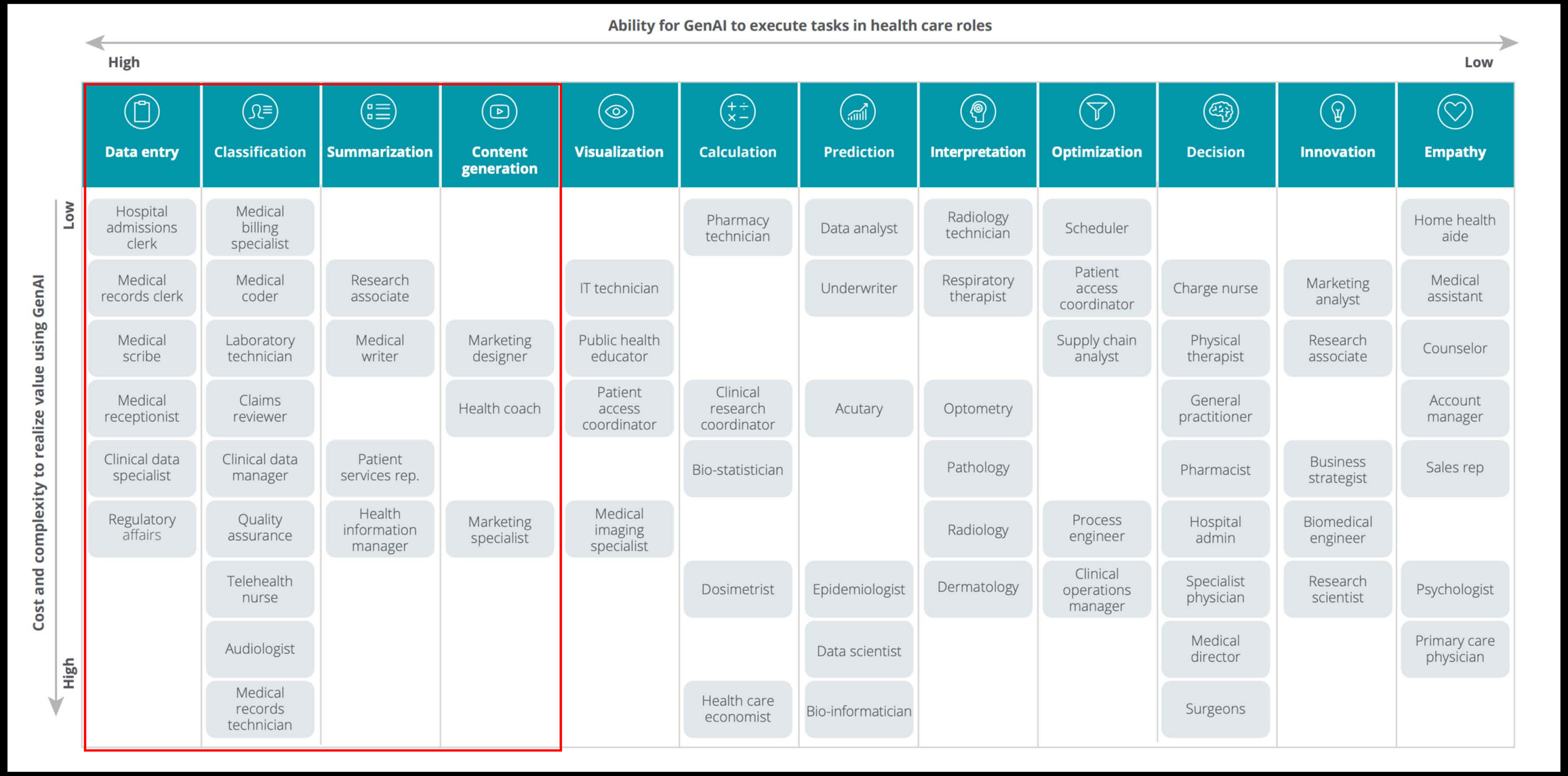
Aerospace & defense	Industrial goods & services	Travel & transport	Telecom	Automotive	Public services
Production floor robotics, supply chain management (**Engineer**)	Production floor robotics, supply chain management (**Engineer**)	Transit flow prediction, customized journey planning (**Analyst**)	Customer care (**Librarian**), network planning (**Engineer**), field service efficiency (**Analyst**)	Production floor robotics, supply chain management (**Engineer**)	Citizen service (**Librarian**), administrative document production (**Scribe**)

Enterprise AI	Finance	Human resources	Knowledge management
	Financial planning (**Analyst**), billing (**Scribe**)	Recruiting, onboarding, training, performance management (**Librarian**, **Scribe**)	Information retrieval, regulatory question & answering (**Librarian**)

Source: Arthur D. Little

https://www.adlittle.com/be-en/insights/report/generative-artificial-intelligence-toward-new-civilization

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@aiilog.com