LLMSec

Security in the Era of LLMs and GenAI

Natalie Pistunovich | **@NataliePis**

Natalie Pistunovich @NataliePis

OpenAI Developer Ambassador

& Independant Consultant

GopherCon Europe

Golang Berlin

Connta

Podcast

SAI: The Security and AI Podcast

Natalie Pistunovich & Ivan Kwiatkowski

Following • • •

Latest episode

Episode 7 - OpenAI cybersecurity grant program

In this episode, Natalie and Ivan, in a conference call with members of GenAI, discuss OpenAI's cybersecurity grant program, and go over the suggestions made to applicants by the company on its webpage.

Oct 1 · 1 hr 9 min

All Episodes

Newest to Oldest

Episode 7 - OpenAI cybersecurity grant program

In this episode, Natalie and Ivan, in a conference call with members of GenAI, discuss OpenAI's cybersecurity grant program, and go over the suggestions made to applicants by the company on its webpage.

Oct 1 · 1 hr 9 min

About

Welcome to the SAI podcast, where intellectual curiosity and technical expertise meet.
Natalie is an OpenAI Developer Ambassador and Ivan is a Senior Security Researcher at Kaspersky.
Together we ask questions and learn, each one bringing their expertise. We're covering the different cross sections: how can AI be used in cybersecurity, how can AI be more secure, what attacks can be done, how to prepare, etc.
All the topics are introduced at a high level and occasionally we dive deeper in.

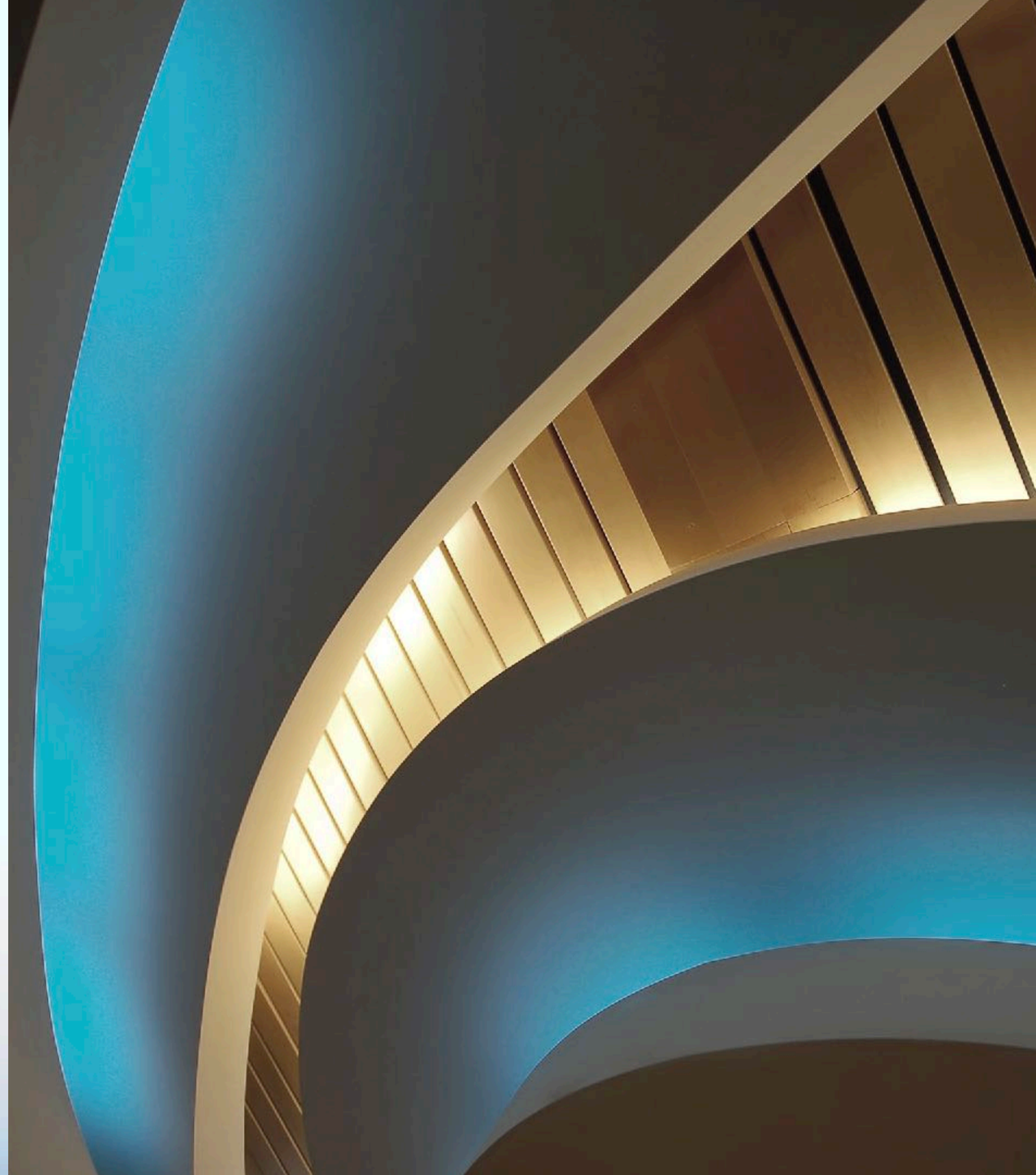
Show less

5 ★ (2)

Technology

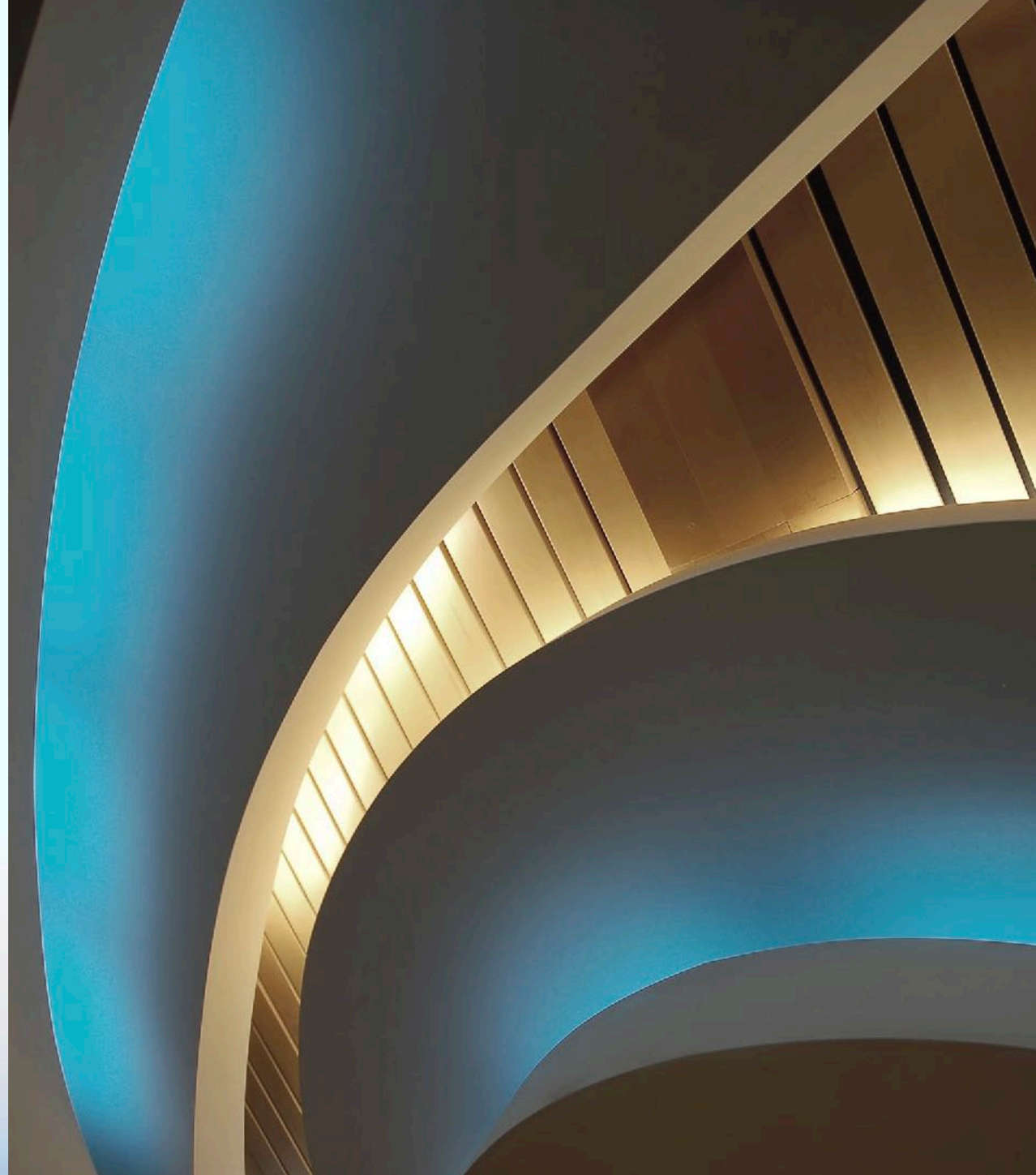
Agenda

- LLMs & ChatGPT
- Voice
- Audio
- Image
- Video
- Multi-Modal
- Autonomous AI Agents



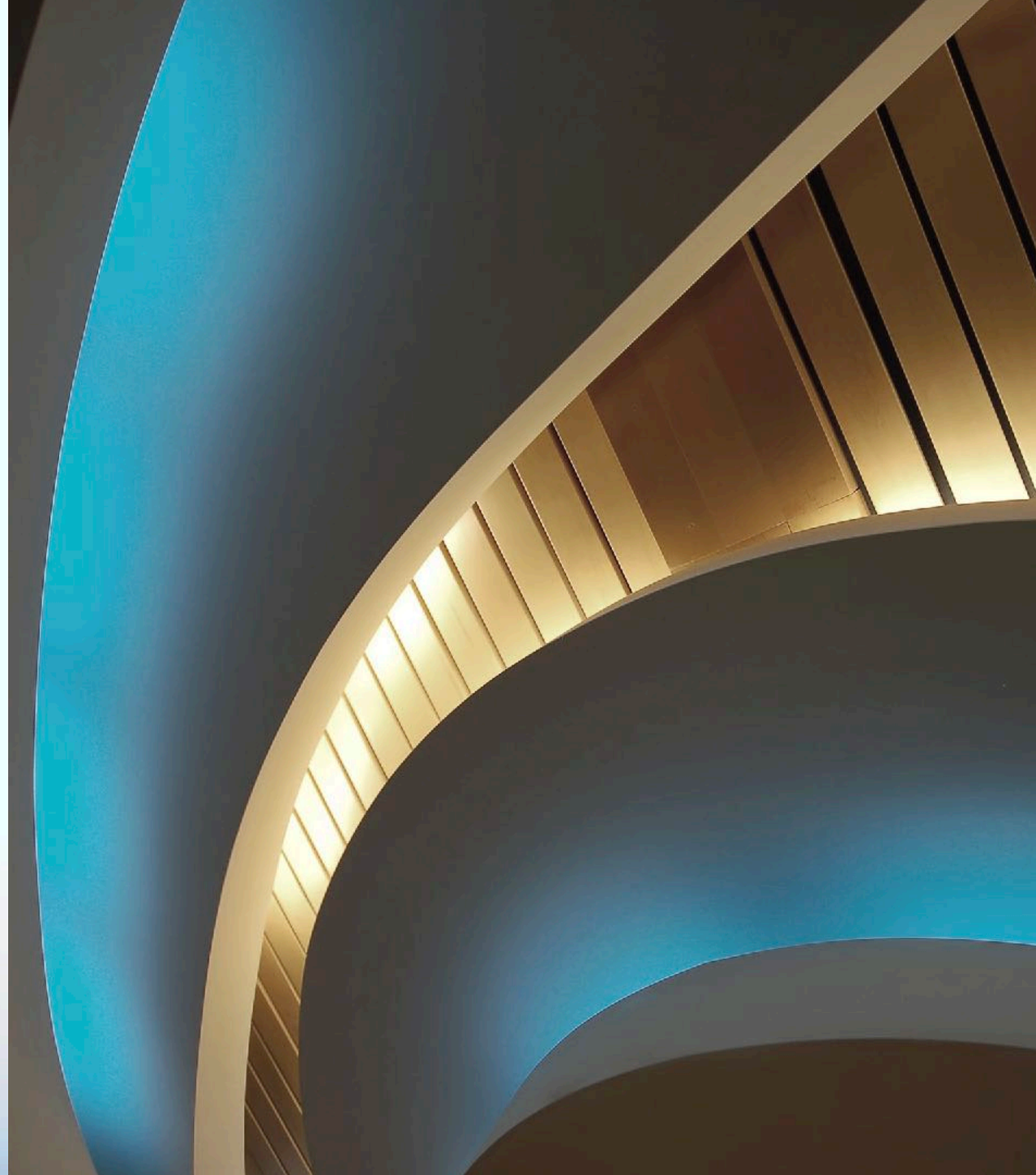
Agenda

- **LLMs & ChatGPT**
- Voice
- Audio
- Image
- Video
- Multi-Modal
- Autonomous AI Agents



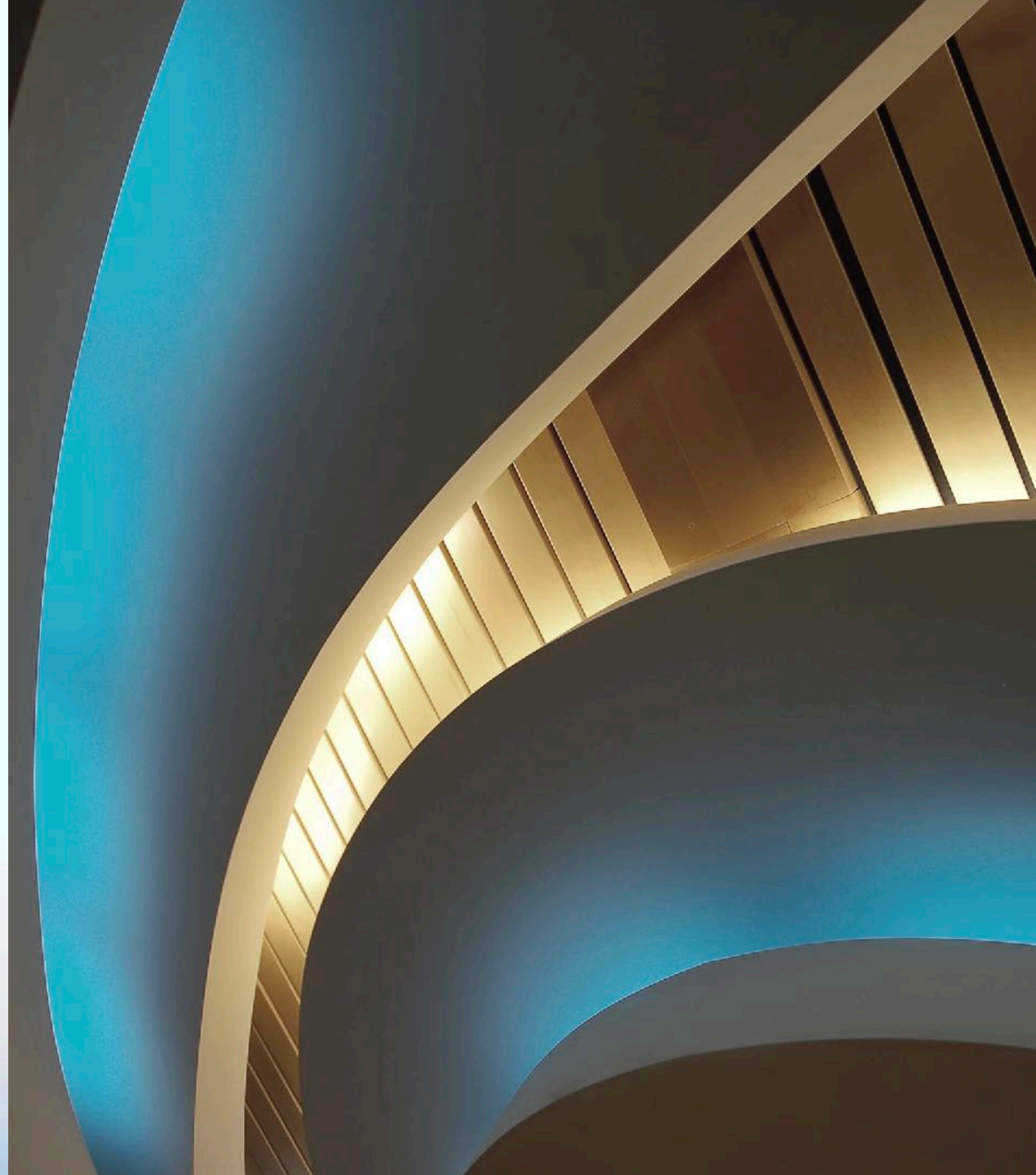
Agenda

- **LLMs & ChatGPT**
  - ChatGPT, Plugins, Functions
  - Fine Tuning, Embeddings, GPTs
  - Ecosystem



Agenda

- **LLMs & ChatGPT**
  - **ChatGPT, Plugins, Functions**
  - Fine Tuning, Embeddings, GPTs
  - Ecosystem



GPT-3.5 | GPT-4 | Alpha

ChatGPT PLUS

Help me study
vocabulary for a college entrance exam

Come up with concepts
for a retro-style arcade game

Give me ideas
for what to do with my kids' art

Suggest fun activities
to do indoors with my high-energy dog

Send a message

?

New chat

Hi in Thai

Model
Plugins ALPHA

Plugins

Updates & FAQ

Log out

Plugin store

Featured plugins | Unverified plugins

OpenTable
Uninstall

Allows you to search for restaurants available for booking dining experiences

Speak
Uninstall

Learn how to say anything in another language with Speak, your AI-powered language tutor.

Klarna Shopping
Uninstall

Search and compare prices from thousands of online shops

KAYAK
Uninstall

Search flights, stays & rental cars or get recommendations where you can go on your budget.

FiscalNote
Uninstall

FiscalNote enables access to select market-leading, real-time data sets for legal, political, and regulatory...

Expedia
Uninstall

Bring your trip plans to life – get there, stay there, find things to see and do.

Milo Family AI
Uninstall

Curating the wisdom of village to give parents ideas that turn any 20 minutes from meh to magic.

Shop
Uninstall

Search for millions of products from the world's greatest brands.

< Prev 1 2 Next >

Install an unverified plugin | Develop your own plugin | About plugins

GPT-3.5 GPT-4 Alpha

Plugin store

Popular | New | All | Installed | Search plugins

A/B Analytics
Install
A/B test and analyze data from Google Analytics, Facebook Ads and more. Powered by Avian.io.
Developer info

A/B JUDGE
Install
Judge whether the A/B test results are superior or not.
Developer info

Aalii FileChat
Install
File management, in-depth analysis, and quick information retrieval.
Developer info

Aardvark AI
Install
Search for Products on Google Shopping in Real-Time, No Ads, Only Trusted Stores.
Developer info

Aardwolf ads
Install
Unlock stunning image ads with just a link. Our Bestever AI scripts, polishes your visuals, and generates magic!
Developer info

Aaron Browser
Install
I'll scrape data from multiple website URLs. Built for Internet crawling, content aggregation, and monitoring.
Developer info

Aaron Build Resume
Install
Create impressive professional resume/CV from scratch or update an existing one. Export as PDF and .docx.
Developer info

Aaron Chat PDF
Install
I'll extract, analyze & chat with multiple PDFs or Google Drive documents. Ask questions, get answers & page...
Developer info

Install an unverified plugin | Develop your own plugin | About plugins

Send a message

Function calling

In an API call, you can describe functions to `gpt-3.5-turbo-0613` and `gpt-4-0613`, and have the model intelligently choose to output a JSON object containing arguments to call those functions. The Chat completions API does not call the function; instead, the model generates JSON that you can use to call the function in your code.

The latest models (`gpt-3.5-turbo-0613` and `gpt-4-0613`) have been fine-tuned to both detect when a function should to be called (depending on the input) and to respond with JSON that adheres to the function signature. With this capability also comes potential risks. We strongly recommend building in user confirmation flows before taking actions that impact the world on behalf of users (sending an email, posting something online, making a purchase, etc).

> Under the hood, functions are injected into the system message in a syntax the model has been trained on. This means functions count against the model's context limit and are billed as input tokens. If running into context limits, we suggest limiting the number of functions or the length of documentation you provide for function parameters.

Function calling allows you to more reliably get structured data back from the model. For example, you can:

Function calling

Who are my top customers?

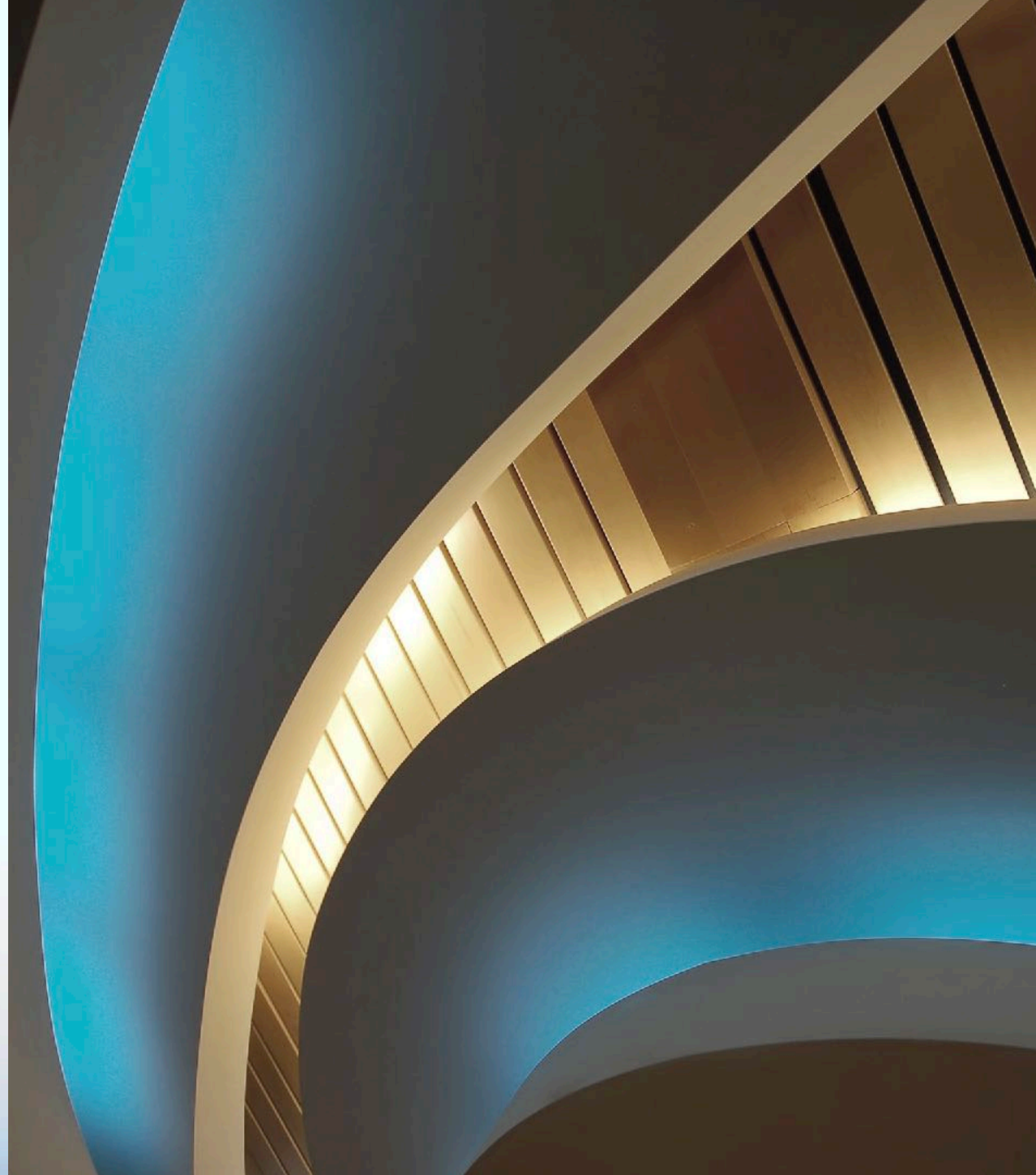
to call the

The latest models (`gpt-3.5-turbo-0613` and `gpt-4-0613`) have been fine-tuned to both detect when a function should to be called (depending on the input) and to respond with JSON that adheres to the

```
get_customers(
min_revenue: int,
created_before: string,
limit: int)
```

Agenda

- **LLMs & ChatGPT**
  - ChatGPT, Plugins, Functions
  - **Fine Tuning, Embeddings, GPTs**
  - Ecosystem



Fine-tuning

Learn how to customize a model for your application.

Introduction

> This guide is intended for users of the new OpenAI fine-tuning API. If you are a legacy fine-tuning user, please refer to our legacy fine-tuning guide.

Fine-tuning lets you get more out of the models available through the API by providing:

1. Higher quality results than prompting
2. Ability to train on more examples than can fit in a prompt
3. Token savings due to shorter prompts
4. Lower latency requests

GPT models have been pre-trained on a vast amount of text. To use the models effectively, we include instructions and sometimes several examples in a prompt. Using demonstrations to show how to perform a task is often called "few-shot learning."

Fine-tuning improves on few-shot learning by training on many more examples than can fit in the prompt, letting you achieve better results on a wide number of tasks. **Once a model has been fine-tuned, you won't need to provide as many examples in the prompt.** This saves costs and enables lower-latency requests.

At a high level, fine-tuning involves the following steps:

1. Prepare and upload training data
2. Train a new fine-tuned model
3. Use your fine-tuned model

GPT-3.5 GPT-4 Alpha

Custom instructions

What would you like ChatGPT to know about you to provide better responses?

i get lost when reading too long and fluffy texts.
im a computer engineer with 10+ years of experience in software and ai.

122/1500

How would you like ChatGPT to respond?

short, to the point, i know you are an LLM model by open ai - no need to say it, clear action items. use less words.

116/1500

Enable for new chats

Cancel Save

York in 3 days

gy
ng free local weekend events

Send a message

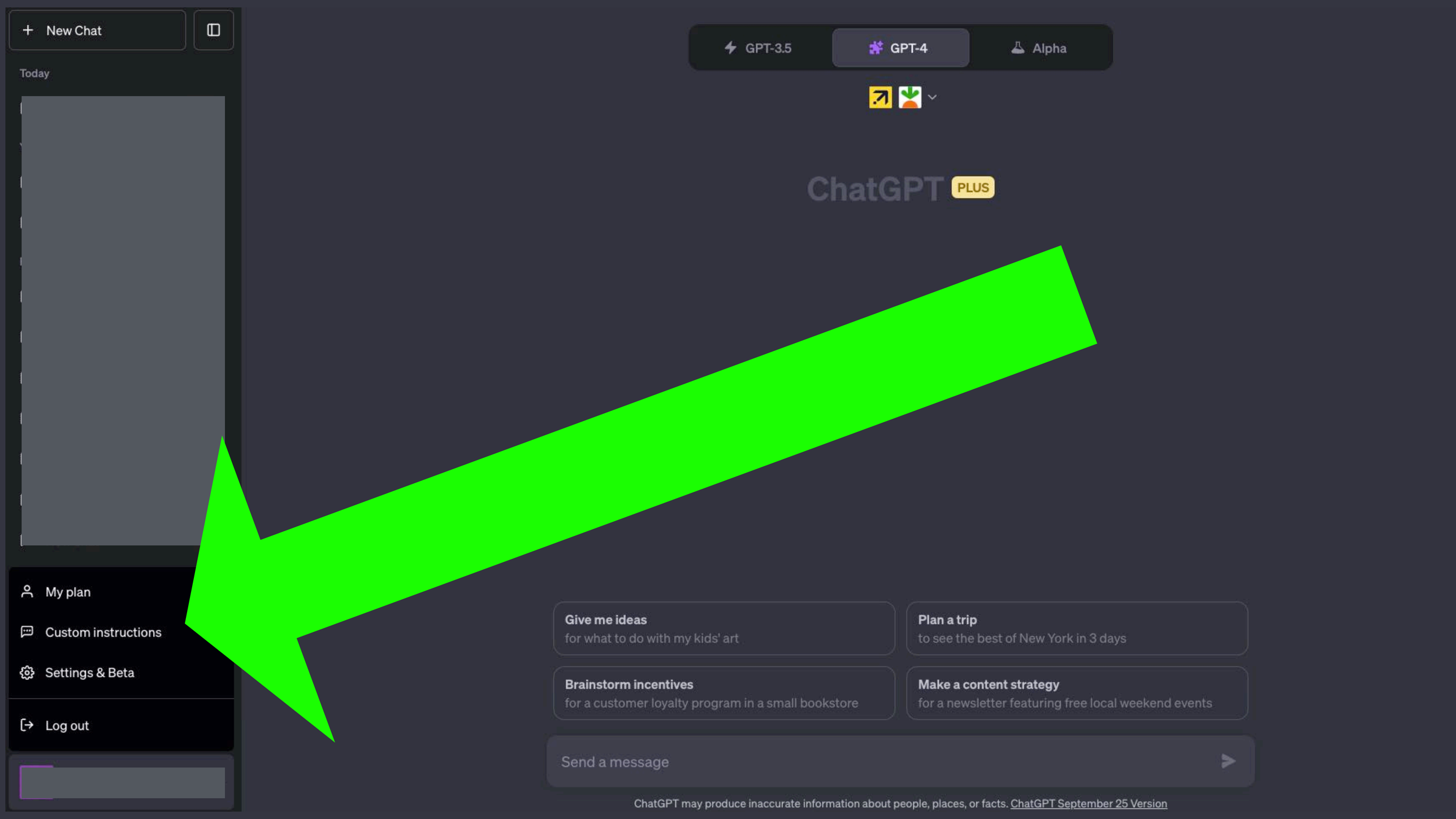
ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT September 25 Version

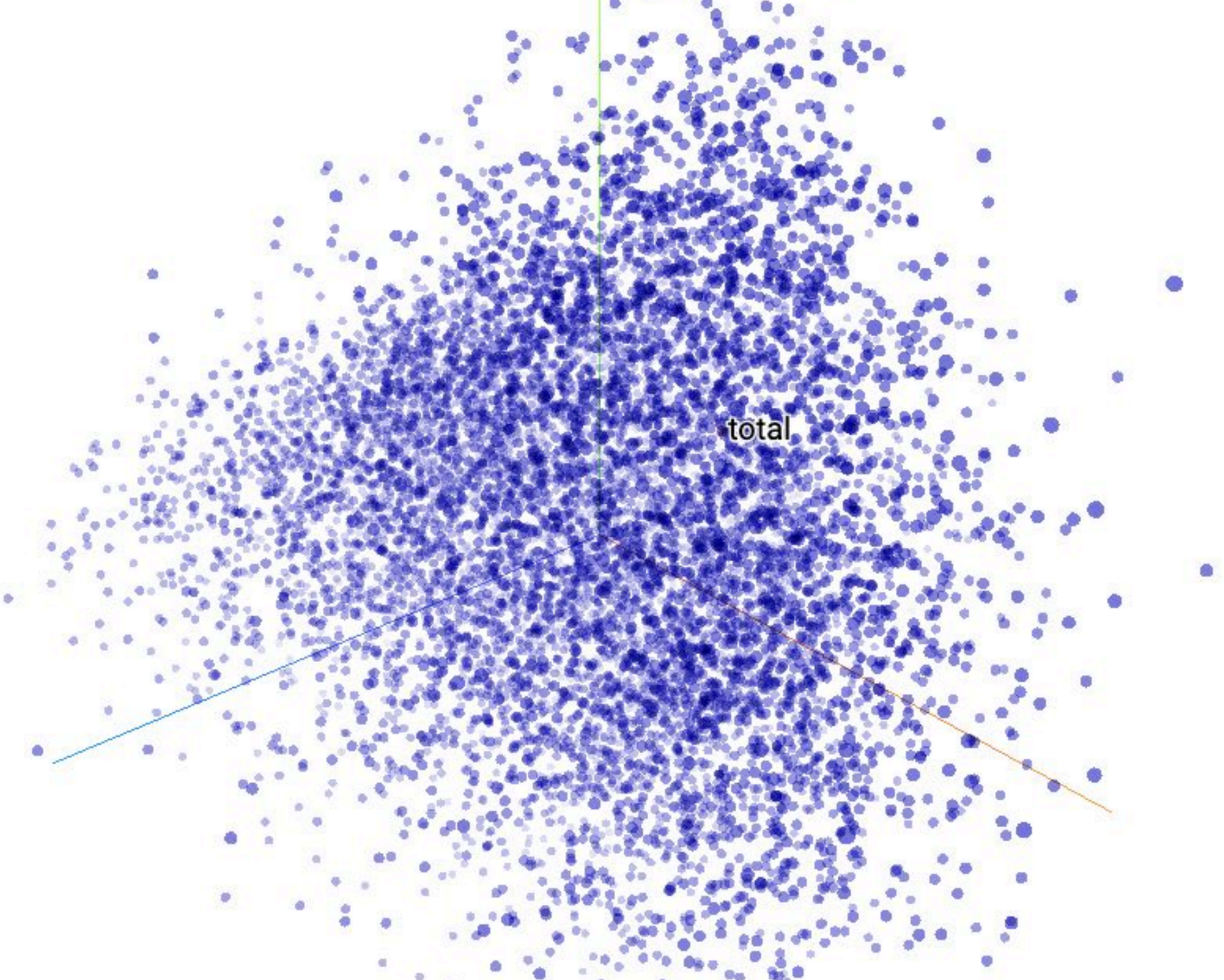
GPT-4

NA

what 3 criteria can i use to evaluate an ai agent

1. Accuracy: Measure how often the AI outputs correct results.
2. Efficiency: Assess computational resources needed for tasks.
3. Adaptability: Evaluate how well the AI adjusts to new data or tasks.


New Chat
Today
My plan
Custom instructions
Settings & Beta
Log out
GPT-3.5
GPT-4
Alpha
ChatGPT
PLUS
Give me ideas
for what to do with my kids' art
Plan a trip
to see the best of New York in 3 days
Brainstorm incentives
for a customer loyalty program in a small bookstore
Make a content strategy
for a newsletter featuring free local weekend events
Send a message
ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT September 25 Version


total

GopherGPT
GPT for the Go programming language
43
Edit

PayGPT-keeps resetting
Pay with Stripe
Edit

CatGPT
Reply with cat language
0
Edit

Recently Used

GopherGPT
GPT for the Go programming language
43
Edit

PayGPT-keeps resetting
Pay with Stripe
Edit

CatGPT
Reply with cat language
0
Edit

Made by OpenAI

DALL·E
Let me turn your imagination into imagery
By ChatGPT

Data Analysis
Drop in any files and I can help analyze and visualize your data
By ChatGPT

ChatGPT Classic
The latest version of GPT-4 with no additional
By ChatGPT

Poe

Assistant | Web-Search | GPT-4 | More

Start a new chat

Official bots

See all

Assistant
General-purpose assistant bot with...

Web-Search
General-purpose assistant bot capab...

GPT-4
OpenAI's most powerful model....

StableDiffusion...
Generates high quality images base...

Claud
Anthrop
model,

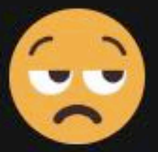
Popular bots

See all

VanGoghPaint
This bot will draw anything you want i...

Midjourney
Midjourney photo prompter

DSLR-SDXL
Generates pictures as if taken by Cano...

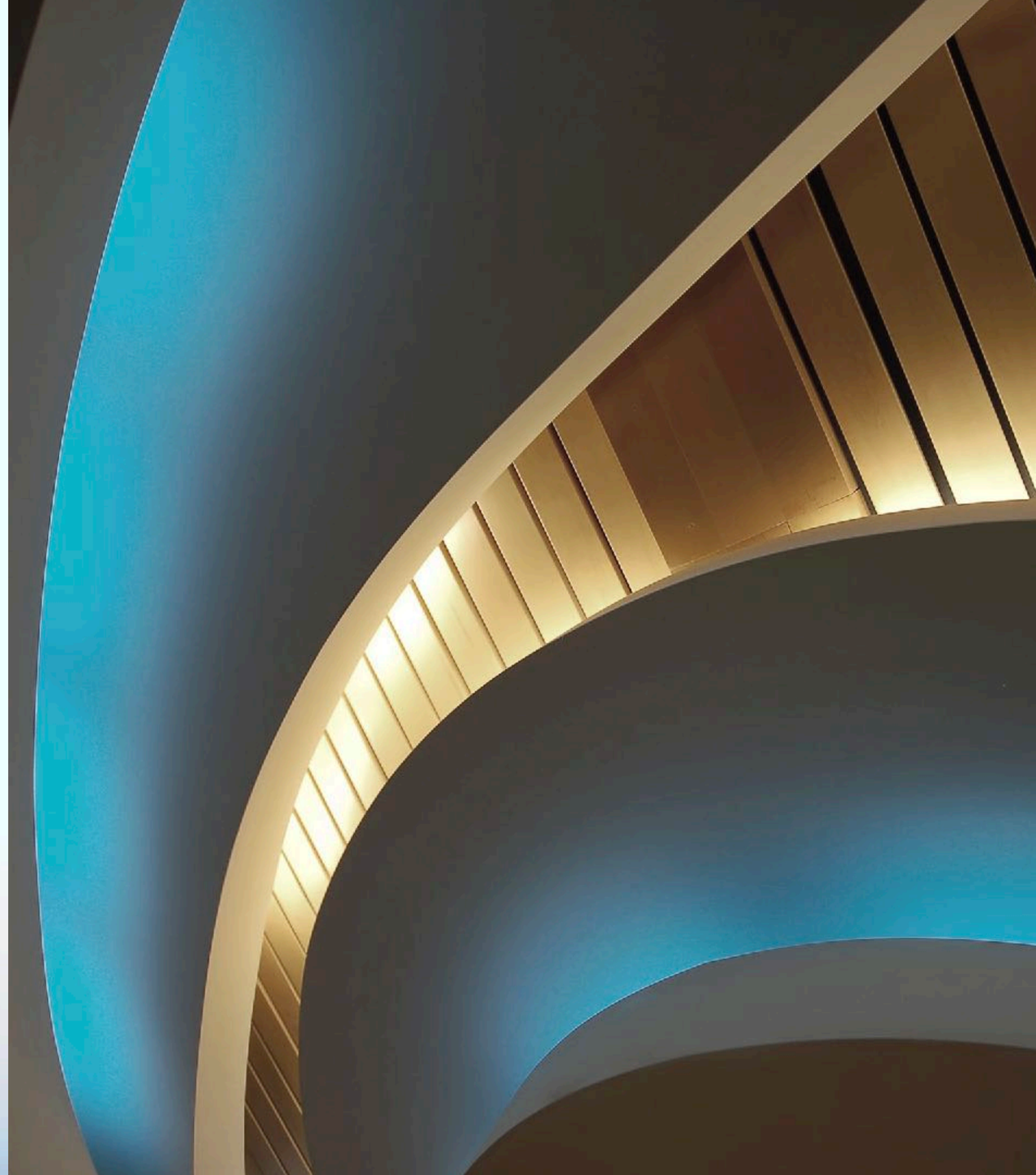


WhateverGPT
After abandoned by creator, this AI...

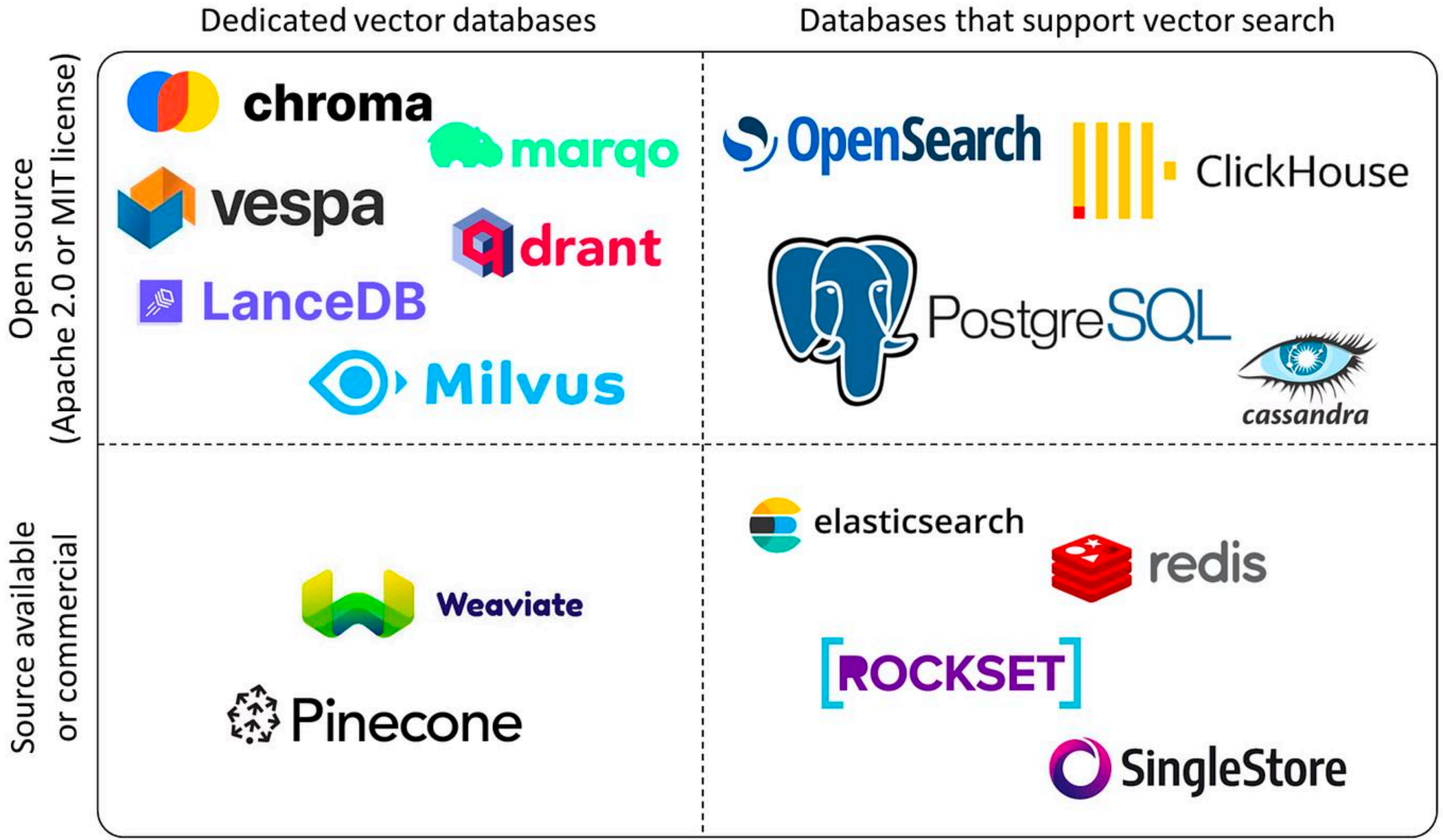
Leona
Leo
Genera

Agenda

- **LLMs & ChatGPT**
  - ChatGPT, Plugins, Functions
  - Fine Tuning, Embeddings, GPTs
  - **Ecosystem**



Vector DB



https://www.datacamp.com/blog/the-top-5-vector-databases

x.ai

OpenAI

ANTHROP\C

Meta AI

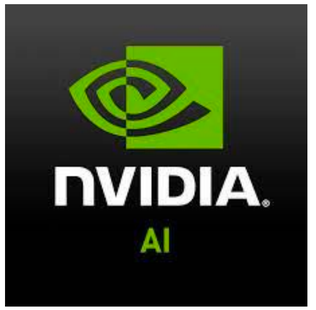
Baidu 百度

Hugging Face

Microsoft Research

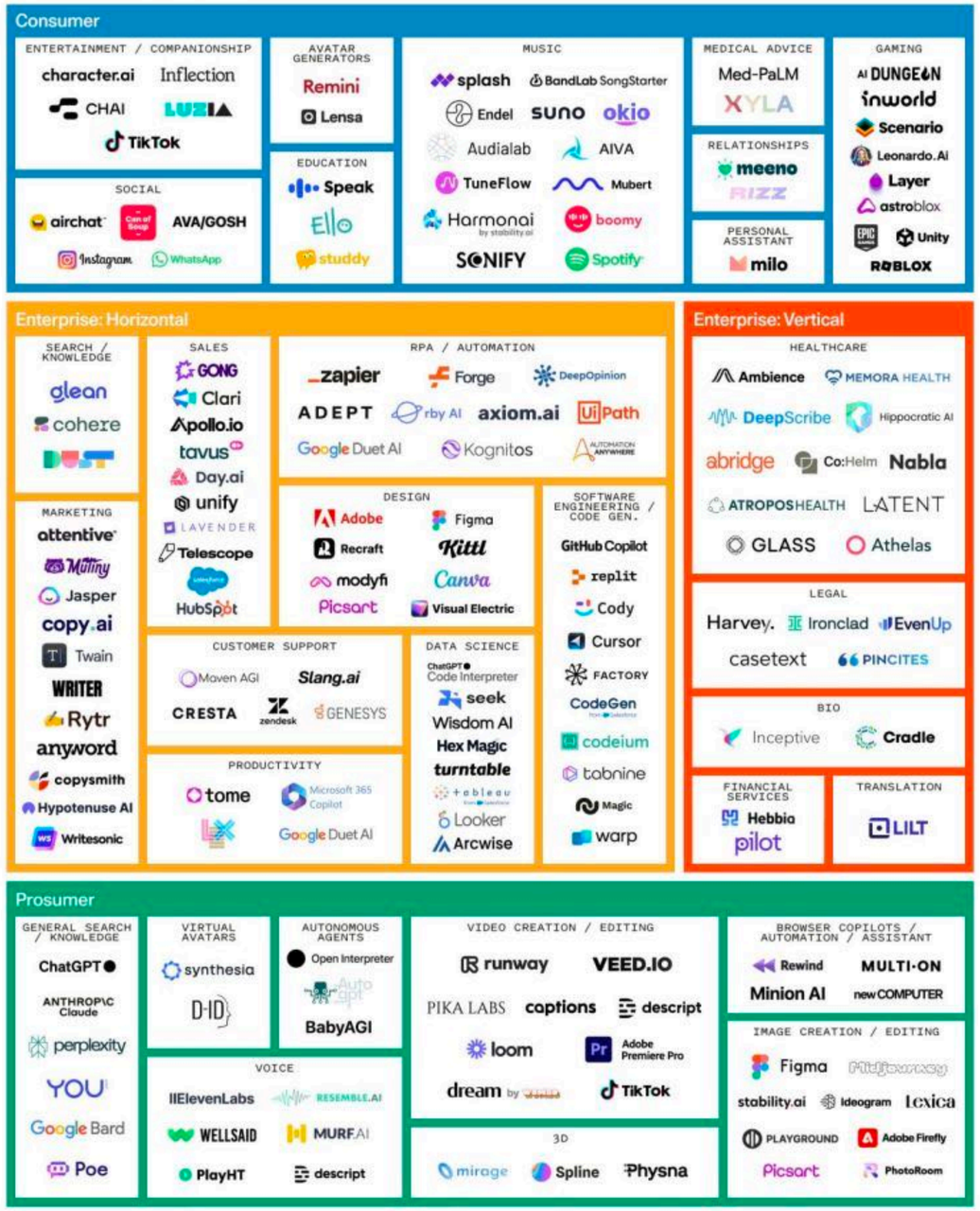
Google DeepMind

LangChain

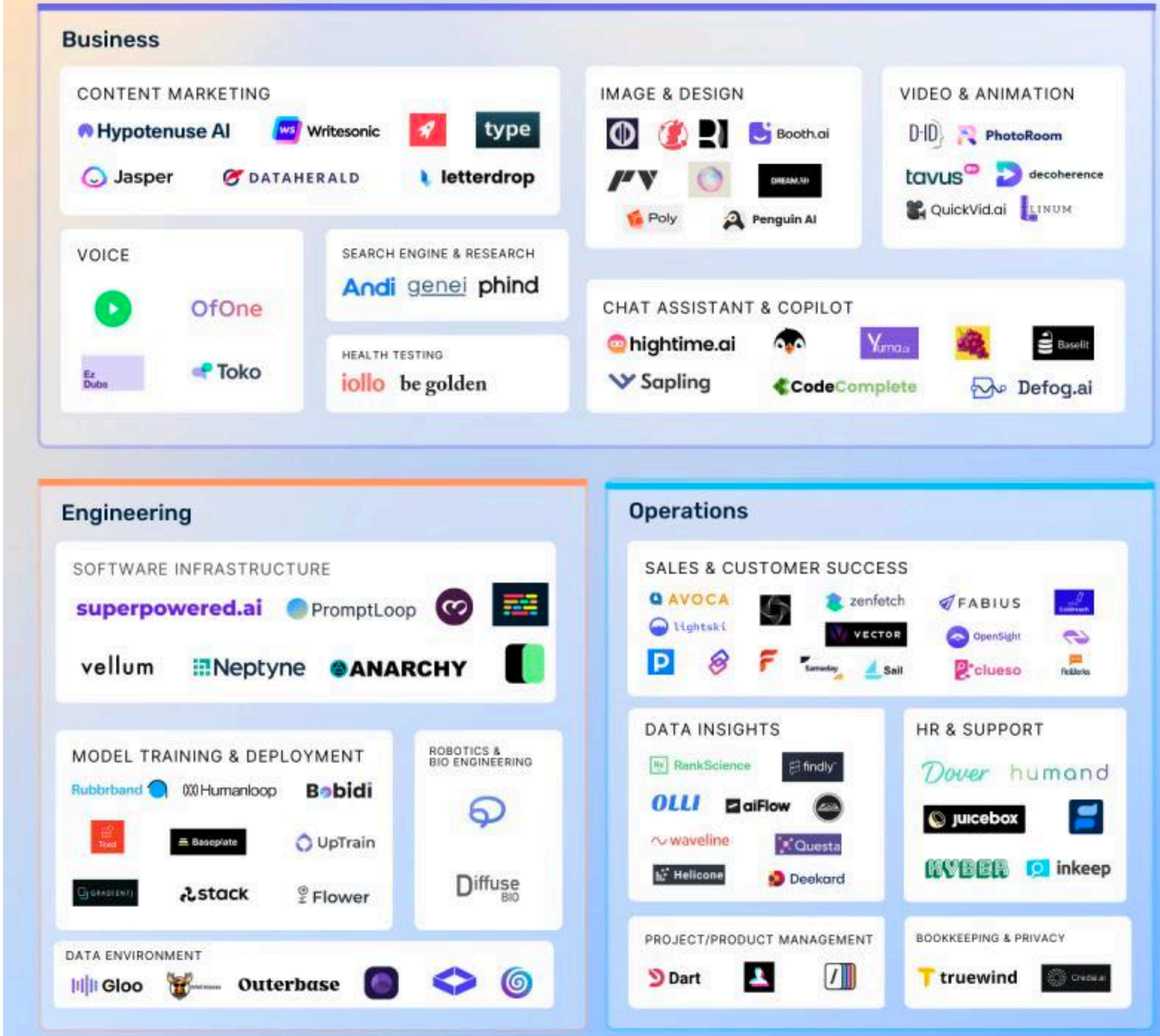


The Generative AI Market Map v3

A work in progress



Y COMBINATOR GENERATIVE AI LANDSCAPE



Three former SpaceX cybersecurity engineers have launched Wraithwatch, an AI-powered security firm that has received $8 million in seed funding.

Wraithwatch was founded by Nik Seetharaman (CEO), Grace Clemente (president), and Carlos Más (CTO). They all previously held important cybersecurity roles at SpaceX, as well as other major companies such as Anduril, Palantir Technologies, Google and Morgan Stanley.

The startup emerged from stealth mode on Thursday with financial backing from Founders Fund, which led the seed funding round, XYZ Capital and Human Capital.

Little information is currently available on Wraithwatch's offering. The company has described worst-case scenarios of AI-powered cyberattacks and claims its goal is "building the defense" against such attacks by using large language mod generate all possible versions of an attack in an effort to predict and prev threats.

Cranium, a company that specializes in securing artificial intelligence (AI) applications and deployments, on Thursday announced raising $25 million in Series A funding.

The funding, which brings the total investment in the company to $32 million, was led by Telstra Ventures, with participation from KPMG and SYN Ventures. The money will be used for R&D and business expansion.

Cranium emerged from stealth earlier this year after spinning out of KPMG. The company helps organizations secure their AI and machine learning (ML) systems, ensuring they are compliant and trustworthy.

Cranium enables organizations to gain visibility, security, and compliance across their AI systems, helping them map, monitor, and manage AI/ML environments.

The Cranium Enterprise software platform can also be used to collect and share information on the trustworthiness and compliance of AI models with third parties, clients, and regulators.

OWASP

Open Web Application Security Project

OWASP for LLMs

- Prompt Injection

OWASP for LLMs

- Prompt Injection
- Insecure Output Handling

OWASP for LLMs

- Prompt Injection
- Insecure Output Handling
- Training Data Poisoning

OWASP for LLMs

- Prompt Injection
- Insecure Output Handling
- Training Data Poisoning
- Model Denial of Service

OWASP for LLMs

- Prompt Injection
- Insecure Output Handling
- Training Data Poisoning
- Model Denial of Service
- Supply Chain Vulnerabilities

OWASP for LLMs

- Prompt Injection
- Insecure Output Handling
- Training Data Poisoning
- Model Denial of Service
- Supply Chain Vulnerabilities
- Sensitive Info Disclosure

OWASP for LLMs

- Prompt Injection
- Insecure Output Handling
- Training Data Poisoning
- Model Denial of Service
- Supply Chain Vulnerabilities
- Sensitive Info Disclosure
- Insecure Plugin Design

OWASP for LLMs

- Prompt Injection
- Insecure Output Handling
- Training Data Poisoning
- Model Denial of Service
- Supply Chain Vulnerabilities
- Sensitive Info Disclosure
- Insecure Plugin Design
- Excessive Agency

OWASP for LLMs

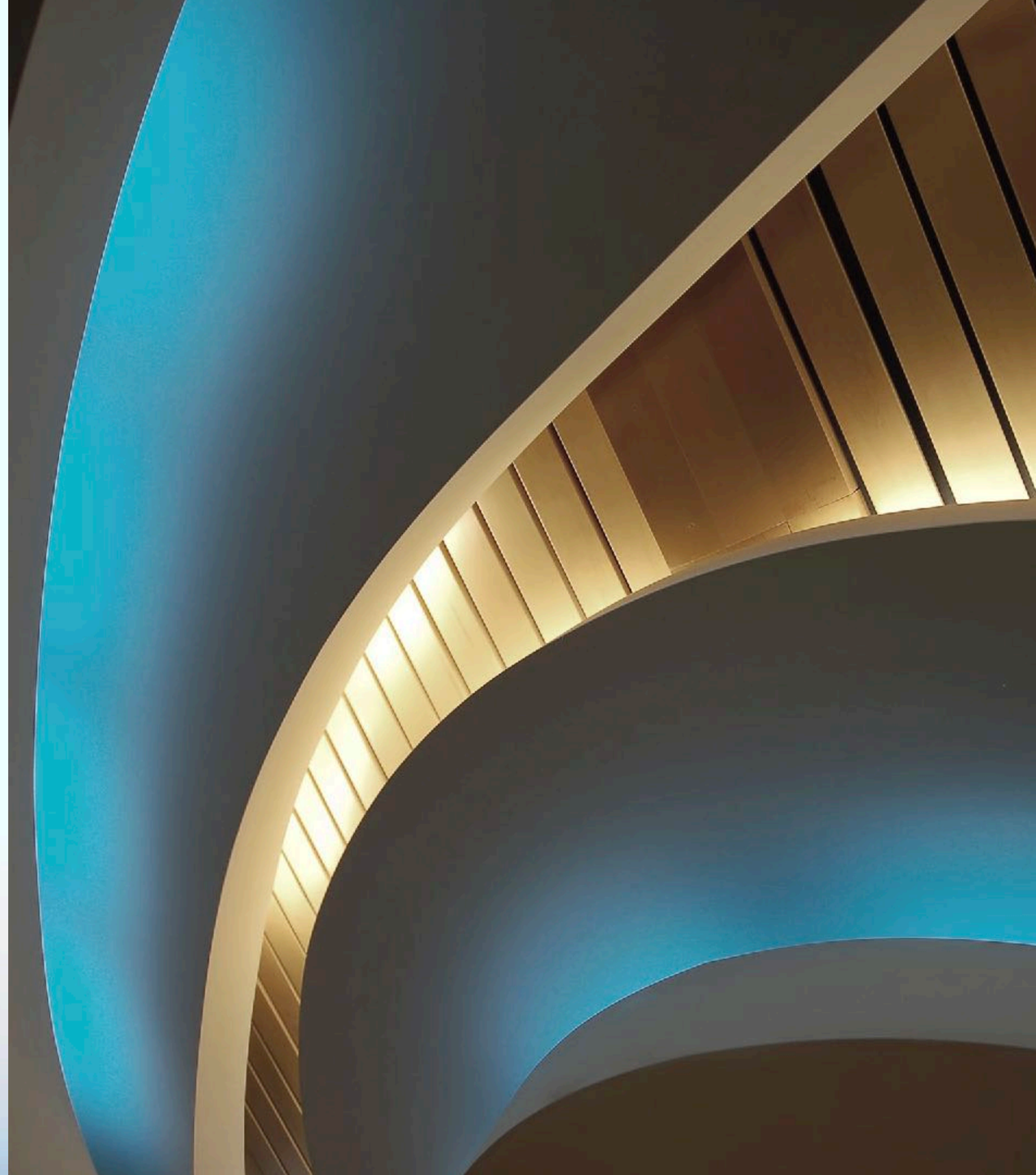
- Prompt Injection
- Insecure Output Handling
- Training Data Poisoning
- Model Denial of Service
- Supply Chain Vulnerabilities
- Sensitive Info Disclosure
- Insecure Plugin Design
- Excessive Agency
- Over-reliance

OWASP for LLMs

- Prompt Injection
- Insecure Output Handling
- Training Data Poisoning
- Model Denial of Service
- Supply Chain Vulnerabilities
- Sensitive Info Disclosure
- Insecure Plugin Design
- Excessive Agency
- Over-reliance
- Model Theft

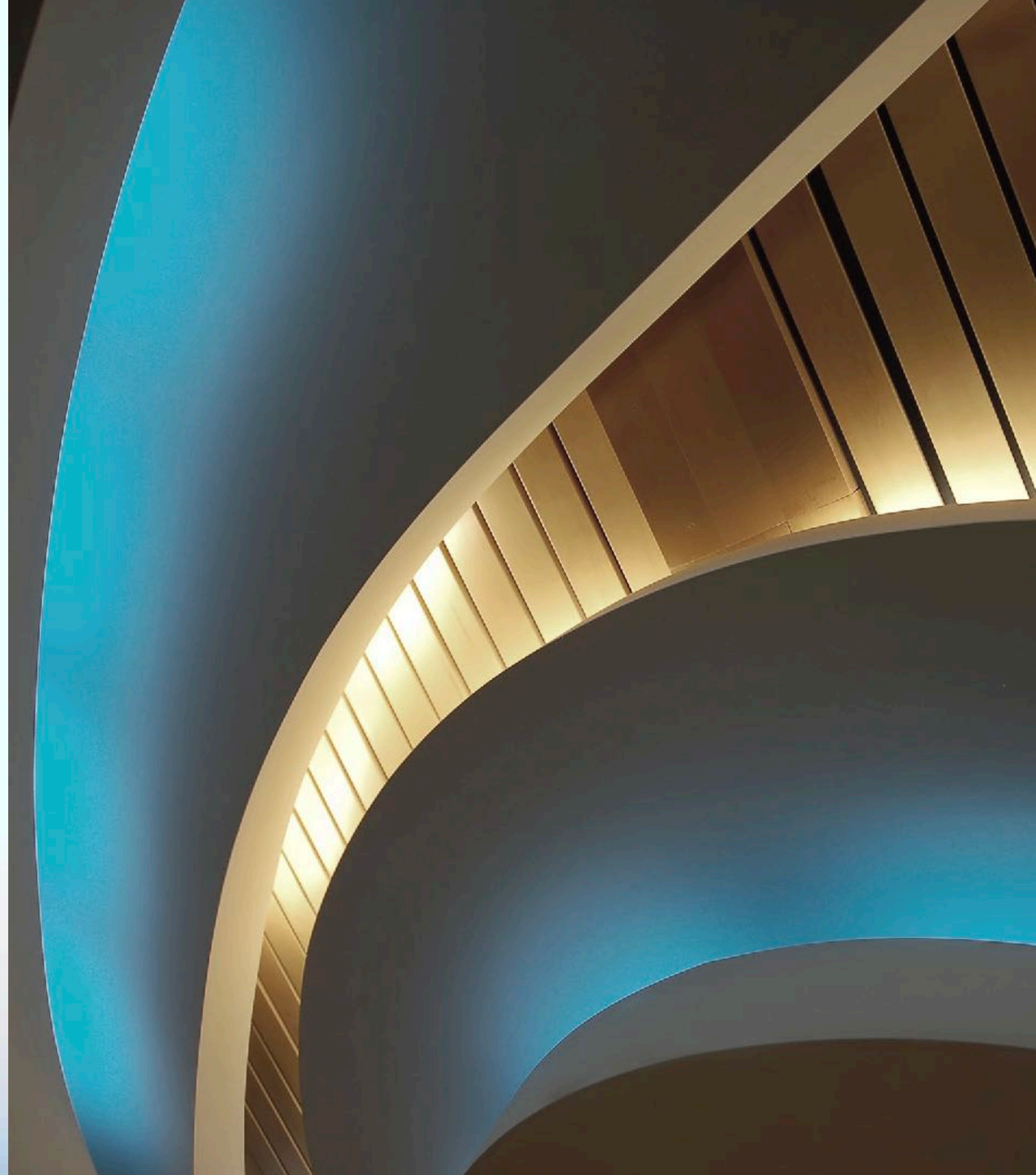
Agenda

- LLMs & ChatGPT
- **Voice**
- Audio
- Image
- Video
- Multi-Modal
- Autonomous AI Agents



Agenda

- **Voice**
  - Voice generation: Eleven Labs
  - Voice transcription: Whisper



IIElevenLabs

Speech to text

Learn how to turn audio into text

Introduction

The speech to text API provides two endpoints, `transcriptions` and `translations`, based on our state-of-the-art open source large-v2 Whisper model. They can be used to:

- Transcribe audio into whatever language the audio is in.
- Translate and transcribe the audio into english.

File uploads are currently limited to 25 MB and the following input file types are supported: `mp3`, `mp4`, `mpeg`, `mpga`, `m4a`, `wav`, and `webm`.

Quickstart

Transcriptions

The transcriptions API takes as input the audio file you want to transcribe and the desired output file format for the transcription of the audio. We currently support multiple input and output file formats.

Transcribe audio — python — Copy

```python
# Note: you need to be using OpenAI Python v0.27.0 for the code below to work
import openai
audio_file= open("/path/to/file/audio.mp3", "rb")
transcript = openai.Audio.transcribe("whisper-1", audio_file)
```

Speech to text

Learn how to turn audio into text

Transcribe 57 languages

+ Translate to English

Transcribe audio

python

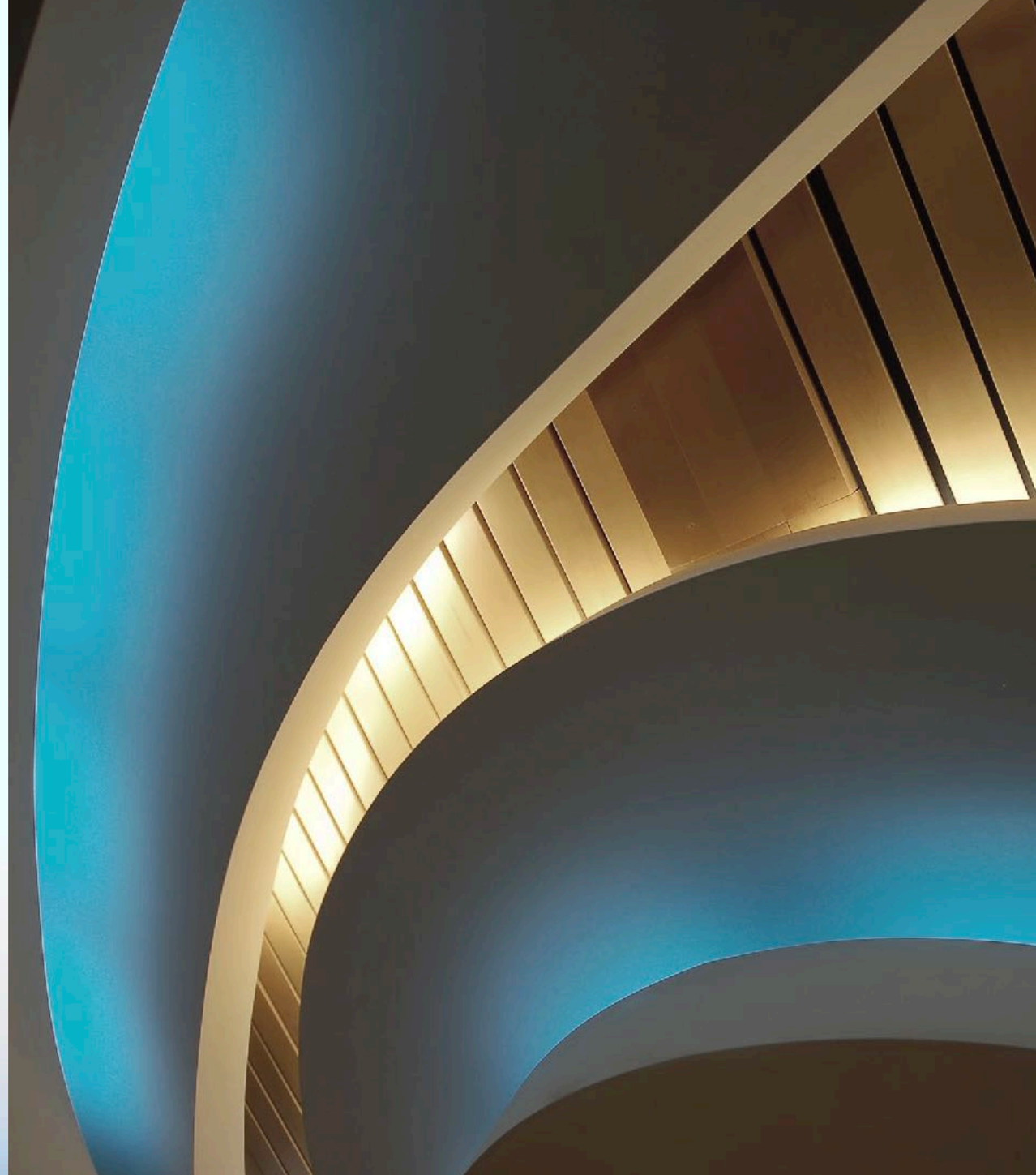
Copy

```python
# Note: you need to be using OpenAI Python v0.27.0 for the code below to work
import openai
audio_file= open("/path/to/file/audio.mp3", "rb")
transcript = openai.Audio.transcribe("whisper-1", audio_file)
```

OpenAI

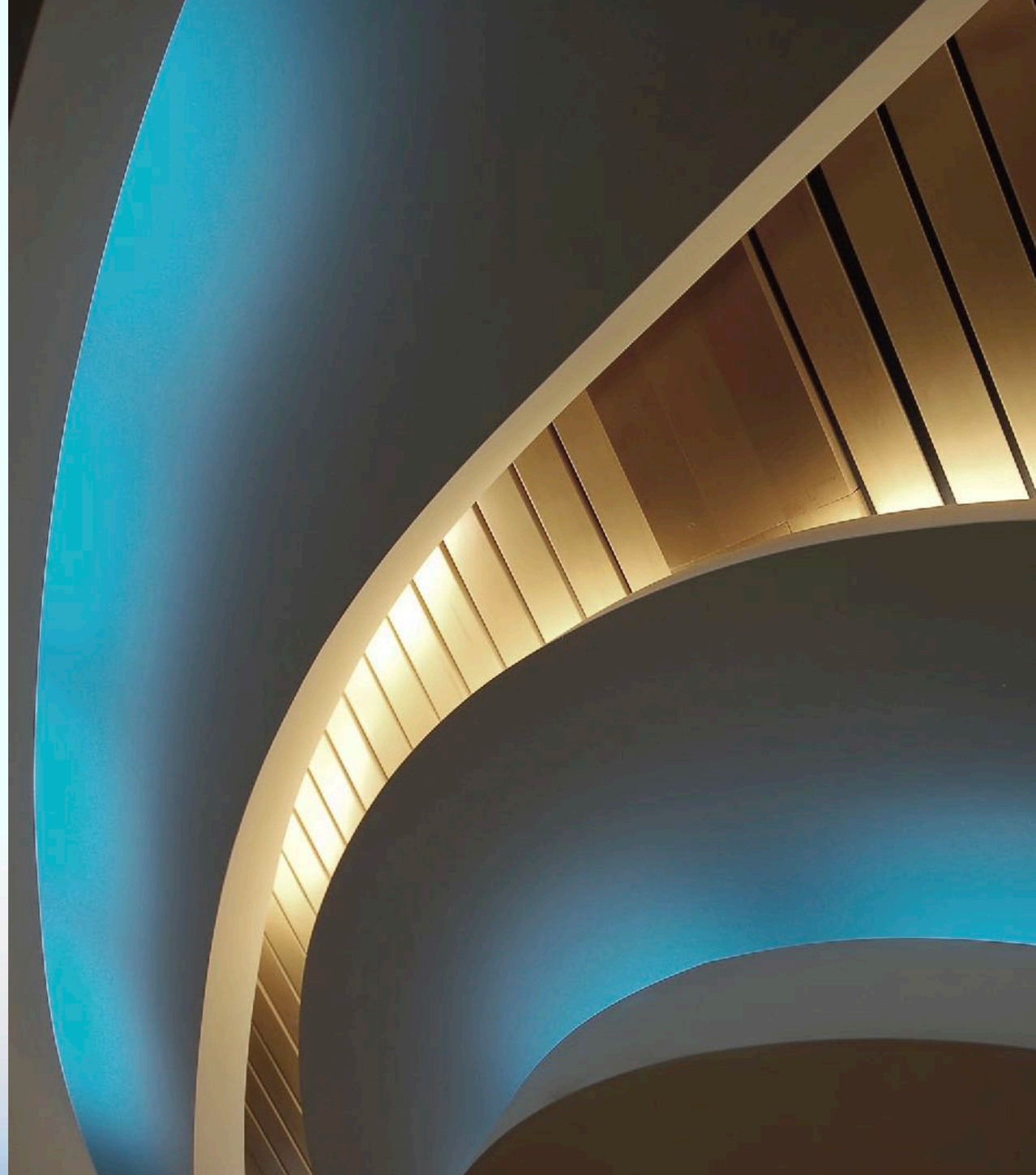
Agenda

- LLMs & ChatGPT
- Voice
- **Audio**
- Image
- Video
- Multi-Modal
- Autonomous AI Agents



Agenda

- **Audio**
  - Meta’s MusicGen



MusicGen

This is the demo for MusicGen, a simple and controllable model for music generation presented at: "Simple and Controllable Music Generation".

Duplicate Space for longer sequences, more control and no queue.

Describe your music

Condition on a melody (optional) File or Mic

file | mic

File

Drop Audio Here
- or -
Click to Upload

Generated Music

Generate

Examples

Describe your music	File
An 80s driving pop song with heavy drums and synth pads in the background	bach.mp3
A cheerful country song with acoustic guitars	bolero_ravel.mp3
90s rock song with electric guitar and heavy drums	
a light and cheerly EDM track, with syncopated drums, aery pads, and strong emotions bpm: 130	bach.mp3
lofi slow bpm electro chill with organic samples	

More details

The model will generate 12 seconds of audio based on the description you provided. You can optionaly provide a reference audio from which a broad melody will be extracted. The model will then try to follow both the description and melody provided. All samples are generated with the `melody` model.

You can also use your own GPU or a Google Colab by following the instructions on our repo.

MusicGen

This is the demo for MusicGen, a simple and controllable model for music generation presented at: "Simple and Controllable Music Generation".

Duplicate Space for longer sequences, more control and no queue.

Describe your music

when the 80s meet electro and metal

Condition on a melody (optional) File or Mic

file | mic

File

Drop Audio Here

- or -

Click to Upload

queue: 1/1 | 35.3/63.4s

Generate

Examples

Describe your music	File
An 80s driving pop song with heavy drums and synth pads in the background	bach.mp3
A cheerful country song with acoustic guitars	bolero_ravel.mp3
90s rock song with electric guitar and heavy drums	
a light and cheerly EDM track, with syncopated drums, aery pads, and strong emotions bpm: 130	bach.mp3
lofi slow bpm electro chill with organic samples	

More details

The model will generate 12 seconds of audio based on the description you provided. You can optionaly provide a reference audio from which a broad melody will be extracted. The model will then try to follow both the description and melody provided. All samples are generated with the `melody` model.

You can also use your own GPU or a Google Colab by following the instructions on our repo.

Verified Artist

Electric Callboy

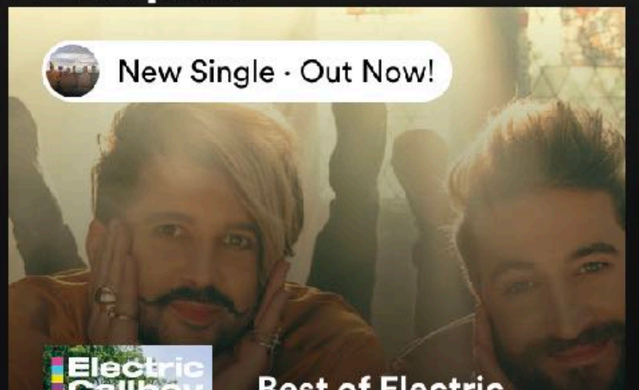
2,063,017 monthly listeners

Follow • • •

Popular

1		We Got the Moves E	58,992,323	3:26
2		Everytime We Touch - ...	11,523,755	3:17
3		Tekkno Train	19,818,459	2:57

Artist pick



MusicGen

This is the demo for MusicGen, a simple and controllable model for music generation presented at: "Simple and Controllable Music Generation".

Duplicate Space for longer sequences, more control and no queue.

Describe your music

when the 80s meet electro and metal

Condition on a melody (optional) File or Mic

file | mic

File

Drop Audio Here
- or -
Click to Upload

Generate

Examples

Describe your music	File
An 80s driving pop song with heavy drums and synth pads in the background	bach.mp3
A cheerful country song with acoustic guitars	bolero_ravel.mp3
90s rock song with electric guitar and heavy drums	
a light and cheerly EDM track, with syncopated drums, aery pads, and strong emotions bpm: 130	bach.mp3
lofi slow bpm electro chill with organic samples	

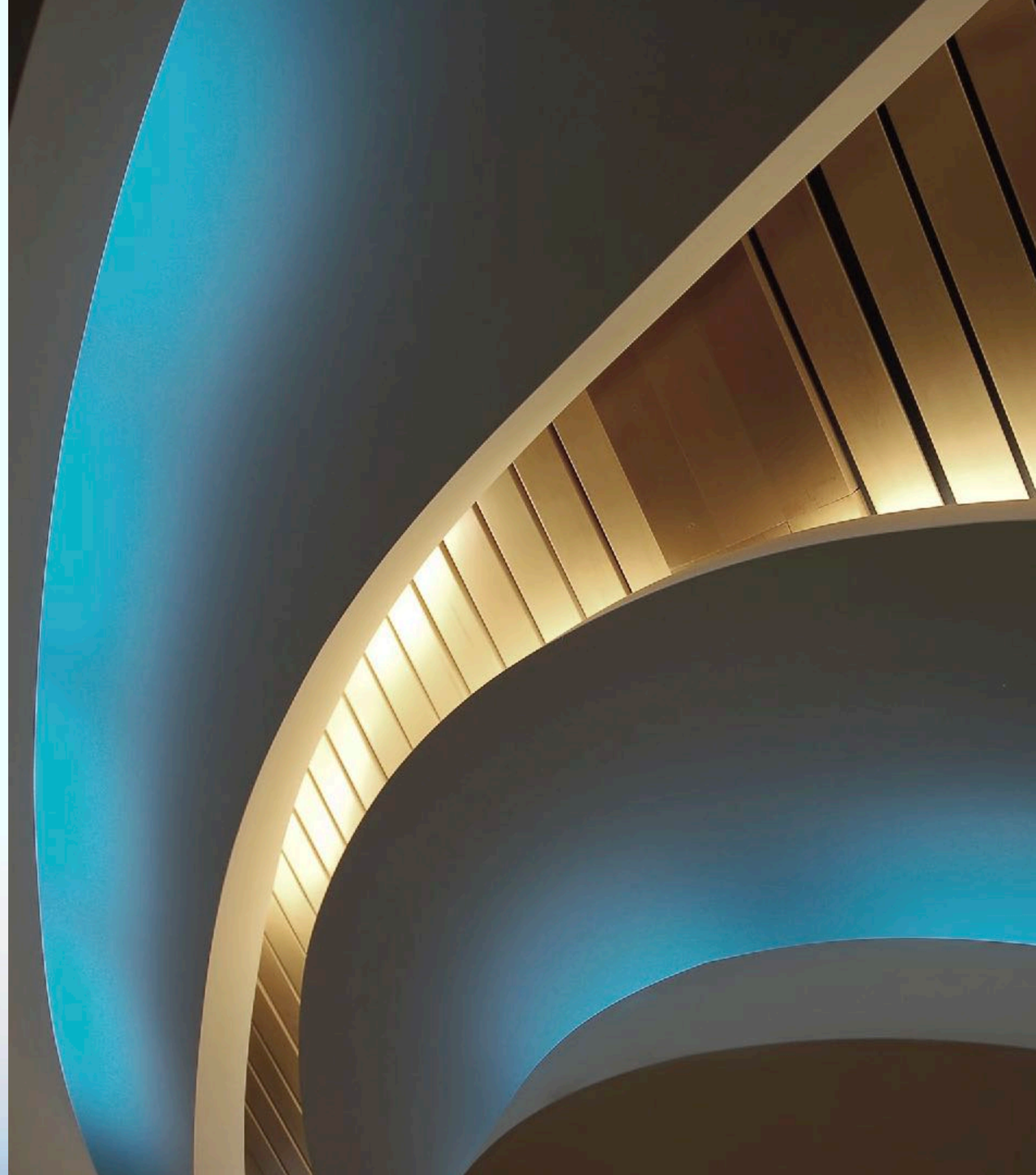
More details

The model will generate 12 seconds of audio based on the description you provided. You can optionaly provide a reference audio from which a broad melody will be extracted. The model will then try to follow both the description and melody provided. All samples are generated with the `melody` model.

You can also use your own GPU or a Google Colab by following the instructions on our repo.

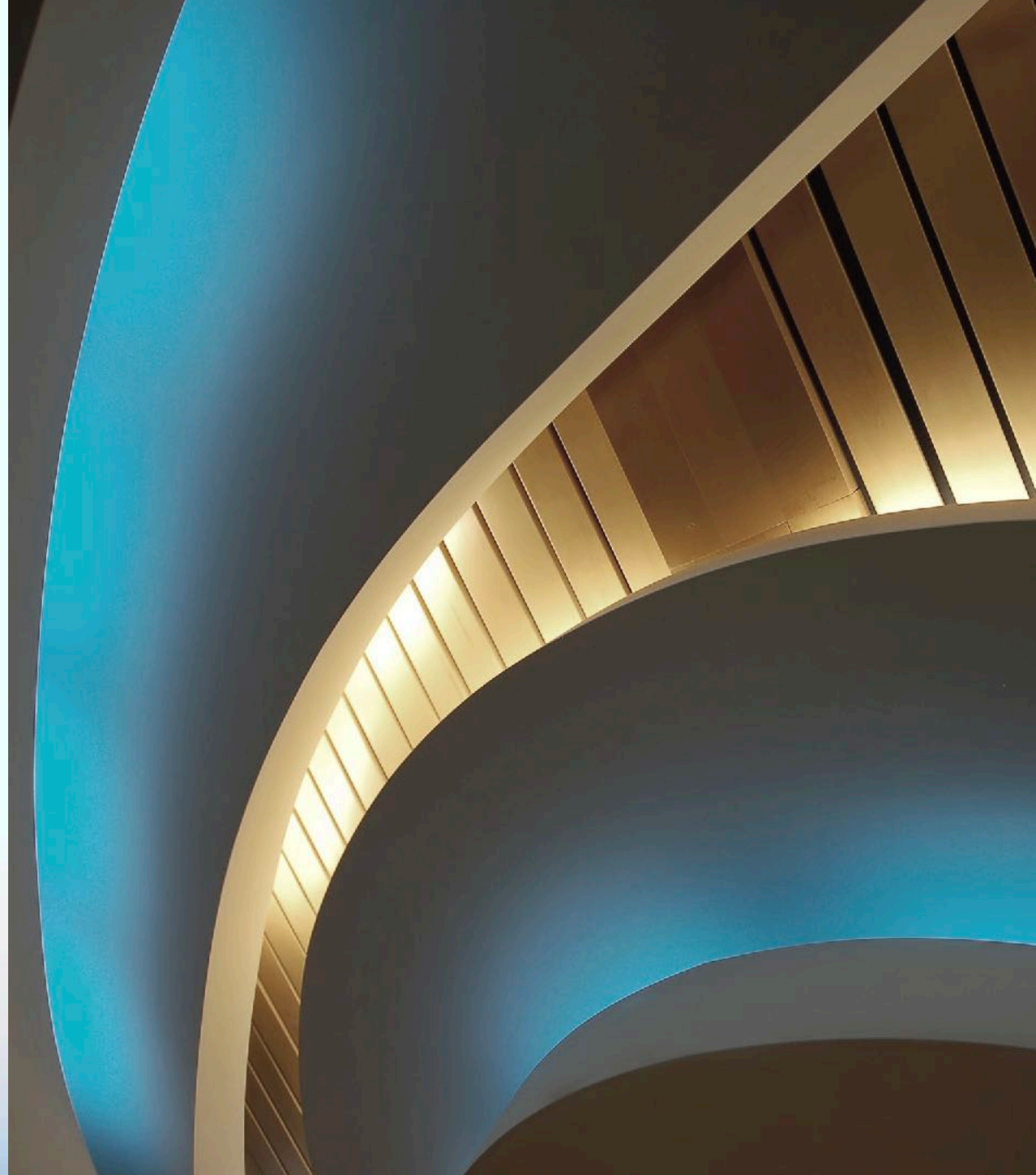
Agenda

- LLMs & ChatGPT
- Voice
- Audio
- **Image**
- Video
- Multi-Modal
- Autonomous AI Agents



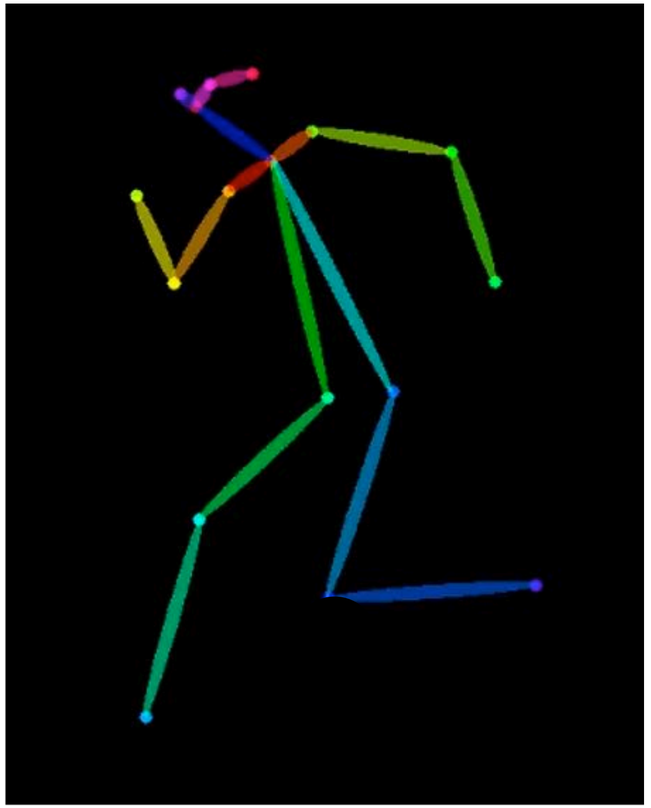
Agenda

- **Image**
  - DALLE-3
  - Stable Diffusion XL
  - Midjourney




stability.ai

input image



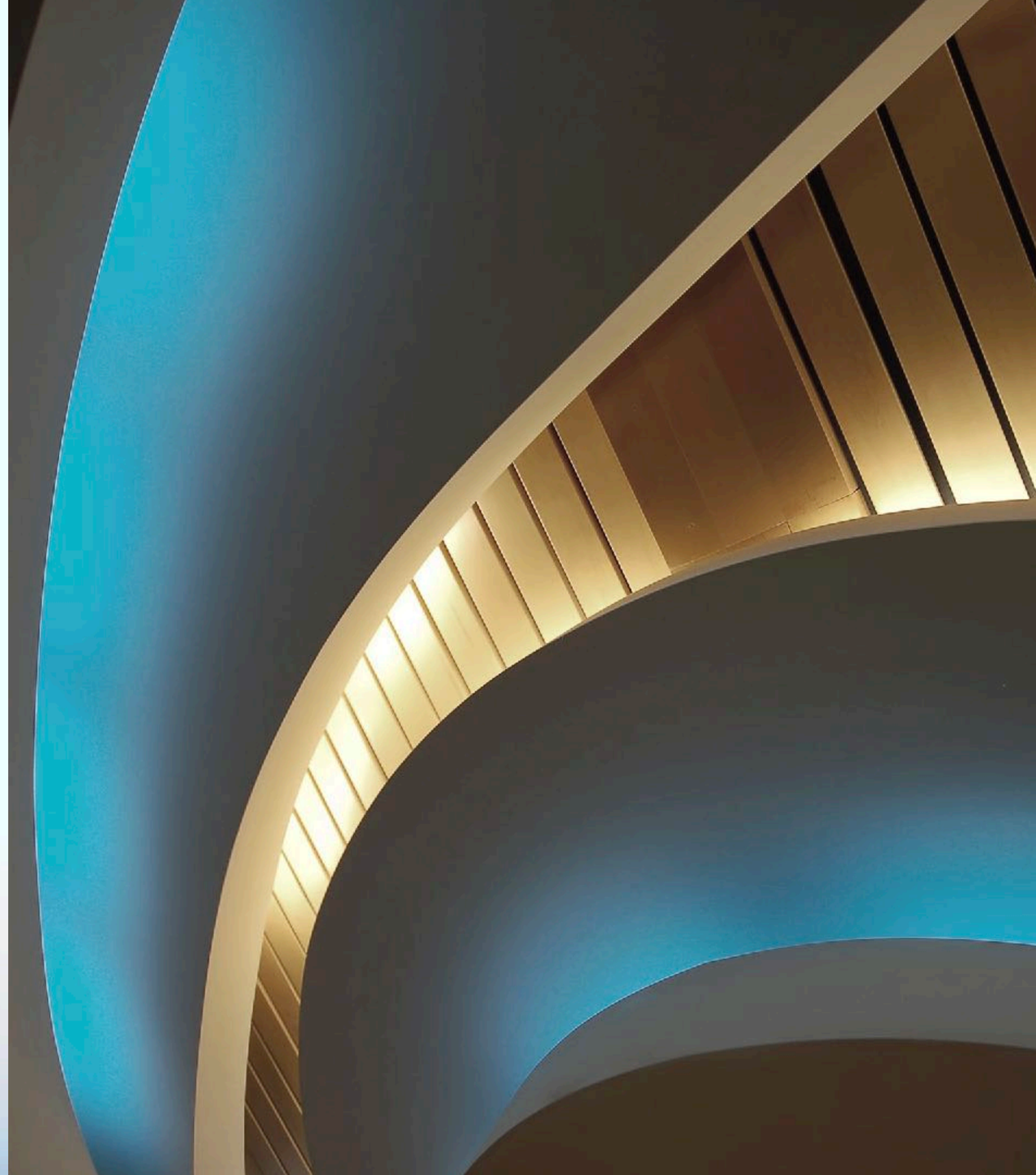
open pose

Stable diffusion


Midjourney

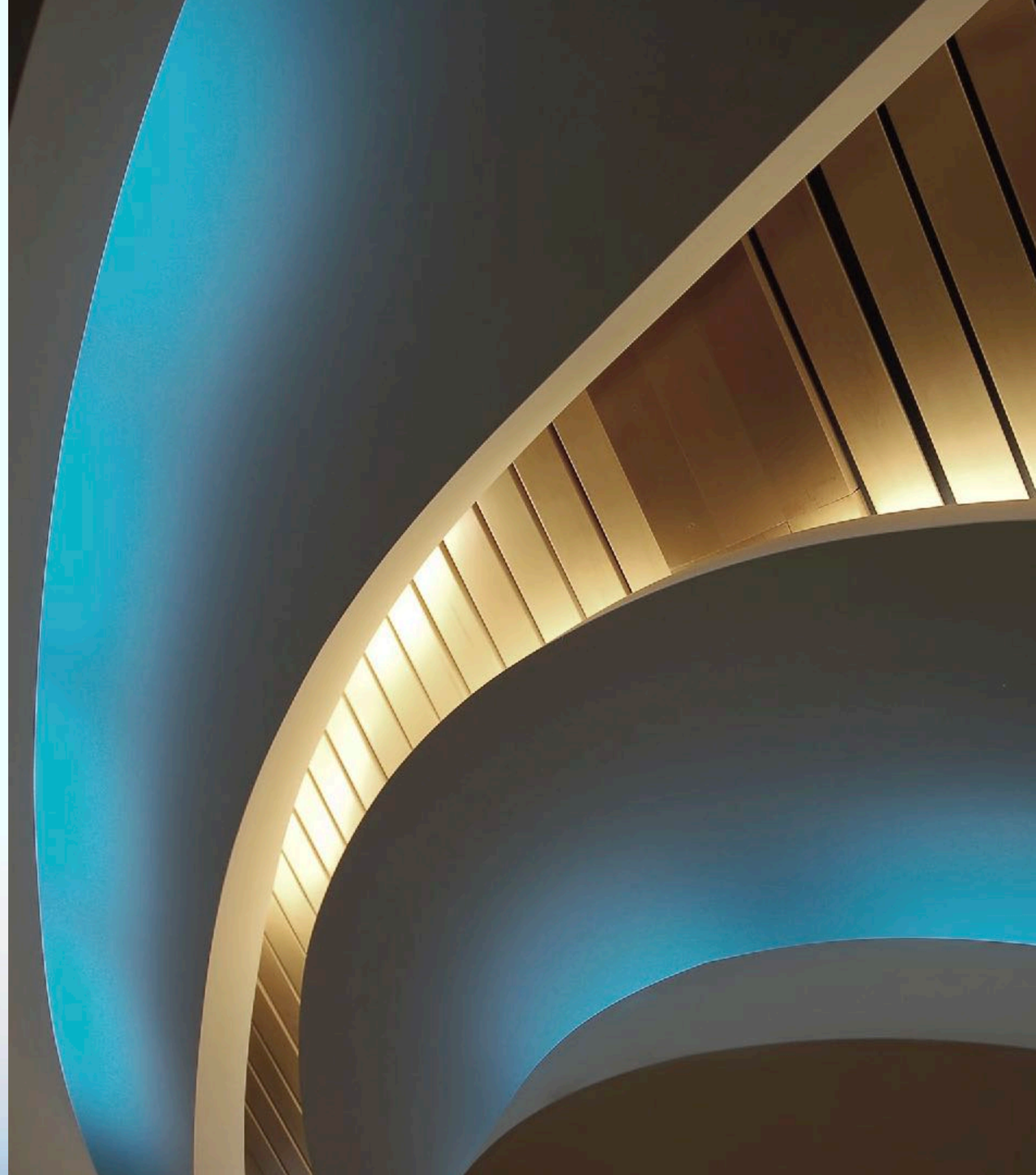
Agenda

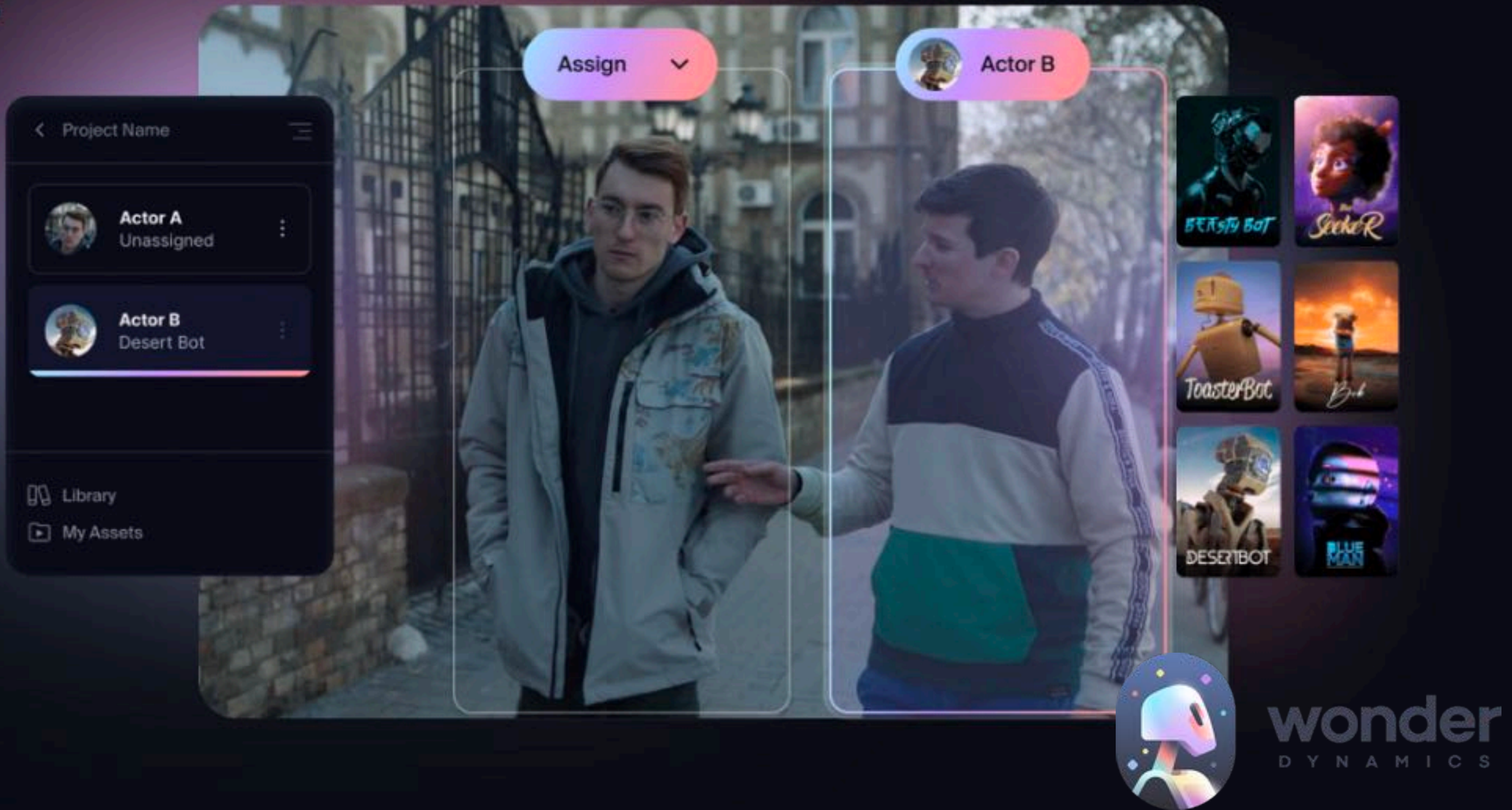
- LLMs & ChatGPT
- Voice
- Audio
- Image
- **Video**
- Multi-Modal
- Autonomous AI Agents

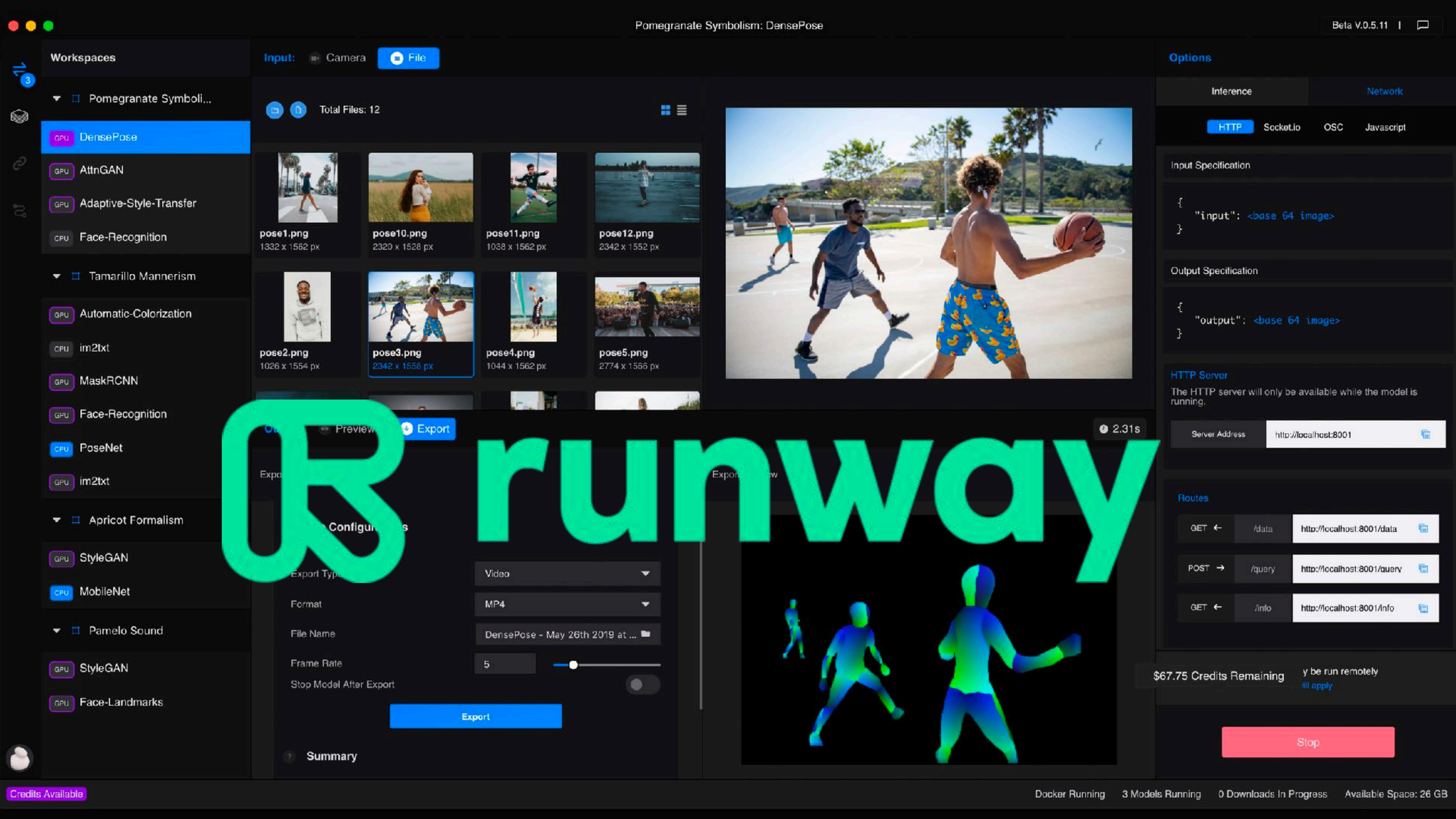


Agenda

- **Video**
  - Wonder Dynamics’ Wonder Studio AI
  - RunwayML

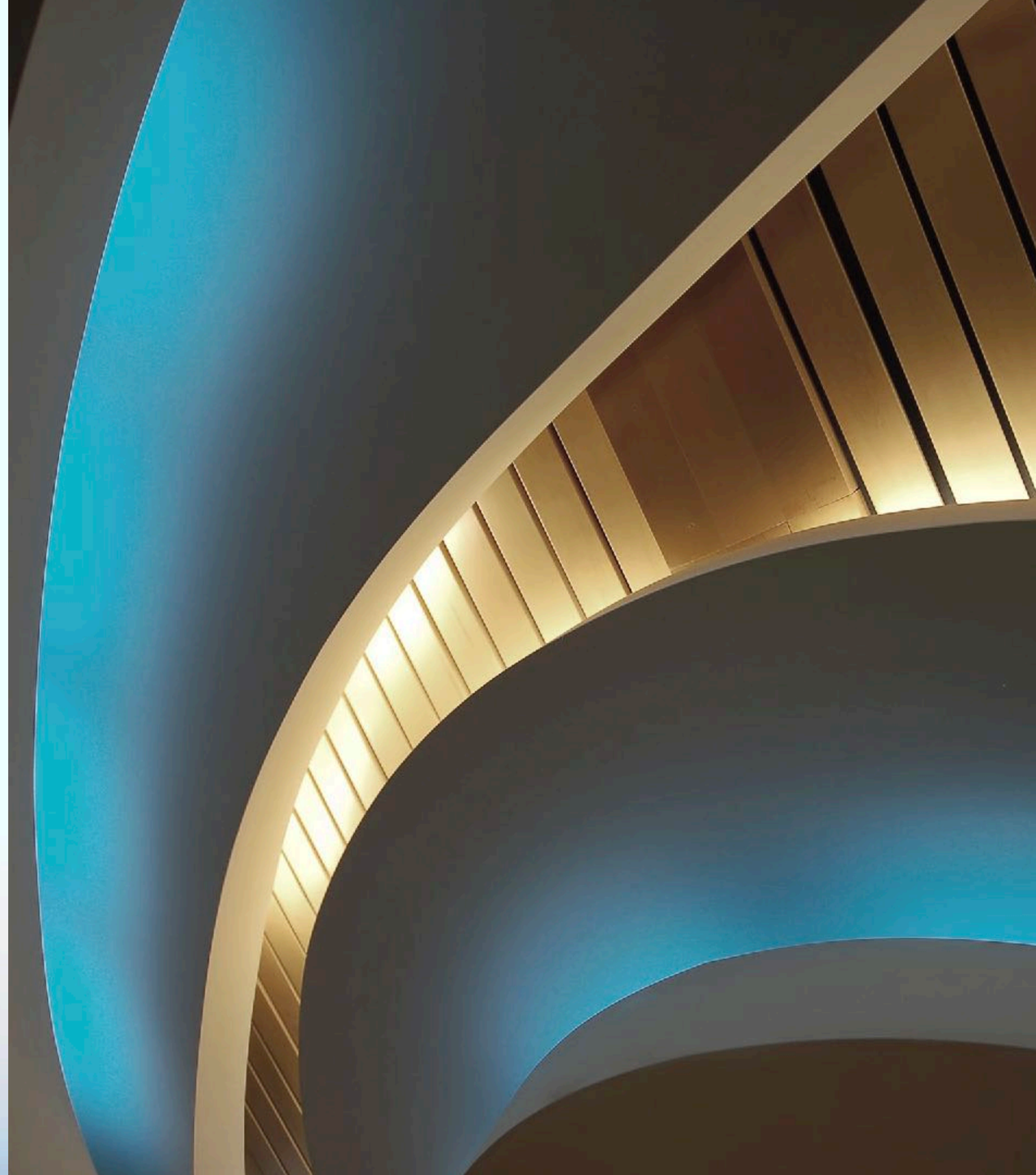



Assign
Actor B
Project Name
Actor A
Unassigned
Actor B
Desert Bot
Library
My Assets
BEASTY BOT
SEEKER
TOASTERBOT
DESERTBOT
BLUE MAN
wonder
DYNAMICS


Pomegranate Symbolism: DensePose
Beta V.0.5.11
Workspaces
Pomegranate Symboli...
DensePose
AttnGAN
Adaptive-Style-Transfer
Face-Recognition
Tamarillo Mannerism
Automatic-Colorization
im2txt
MaskRCNN
Face-Recognition
PoseNet
im2txt
Apricot Formalism
StyleGAN
MobileNet
Pamelo Sound
StyleGAN
Face-Landmarks
Input:
Camera
File
Total Files: 12
pose1.png
1332 x 1562 px
pose10.png
2320 x 1528 px
pose11.png
1038 x 1562 px
pose12.png
2342 x 1552 px
pose2.png
1026 x 1554 px
pose3.png
2342 x 1558 px
pose4.png
1044 x 1562 px
pose5.png
2774 x 1556 px
Preview
Export
2.31s
Export Type
Video
Format
MP4
File Name
DensePose - May 26th 2019 at ...
Frame Rate
5
Stop Model After Export
Export
Summary
Options
Inference
Network
HTTP
Socket.io
OSC
Javascript
Input Specification
{
"input": <base 64 image>
}
Output Specification
{
"output": <base 64 image>
}
HTTP Server
The HTTP server will only be available while the model is running.
Server Address
http://localhost:8001
Routes
GET ← /data http://localhost:8001/data
POST → /query http://localhost:8001/query
GET ← /info http://localhost:8001/info
$67.75 Credits Remaining
Stop
runway
Credits Available
Docker Running
3 Models Running
0 Downloads In Progress
Available Space: 26 GB

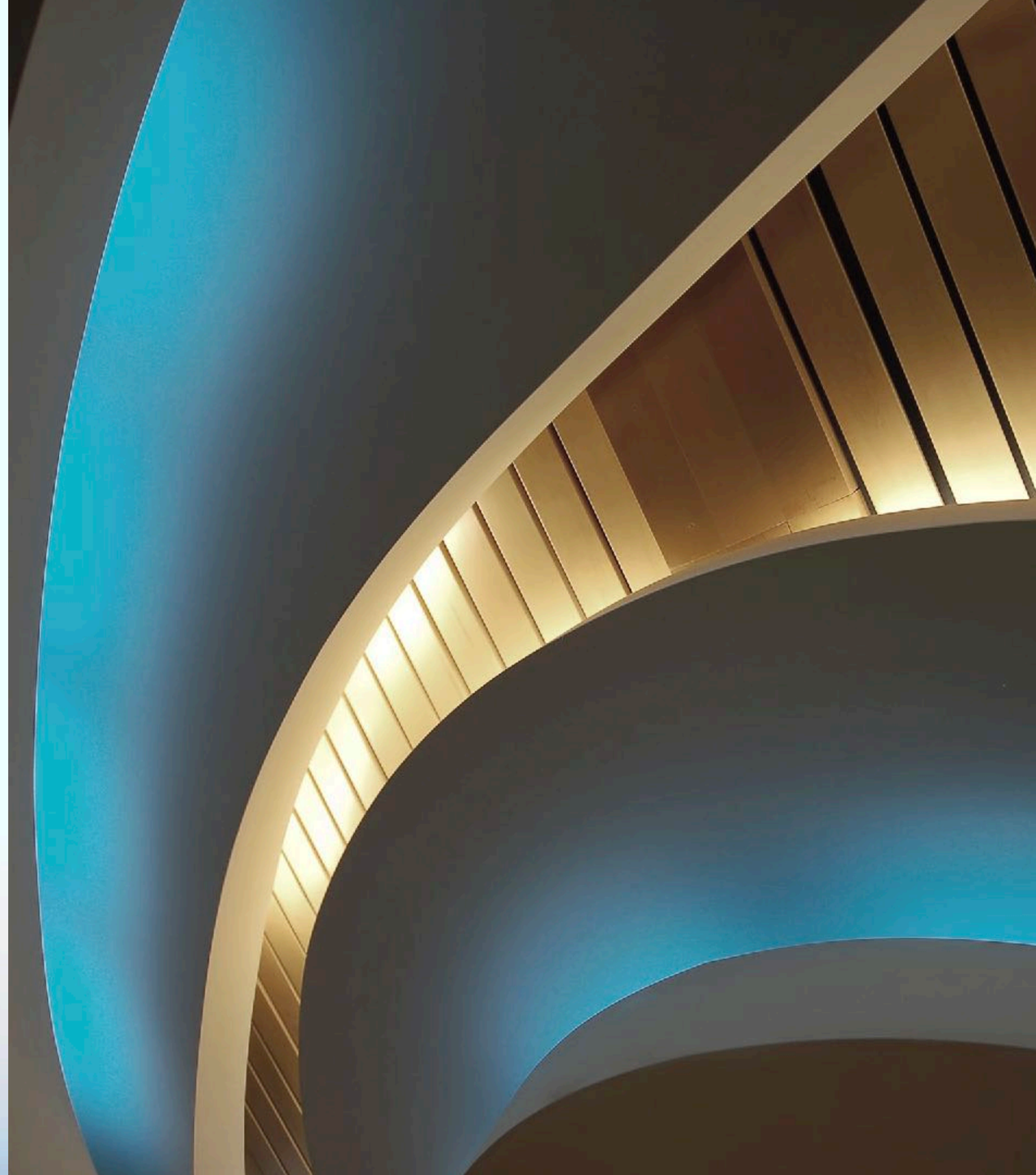
Agenda

- LLMs & ChatGPT
- Voice
- Audio
- Image
- Video
- **Multi-Modal**
- Autonomous AI Agents



Agenda

- **Multi-Modal**
  - ChatGPT does it all



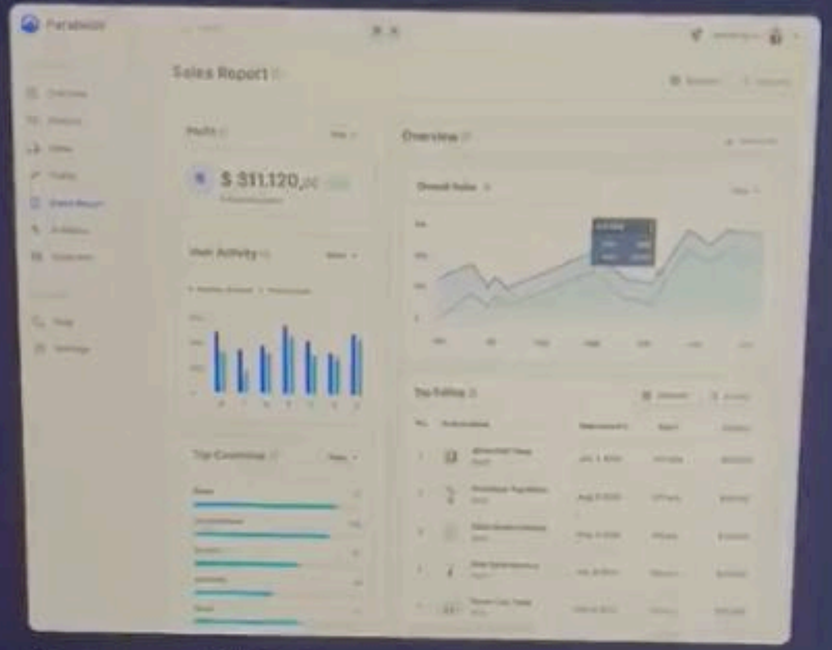
GPT-4 with vision ALPHA

NA

Is this good for my burn

Yes, the cream is for soothing burns, sunburns, and wounds. Use as directed. If concerned, consult a healthcare professional.

GPT-4



You are an expert UI/UX designer and software developer. Break down this SaaS dashboard into components. You are going to pass this plan to a software dev on your team to build.

Certainly! Let's break down the SaaS dashboard into its main components and sub-components:

1. Sidebar Menu:

- **Category List**: A vertical list of categories.
  - **Icons**: Representing each category.
  - **Labels**: Textual representation of each category.

2. Top Navigation Bar:

Regenerate

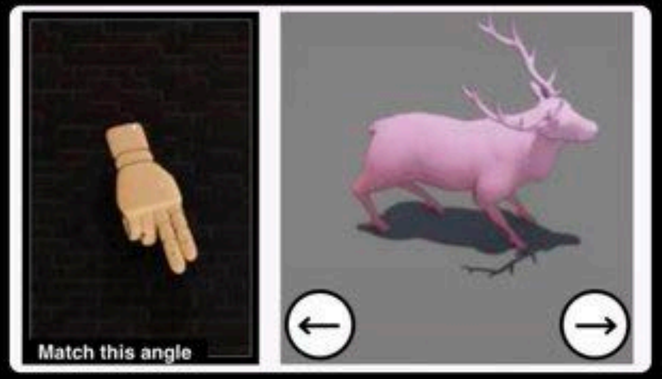
Send a message

ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT September 25 Version

It's Wednesday at 4 pm. Can I park at this spot right now? Tell me in 1 line.

Yes, you can park for up to 1 hour starting at 4 pm.

ME

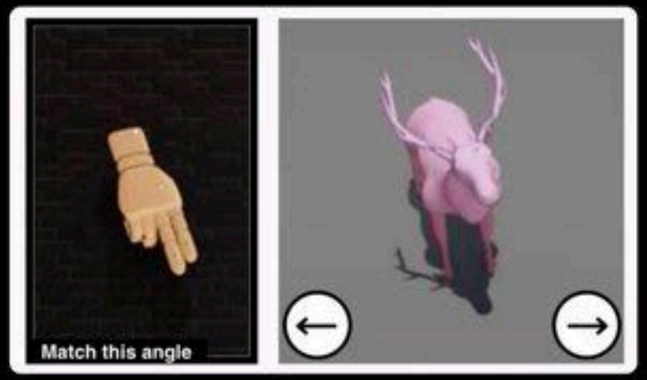


Does the angle of the image on the right match the angle of the hand on the left? Yes or no.

CHATGPT

No.

ME



How about this one?

CHATGPT

Yes.

17:41

ME

Attached is a 3 by 3 image matrix. Let's number them from left to right as 1-9. Please tell me which of them contains stairs.

CHATGPT

The images containing stairs are numbers 2, 3, and 5.

GPT-3.5 GPT-4

ChatGPT

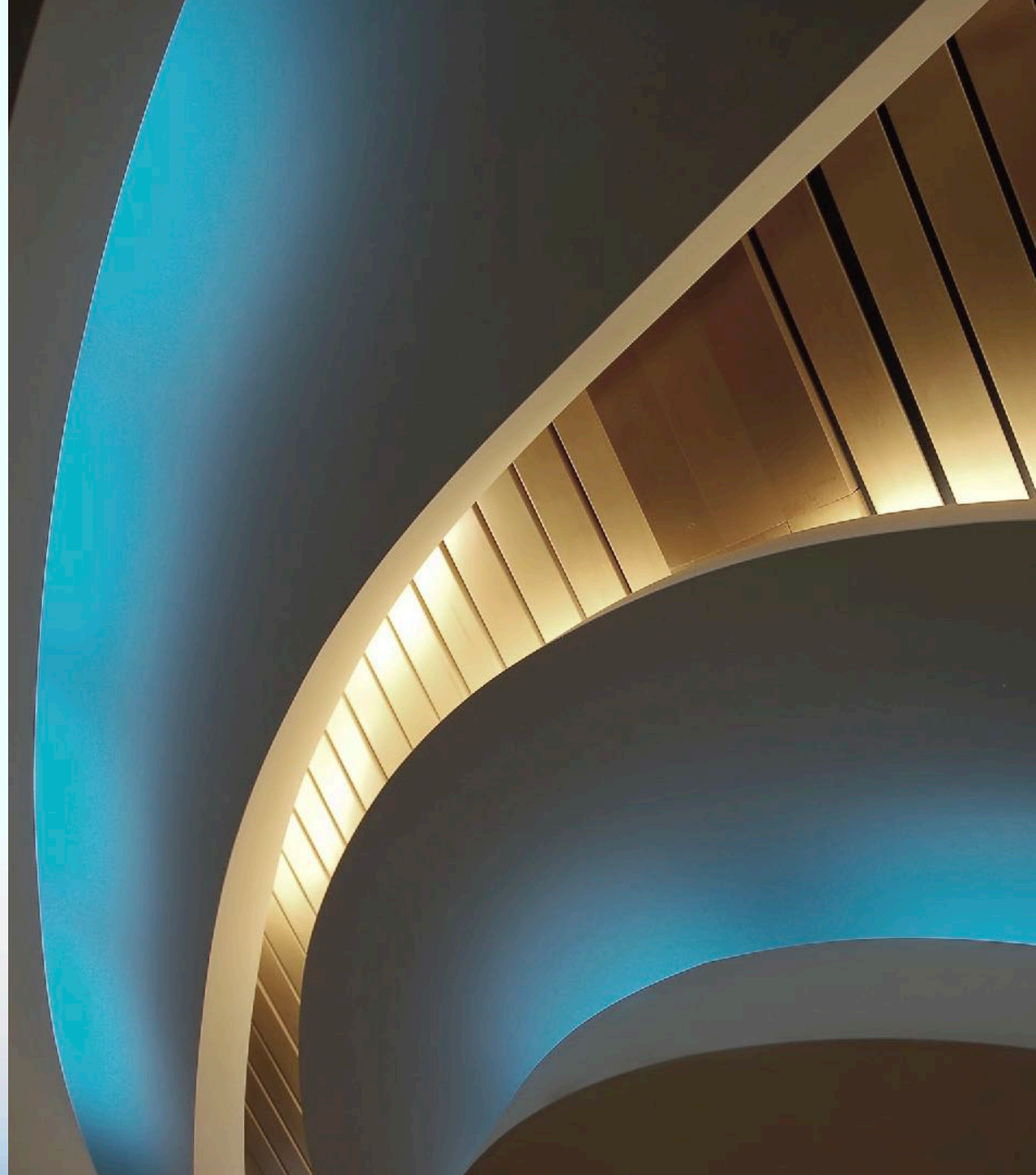
Come up with concepts
for a retro-style arcade game

Design a data
for an online m

Message

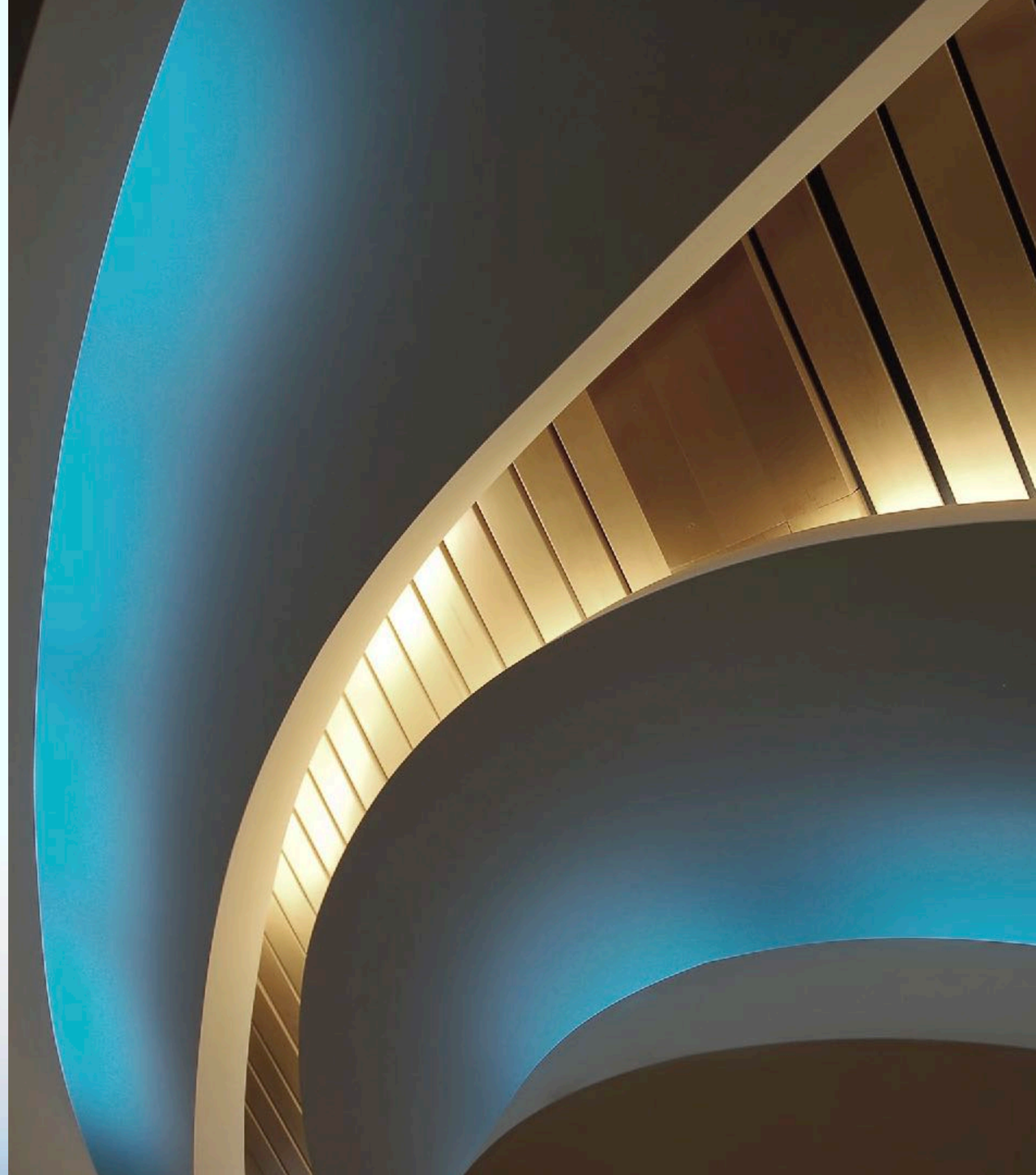
Agenda

- LLMs & ChatGPT
- Voice
- Audio
- Image
- Video
- Multi-Modal
- **Autonomous AI Agents**



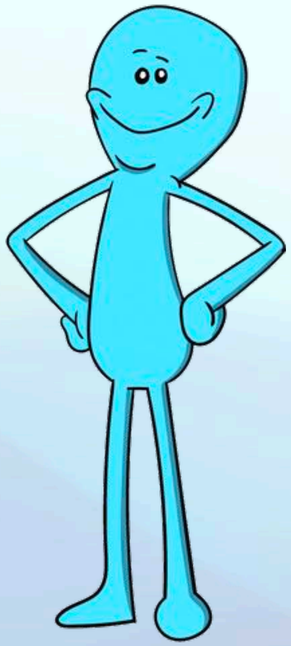
Agenda

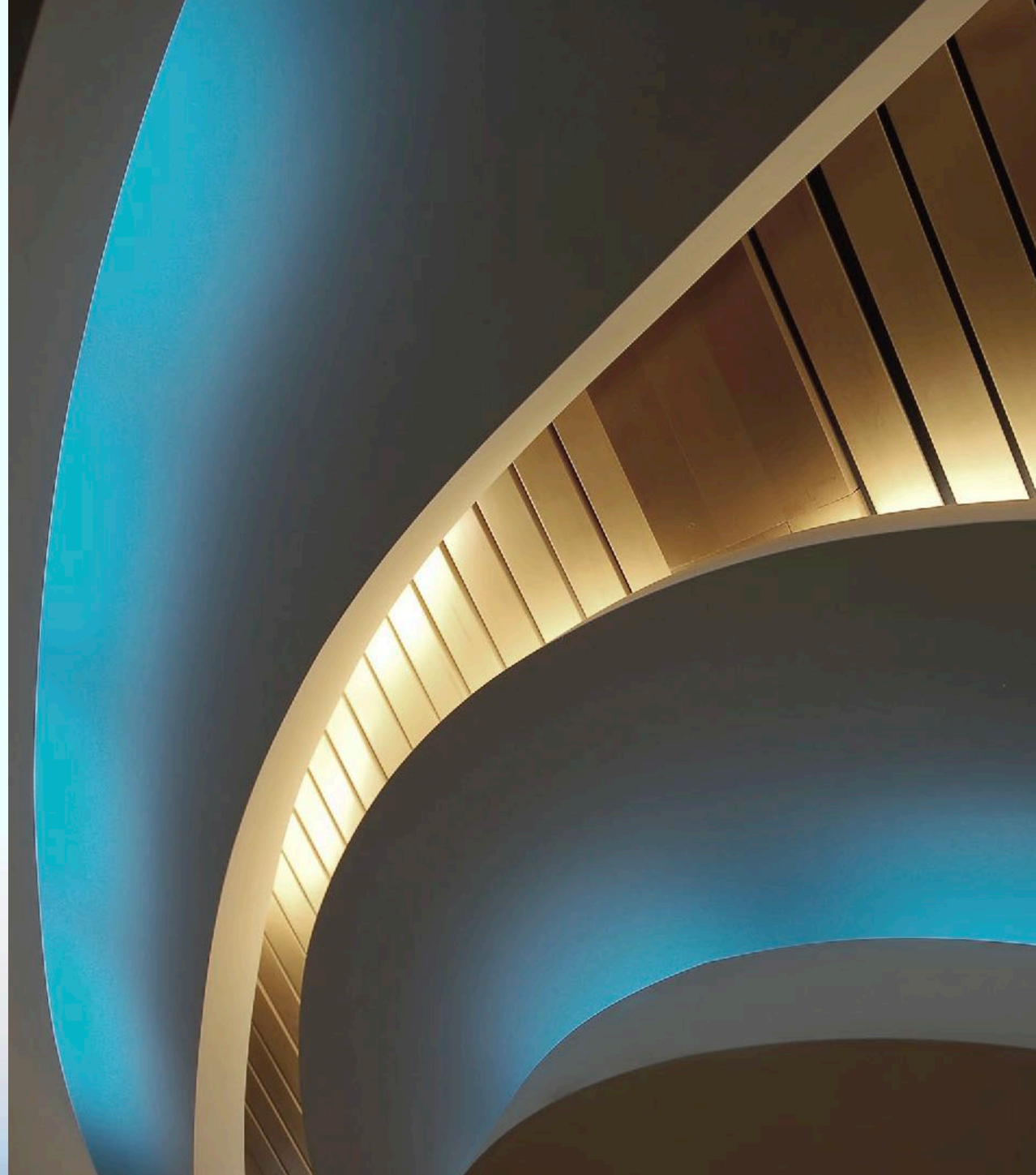
- **Autonomous AI Agents**
  - AutoGPT
  - BabyAGI



Agenda

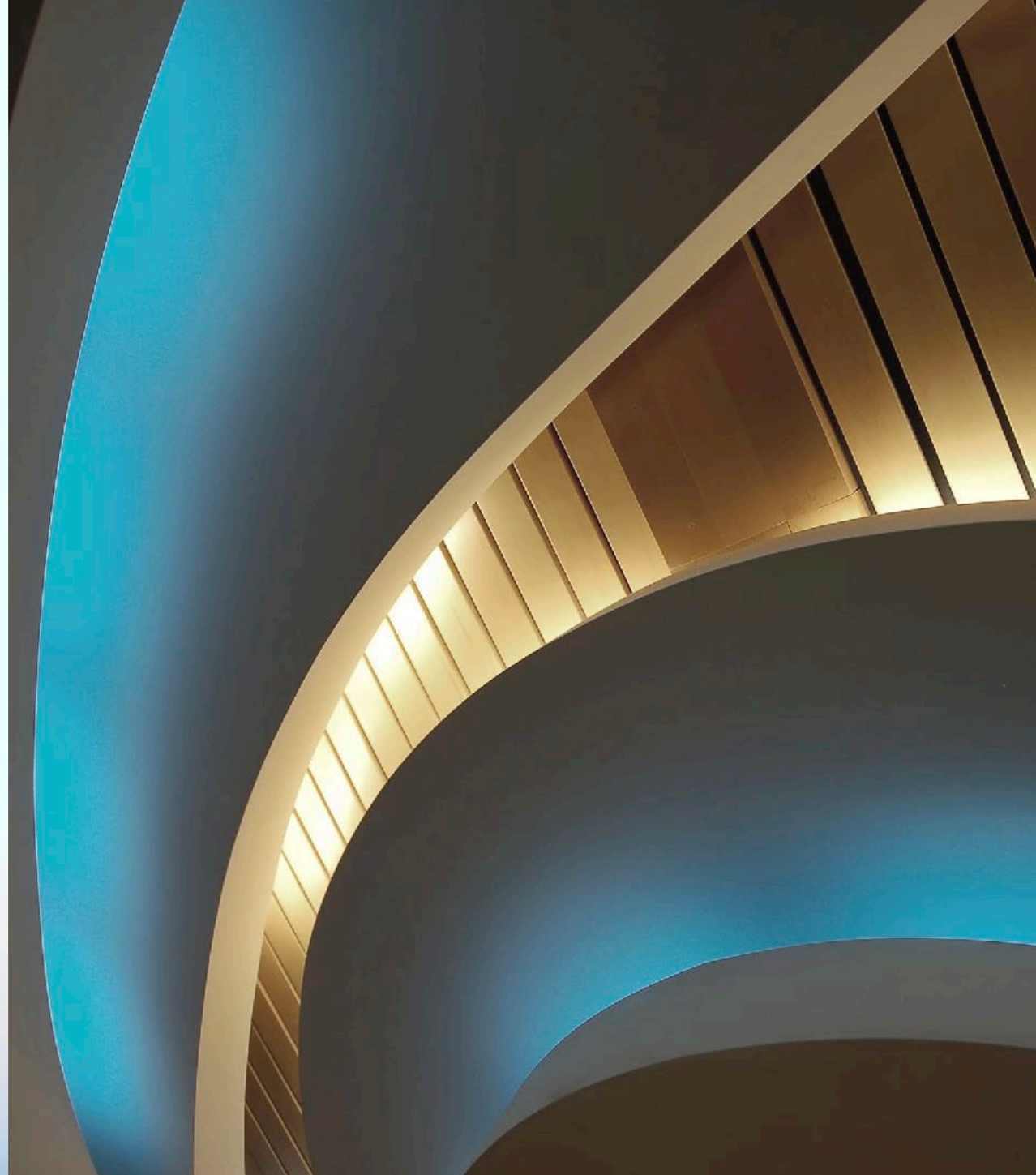
- **Autonomous AI Agents**
  - AutoGPT
  - BabyAGI





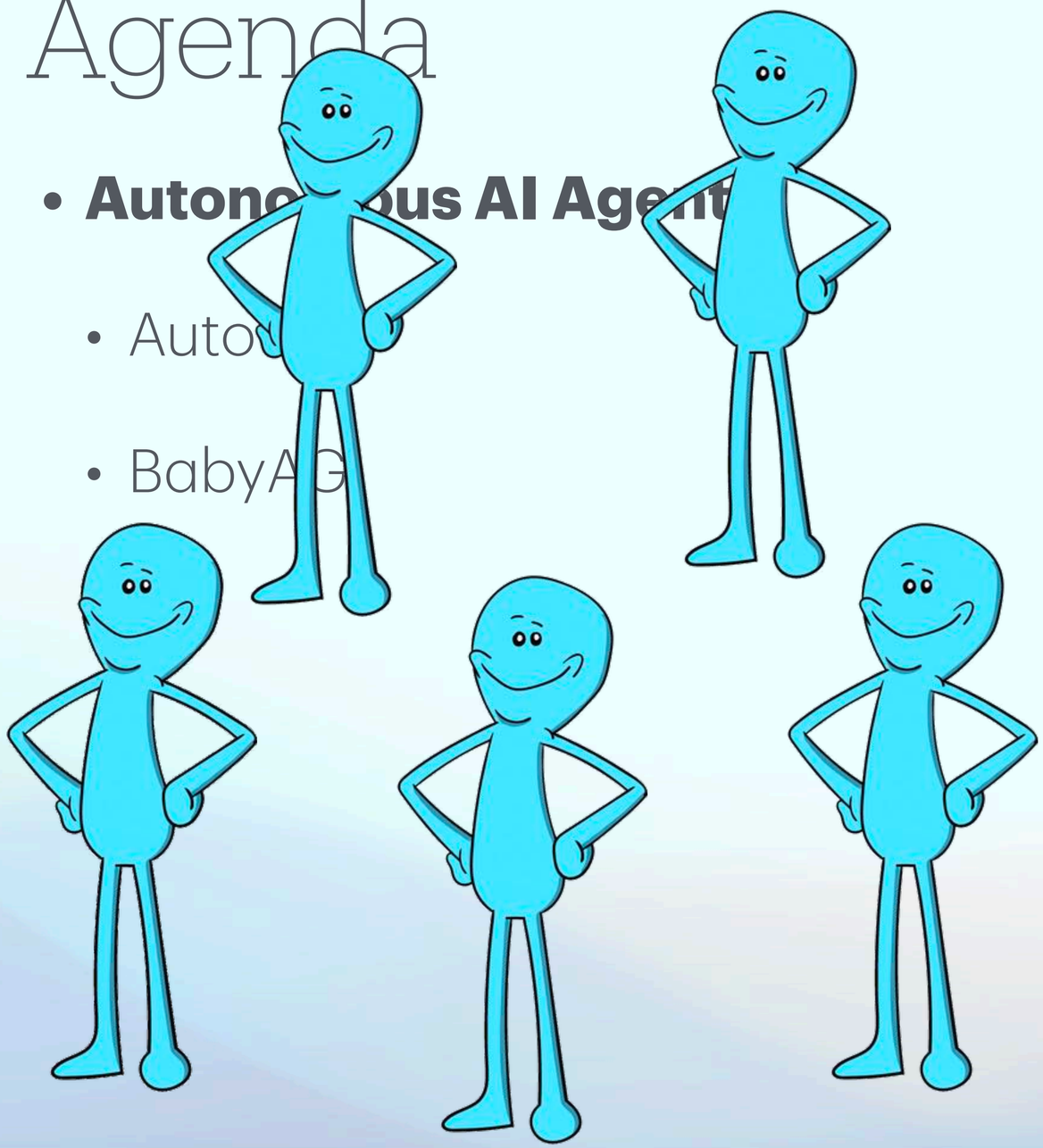
Agenda

- **Autonomous AI Agents**
  - AutoGPT
  - BabyAGI



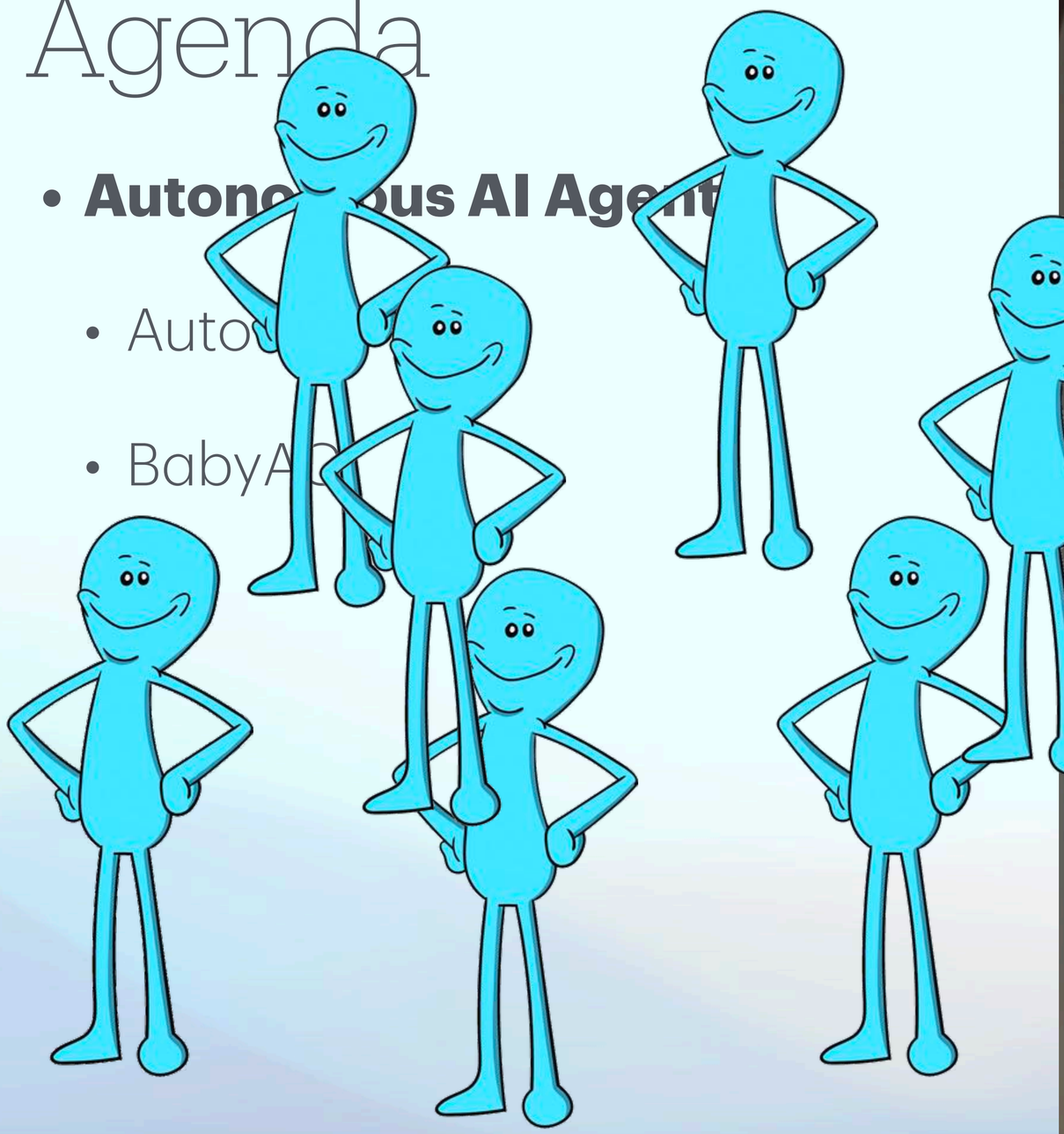
Agenda

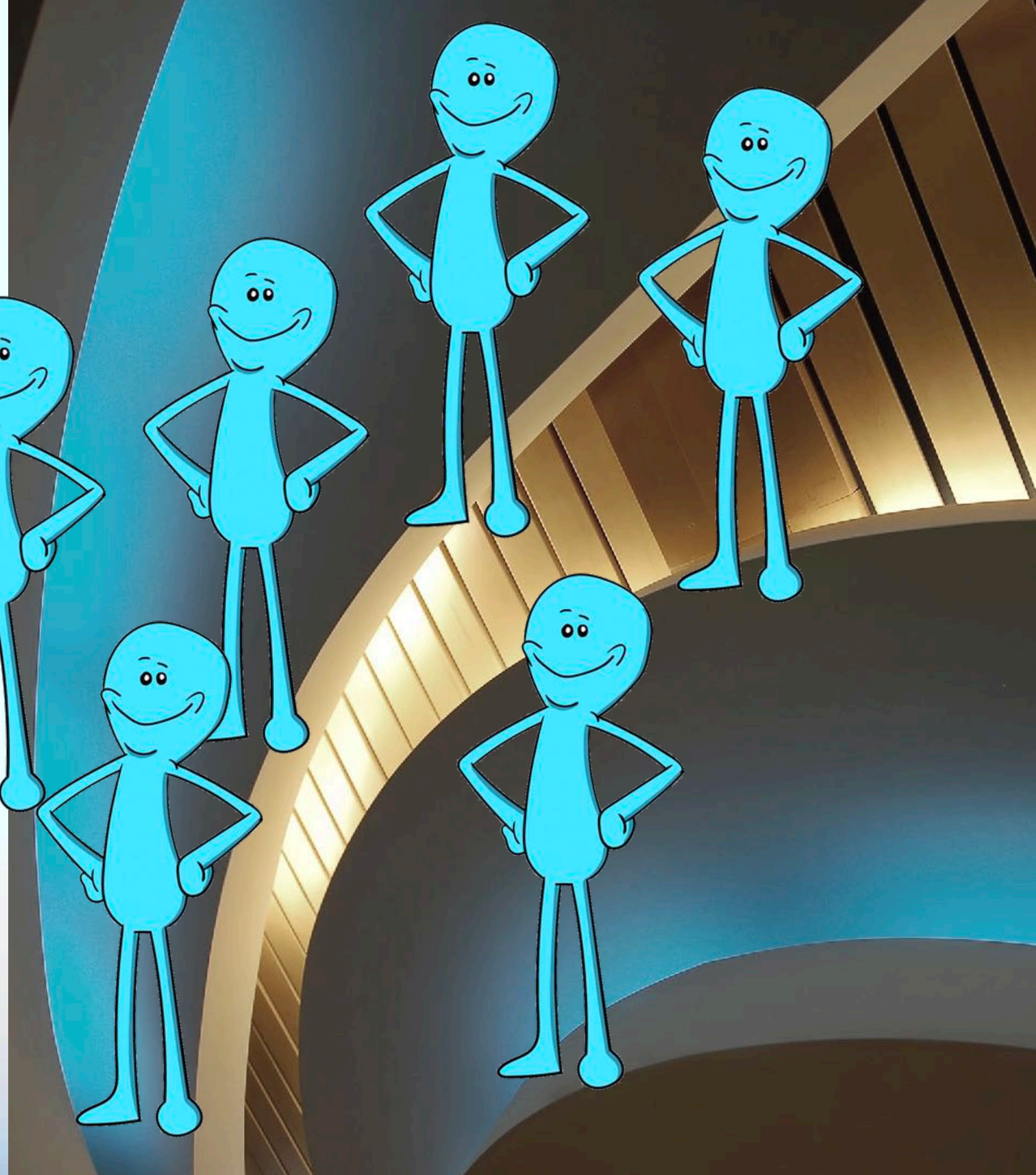
- **Autonomous AI Agents**
  - Auto
  - BabyAG



Agenda

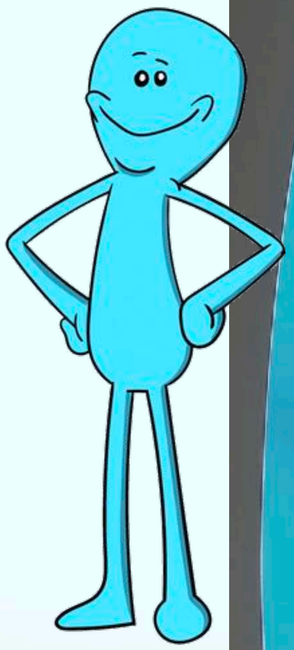
- **Autono[illegible]ous AI Agent**
  - Auto[illegible]
  - BabyA[illegible]

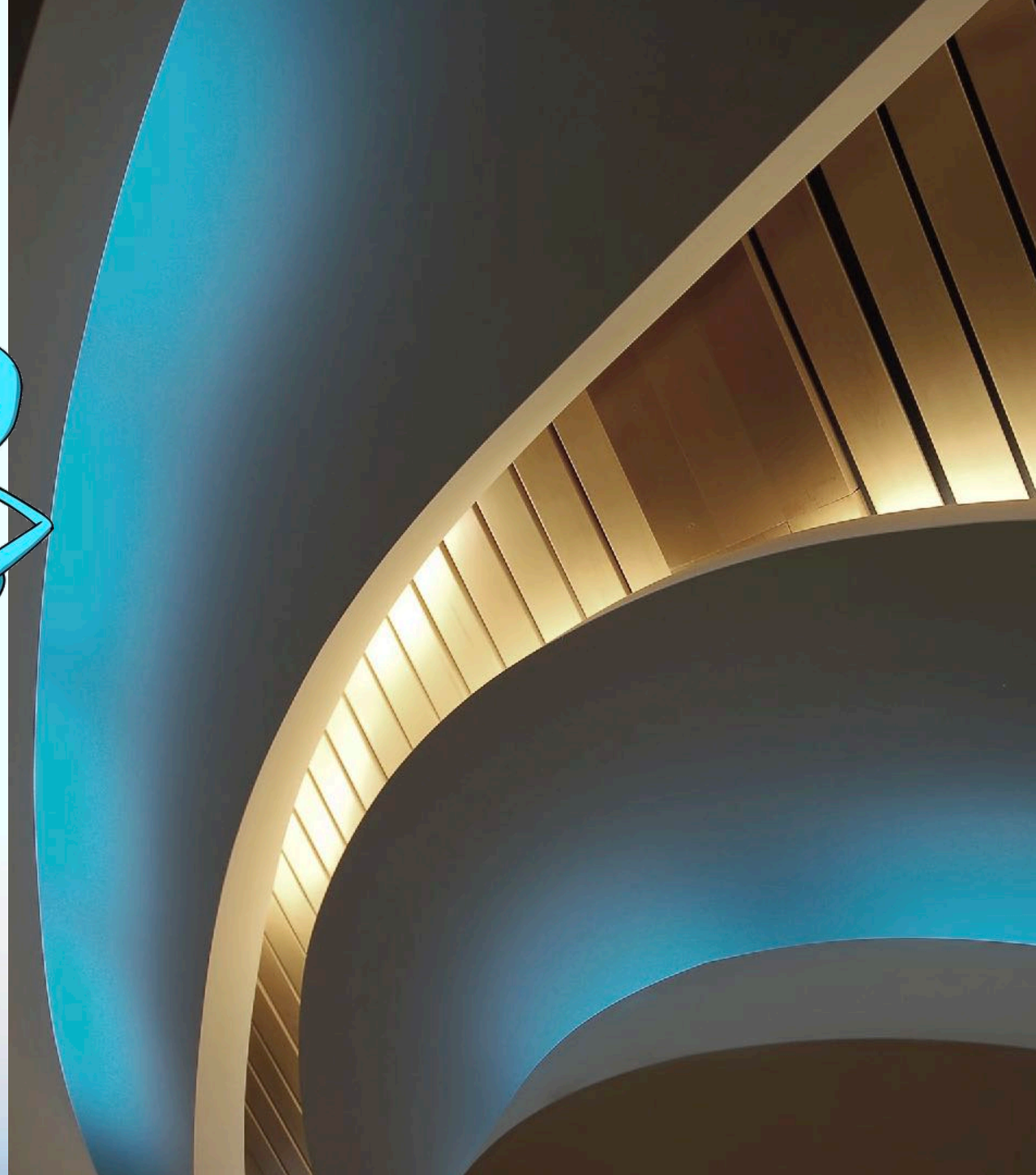




Agenda

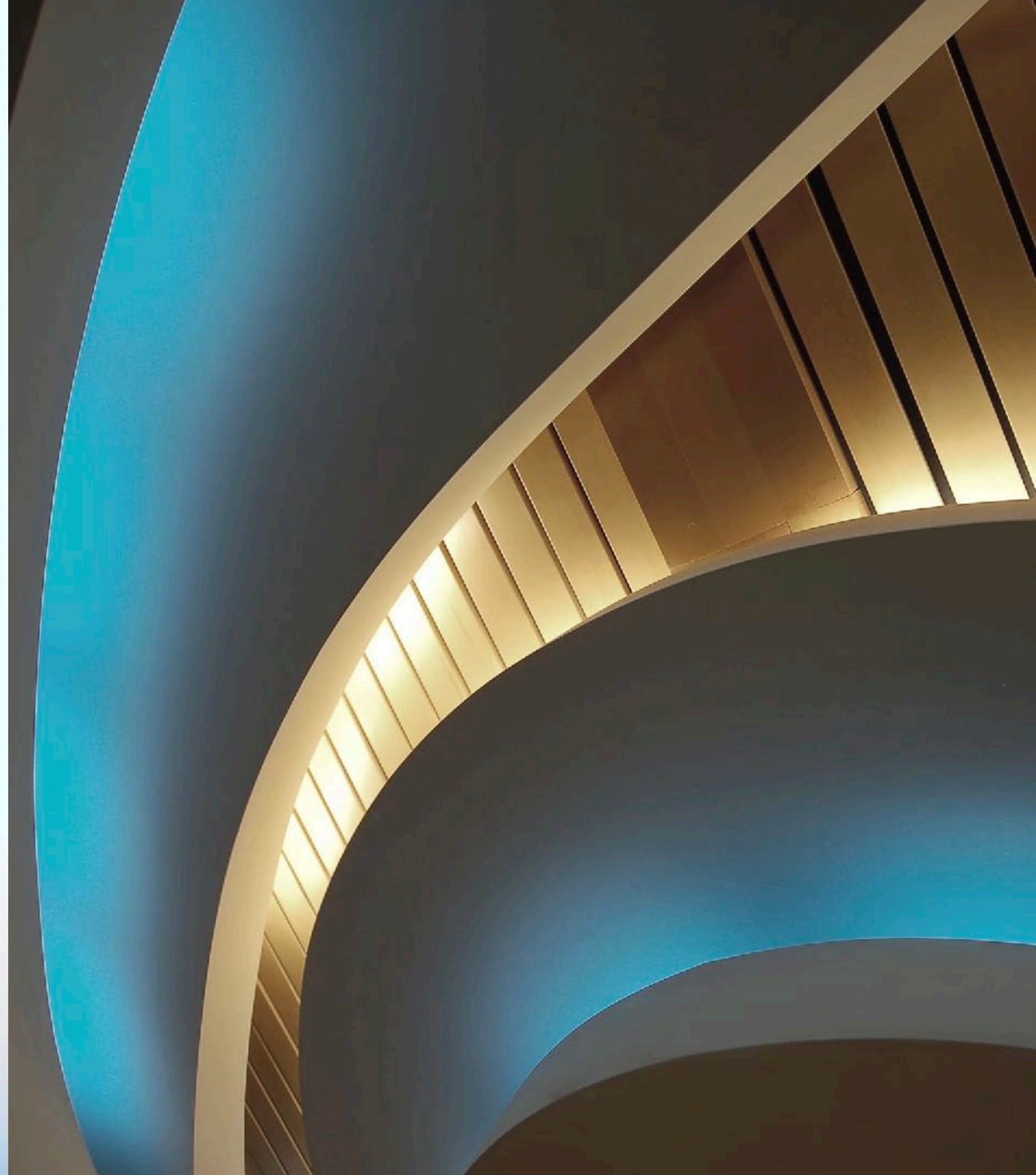
- **Autonomous AI Agents**
  - AutoGPT
  - BabyAGI





Agenda

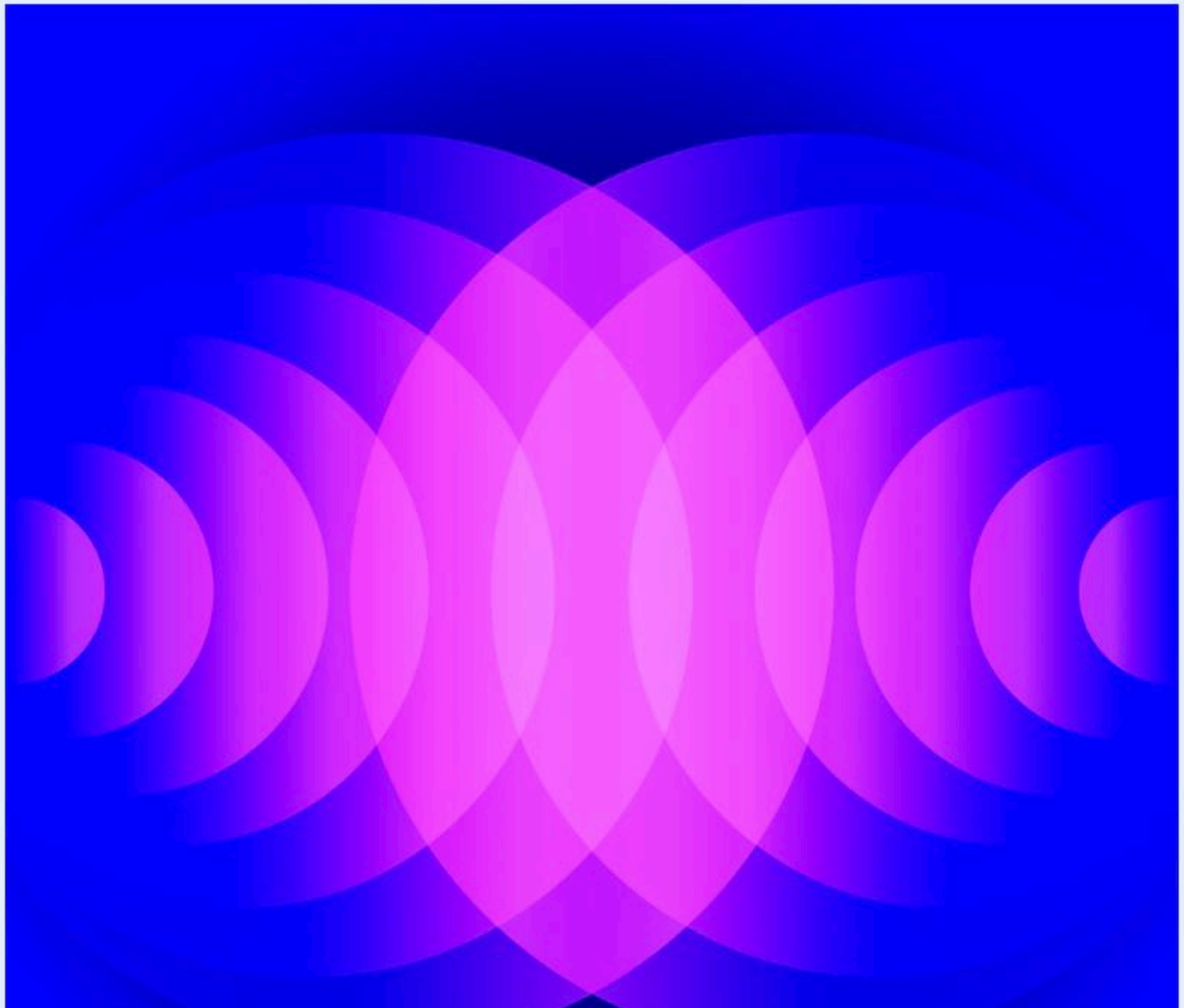
- **Autonomous AI Agents**
  - AutoGPT
  - BabyAGI



BabyAGI

OpenAI cybersecurity grant program

Our goal is to facilitate the development of AI-powered cybersecurity capabilities for defenders through grants and other support.

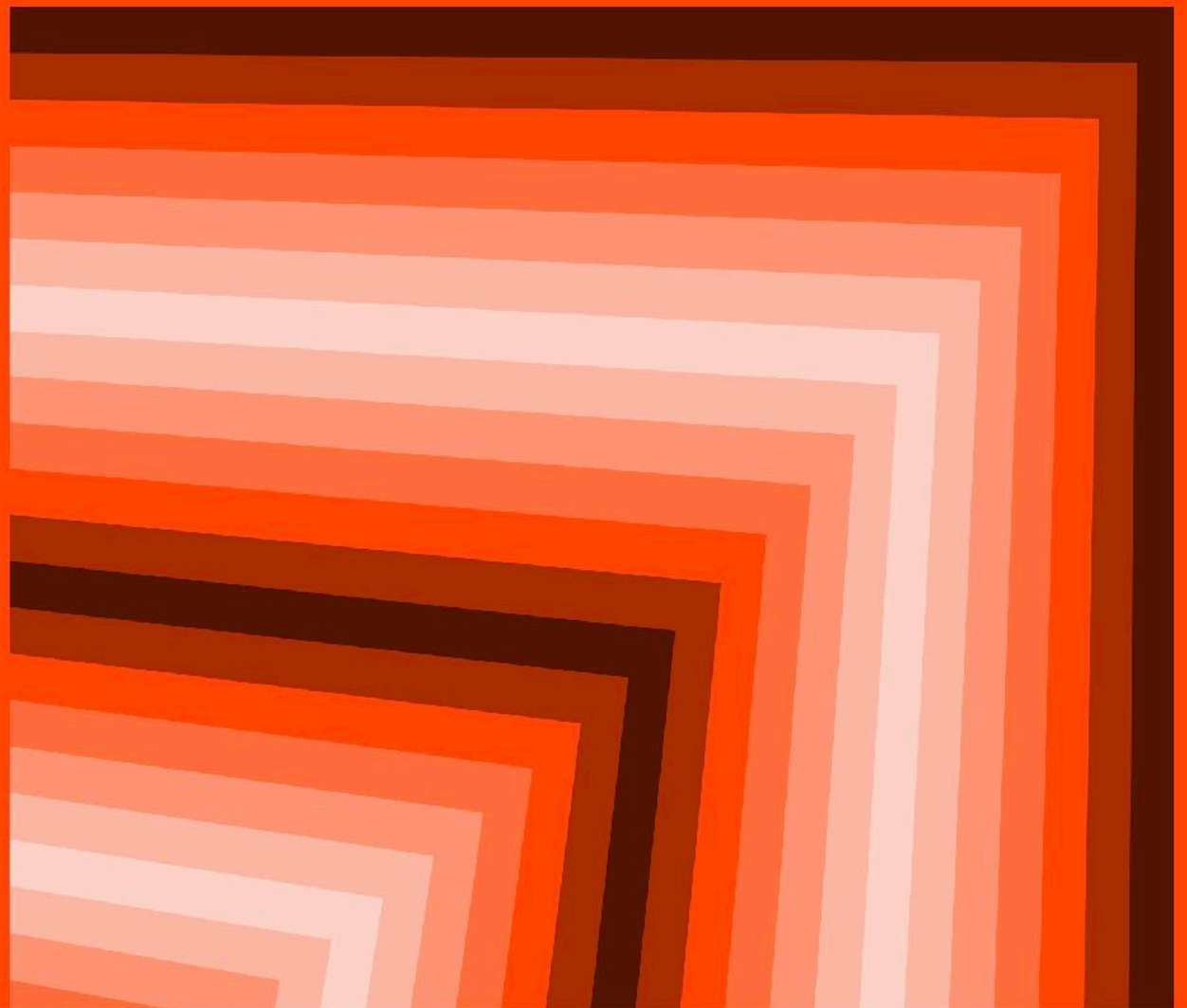


Blog

OpenAI Red Teaming Network

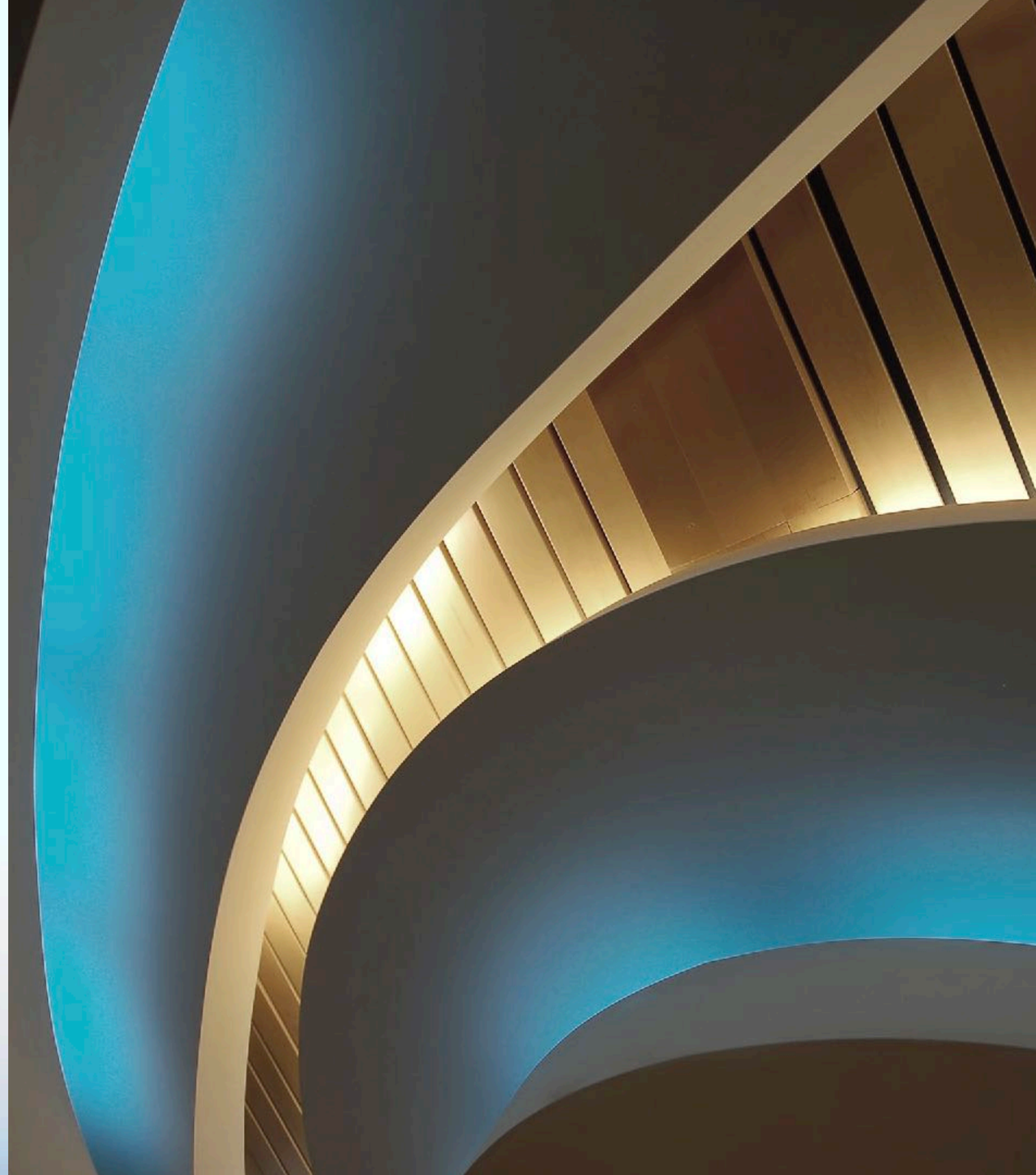
We're announcing an open call for the OpenAI Red Teaming Network and invite domain experts interested in improving the safety of OpenAI's models to join our efforts.

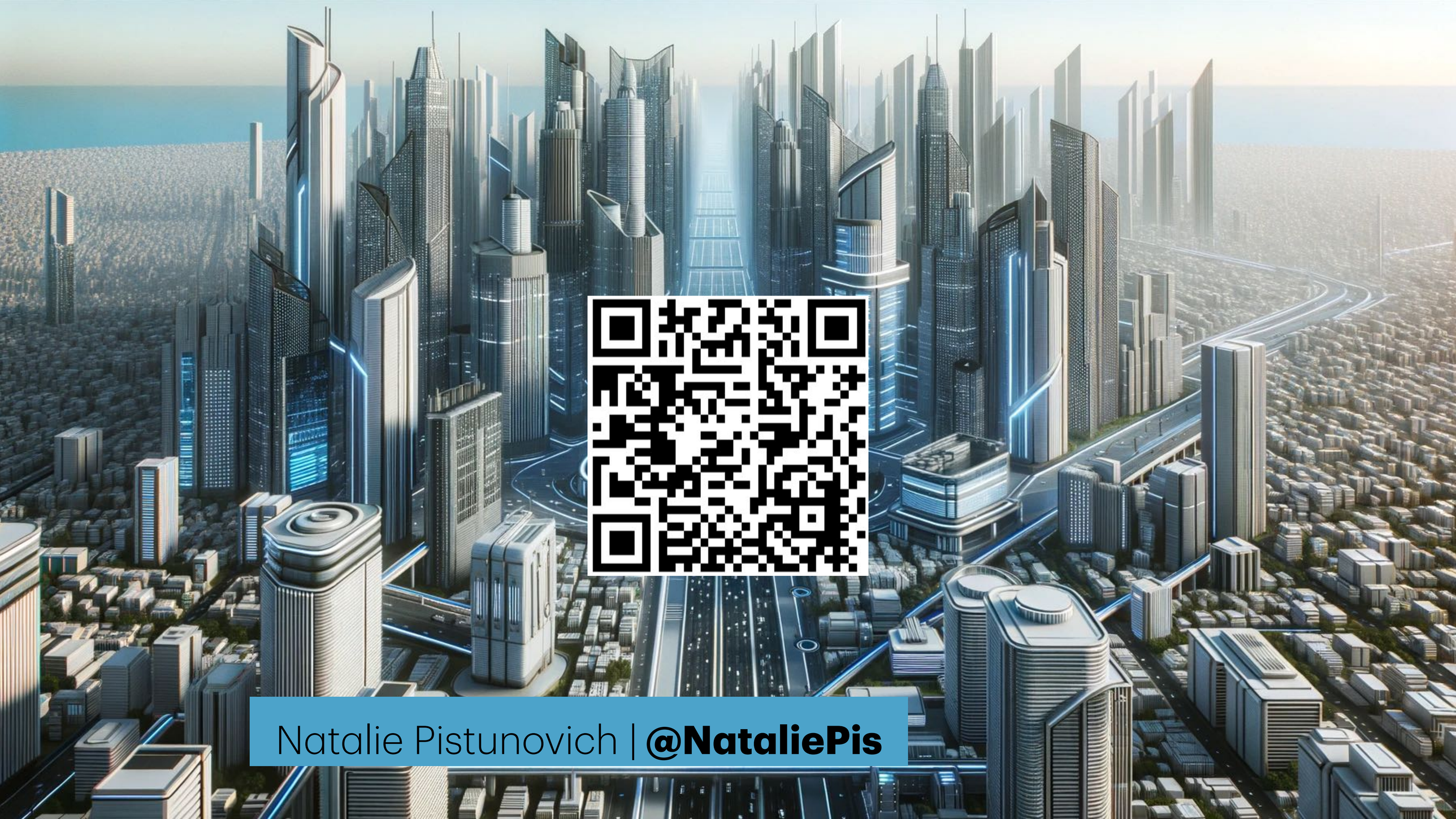
Apply to join



Recap

- LLMs & ChatGPT
- Voice
- Audio
- Image
- Video
- Multi-Modal
- Autonomous AI Agents





Natalie Pistunovich | **@NataliePis**