

The Era of Large Models and Expanding Horizons

– Applications and Challenges –

Elliott Wu

GPT-4



SORA



SAM 3



SAM 3D



Genie



Applications and Challenges – Outline

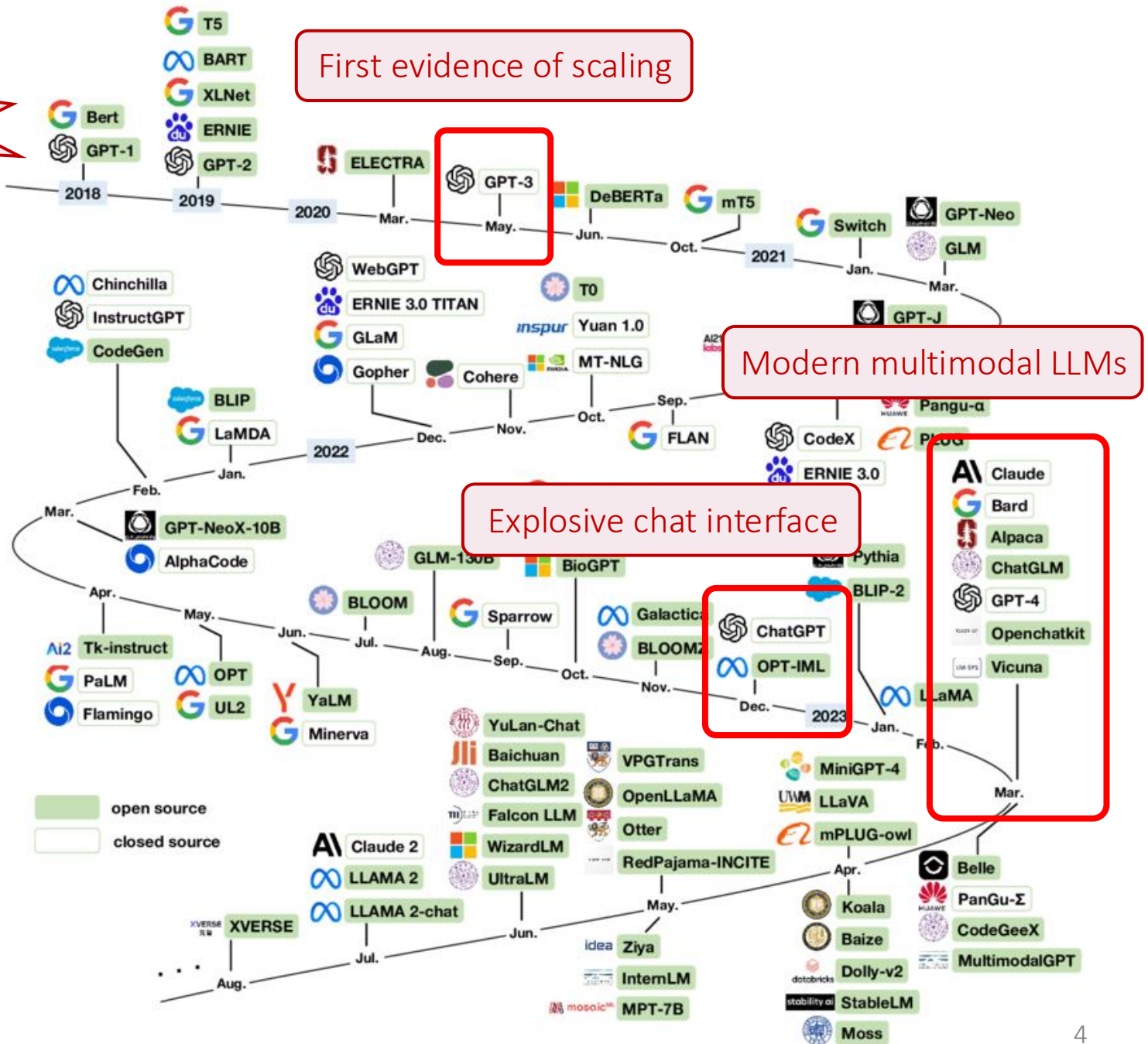
- Part 1 – The Era of Large Models
- Part 2 – Expanding Horizons

Part 1 – The Era of Large Models

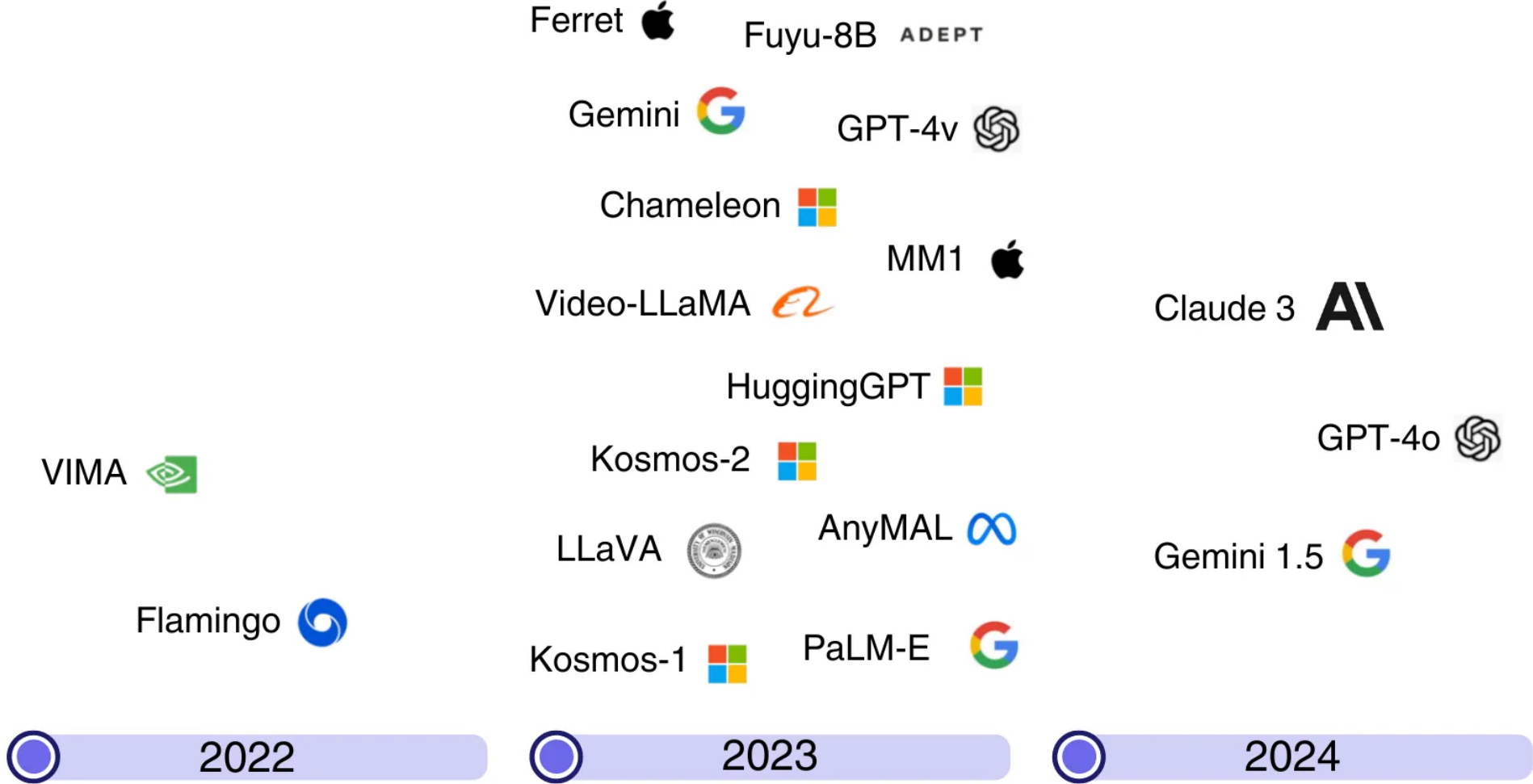


A Brief History of LLMs (2018-2023)

- GPT-1 (2018): 117M params
- GPT-2 (2019): 1.5B params
- GPT-3 (2020): 175B params
- GPT-4 (2023): >1T params



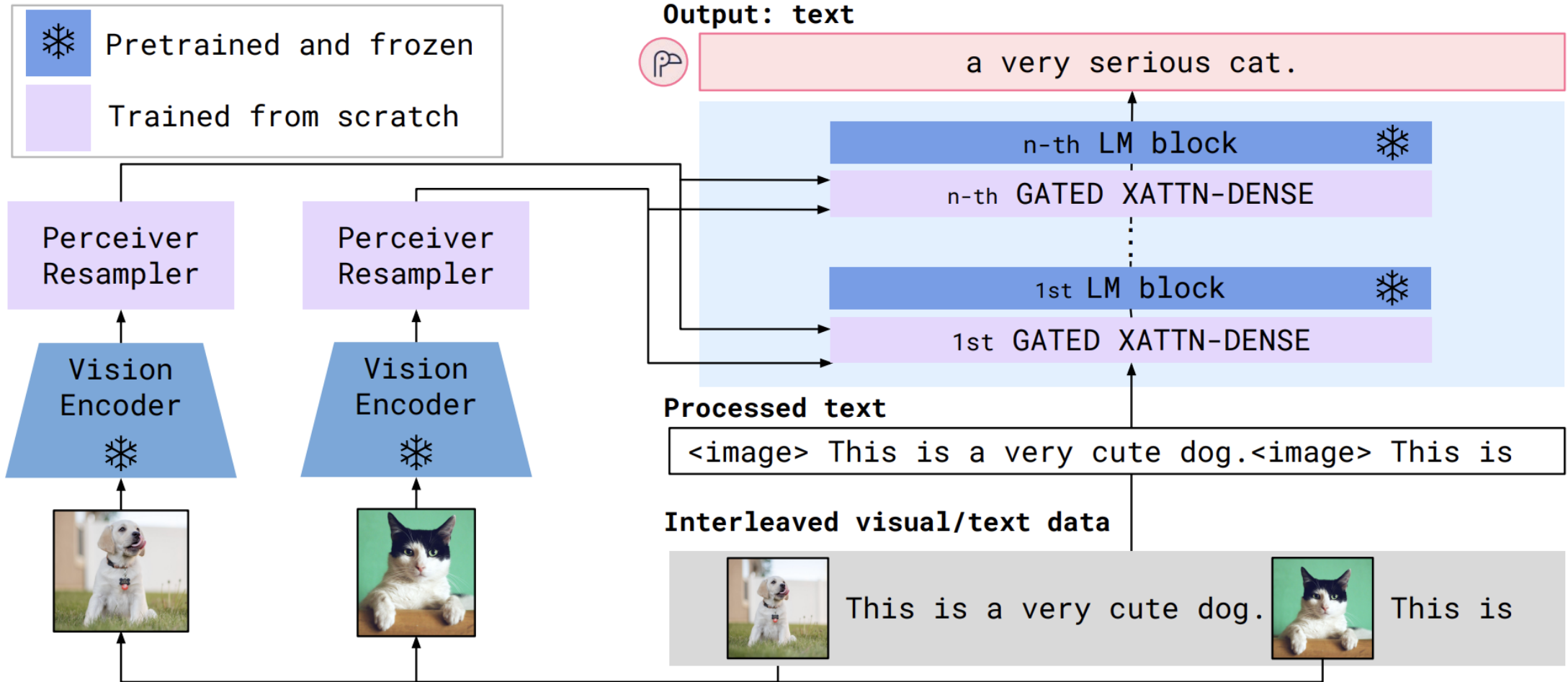
Multimodal Language Models (2022-2024)



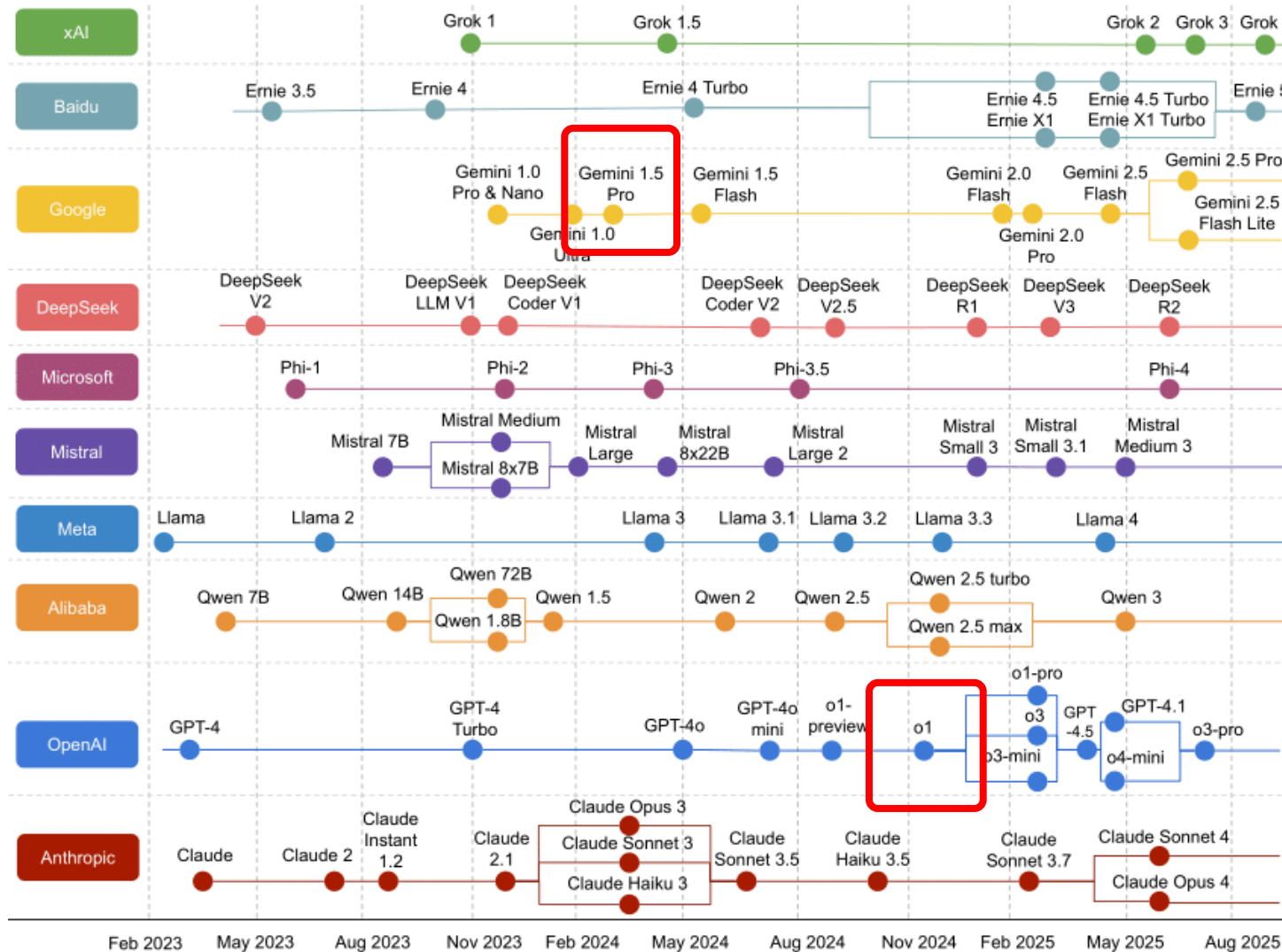
https://medium.com/@tenyks_blogger/multimodal-large-language-models-mlms-transforming-computer-vision-76d3c5dd267f



Flamingo: a Visual Language Model for Few-Shot Learning



Long Context and Reasoning (2024-2025)



1M token context window
(hour-long video)

Chain-of-Thought (CoT)
reasoning, solving 83%
of IMO questions

As of 5 March 2026

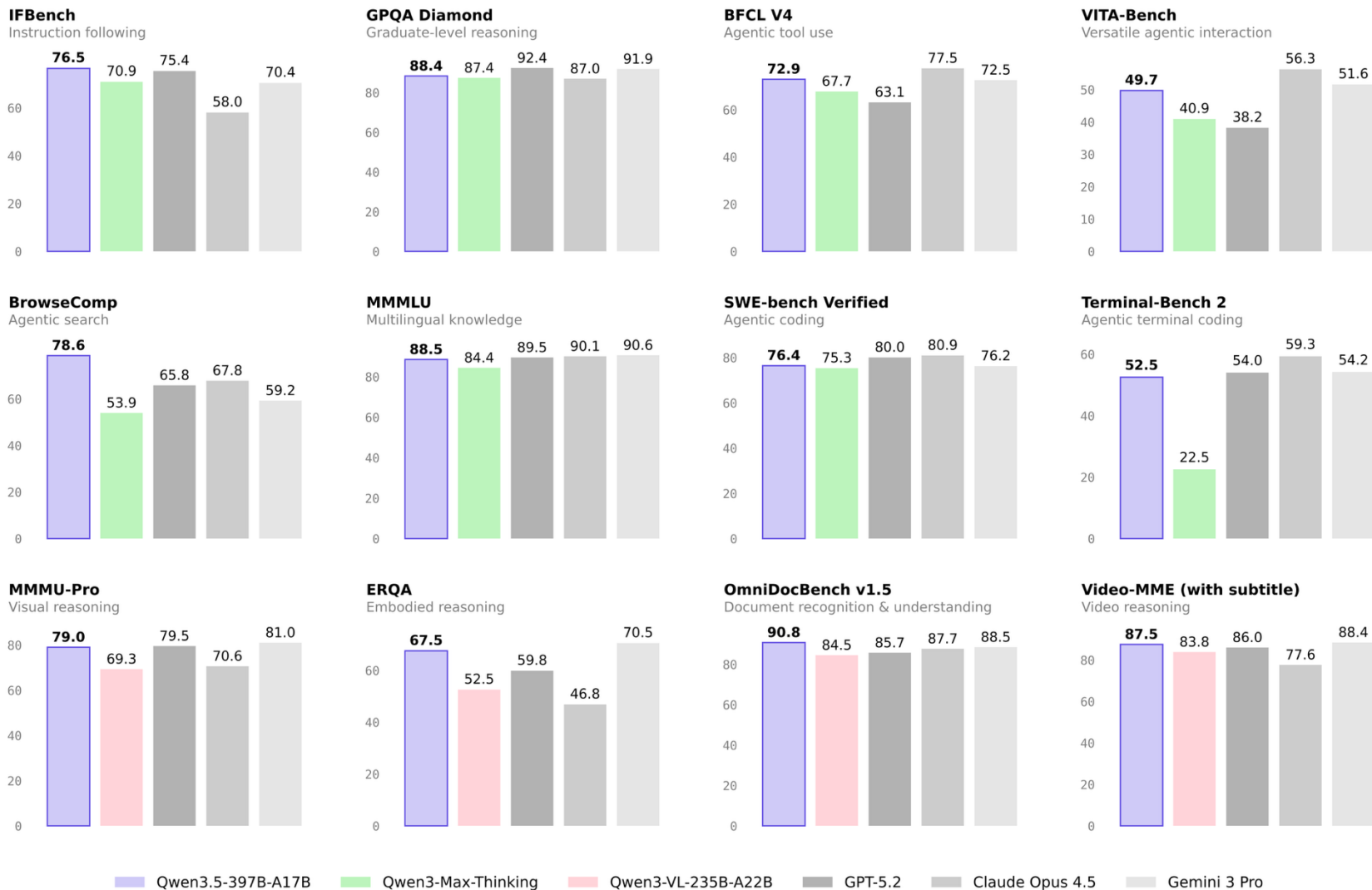
Closed-source (proprietary)

- Gemini 3.1 Pro (Google)
- GPT 5.4 (OpenAI)
- Claude 4.6 (Anthropic)
- Grok 4.1 (xAI)
- Seed 2.0 (ByteDance)
- ...

Open-weight

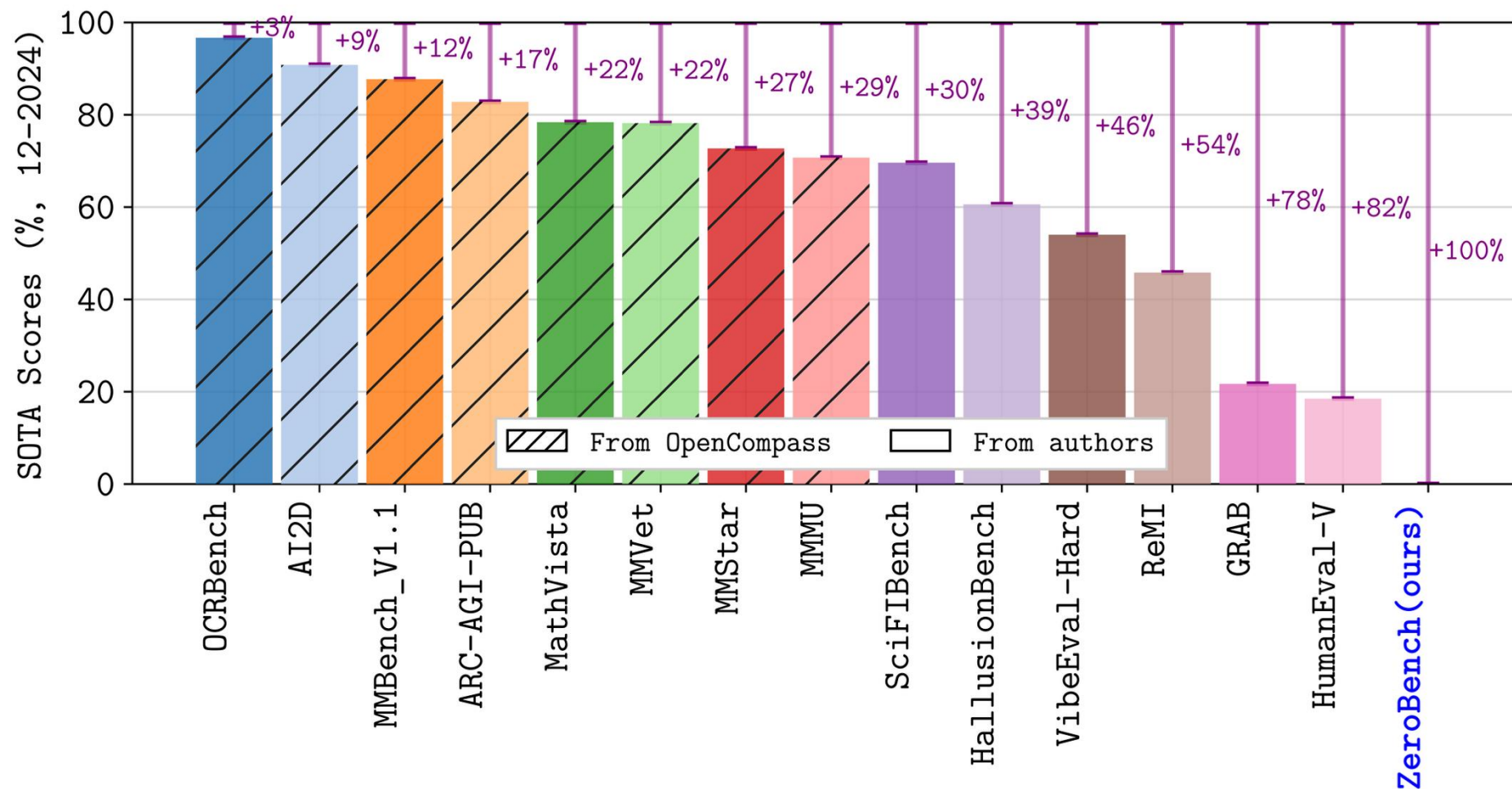
- Qwen 3.5 (Alibaba)
- Kimi K2.5 (Moonshot AI)
- DeepSeek V3.2 (DeepSeek)
- GLM 5 (Zhipu)
- Gemma 3 (Google)
- Mistral 3 (Mistral AI)
- Llama 4 (Meta)
- GPT-OSS-120B (OpenAI)
- ...

Evaluation Benchmarks



ZeroBench

An Impossible* Visual Benchmark for Contemporary Large Multimodal Models



ZeroBench

An Impossible* Visual Benchmark for Contemporary Large Multimodal Models

ZeroBench					
Models	Main questions (100)			Subquestions (334)	
	k/k [%] (n)	pass@k [%] (n)		pass@k [%] (SE_{CLT})	Num. correct
	k=5	k=1	k=5	k=1	k=1
Reasoning LMMs					
o1 pro ^{o,s}	0.0 (0)	0.0 (0)	-	22.40 (2.48)	75
o1 ^o	0.0 (0)	0.0 (0)	0.0 (0)	20.21 (2.33)	68
Gemini 2 Flash Thinking	0.0 (0)	0.0 (0)	5.0 (5)	20.51 (2.60)	69
QVQ	0.0 (0)	0.0 (0)	4.0 (4)	20.47 (2.38)	70
Proprietary LMMs					
GPT-4o	0.0 (0)	0.0 (0)	0.0 (0)	19.60 (2.37)	67
GPT-4o mini	0.0 (0)	0.0 (0)	2.0 (2)	16.58 (2.41)	54
Gemini 2 Flash	0.0 (0)	0.0 (0)	3.0 (3)	23.24 (2.85)	75
Gemini 1.5 Pro	0.0 (0)	0.0 (0)	2.0 (2)	20.88 (2.47)	74
Gemini 1.5 Flash	0.0 (0)	0.0 (0)	2.0 (2)	17.87 (2.41)	63
Gemini 1 Pro Vision	0.0 (0)	0.0 (0)	2.0 (2)	12.36 (2.08)	46
Claude 3.5 Sonnet v2	0.0 (0)	0.0 (0)	2.0 (2)	25.50 (2.67)	82
Claude 3.5 Sonnet	0.0 (0)	0.0 (0)	1.0 (1)	20.71 (2.48)	72
Claude 3 Opus	0.0 (0)	0.0 (0)	0.0 (0)	15.10 (2.16)	45
Claude 3 Sonnet	0.0 (0)	0.0 (0)	1.0 (1)	16.08 (2.25)	49
Claude 3 Haiku	0.0 (0)	0.0 (0)	0.0 (0)	12.27 (2.05)	41
Reka Edge	0.0 (0)	0.0 (0)	0.0 (0)	3.74 (0.96)	13
Open-weight LMMs					
Llama 3.2 90B	0.0 (0)	0.0 (0)	0.0 (0)	13.26 (1.92)	48
Qwen2-VL-72B-Instruct	0.0 (0)	0.0 (0)	2.0 (2)	13.00 (2.32)	43
NVLM-D-72B	0.0 (0)	0.0 (0)	1.0 (1)	14.91 (2.36)	51
Pixtral-Large	0.0 (0)	0.0 (0)	3.0 (3)	18.68 (2.26)	62

ZeroBench

An Impossible* Visual Benchmark for Contemporary Large Multimodal Models

Model	Main questions (100)			Subquestions (334)	
	pass@1	pass@5	pass^5	pass@1	n correct
GPT-5.4 (xhigh) 🔥	-	23.0	8.0	-	-
Gemini 3.1 Pro 🔥	-	19.0	7.0	-	-
Gemini 3 Pro 🔥	-	19.0	5.0	-	-
GPT-5.2 (medium reasoning) 🔥	-	17.0	6.0	-	-
Gemini 3 Flash 🔥	-	13.0	2.0	-	-
Claude Opus 4.5 🔥	-	10.0	1.0	-	-
o4-mini 🔥	2.0	10.0	0.0	29.05	89
GPT-5-mini (high reasoning) 🔥	4.0	9.0	3.0	27.79	95
Claude Opus 4.1 (thinking) 🔥	5.0	8.0	1.0	24.38	80
GPT-5 (medium reasoning) 🔥	1.0	7.0	0.0	26.19	93
GPT 4.5 🔥	1.0	7.0	0.0	27.01	95
Claude Sonnet 4 (thinking) 🔥	3.0	6.0	0.0	24.60	80
o3 🔥	3.0	6.0	0.0	25.51	85
GPT-5.1 (medium reasoning) 🔥	-	5.0	0.0	-	-
Claude Opus 4 (thinking) 🔥	4.0	5.0	1.0	25.06	81
Claude Sonnet 4 🔥	2.0	5.0	1.0	23.01	77
Gemini 2.5 Pro 🔥	3.0	5.0	1.0	26.04	90
Claude Opus 4.1 🔥	1.0	4.0	1.0	25.30	81
Grok 4 🔥	1.0	4.0	0.0	21.58	71

ZeroBench

An Impossible* Visual Benchmark for Contemporary Large Multimodal Models

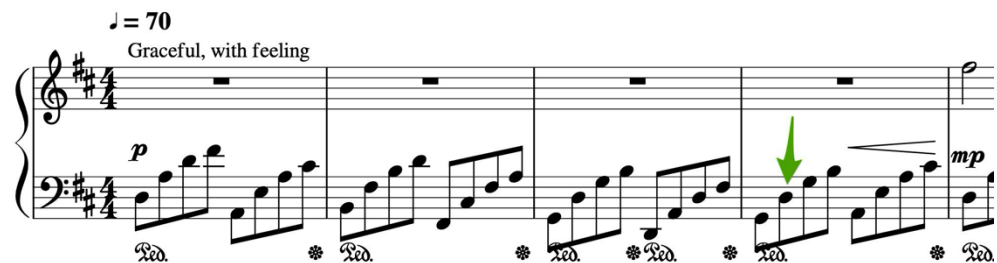


Question:

In the image there is a stationary pot (containing different pens and a letter opener) and other things. Consider: (1) The typical number of legs of the species visible in the picture (2) The total number of pens (3) The total number of pen nibs exposed and visible. What is the value of the product of (1), (2) and (3)?

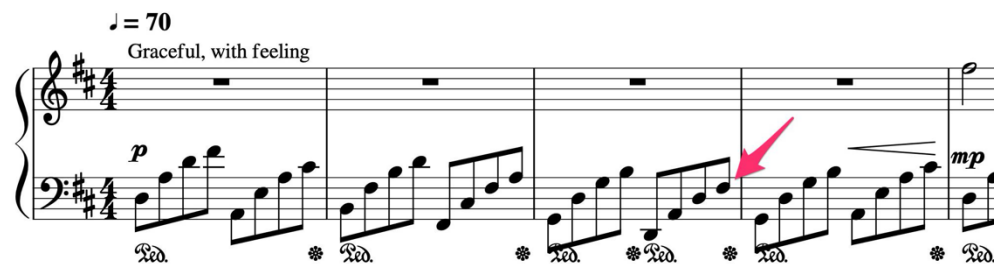
ZeroBench

An Impossible* Visual Benchmark for Contemporary Large Multimodal Models



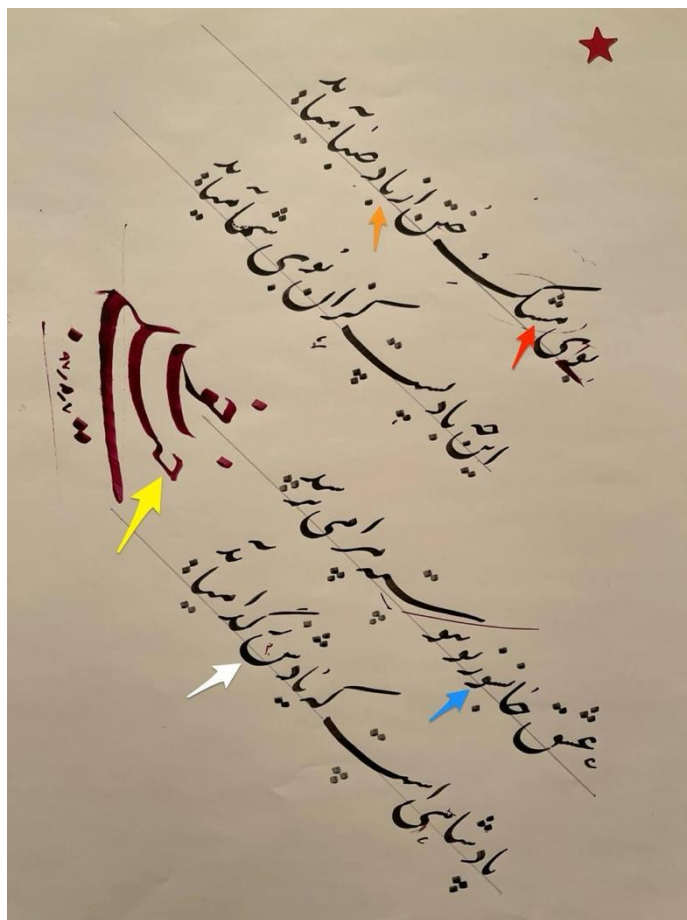
Question:

Read the note pointed at by the pink arrow. Then, read the note pointed at by the green arrow. Return the two notes in that order without a space or comma.



ZeroBench


An Impossible* Visual Benchmark for Contemporary Large Multimodal Models











Question:


The image shows calligraphic writing in Farsi, with some letters annotated with arrows. If we put these letters together, what would the message be when translated into English? Use the colors in the following order: red, orange, yellow, blue, and white.





Qwen 3.5 Small Models (3 March 2026)





 Qwen/Qwen3.5-0.8B NEW


 Image-Text-to-Text  Safetensors  PyTorch 





 Qwen |  2026.03.03 |  15.2k |  10





 Qwen/Qwen3.5-2B NEW


 Image-Text-to-Text  Safetensors  PyTorch 





 Qwen |  2026.03.03 |  4.2k |  3





 Qwen/Qwen3.5-4B NEW

 Image-Text-to-Text  Safetensors  PyTorch 

 Qwen |  2026.03.03 |  19.2k |  11

 Qwen/Qwen3.5-9B NEW

 Image-Text-to-Text  Safetensors  PyTorch 

 Qwen |  2026.03.03 |  19.2k |  15

Qwen3.5-0.8B WebGPU demo



r/LocalLLaMA · 2d ago
HighFlyingB1rd

Running Qwen3.5-0.8B on my 7-year-old Samsung S10E

Discussion

```
llama.cpp

build      : b8188-36a7a6589
model     : Qwen3.5-0.8B-Q4_K_M.gguf
modalities : text

available commands:
/exit or Ctrl+C  stop or exit
/regen          regenerate the last response
/clear         clear the chat history
/read         add a text file

> Explain how a car engine works in 3 sentences.

A car engine operates by converting chemical energy from burning fuel into mechanical energy, which is then transformed into rotational motion to drive wheels and move the vehicle. The combustion of fuel creates high-pressure gas that expands, pushing pistons back and forth across engine cylinders to generate power. Finally, this mechanical energy is transferred to the car's drivetrain to propel the vehicle forward, utilizing friction and torque from the wheels.

[ Prompt: 51.6 t/s | Generation: 12.2 t/s ]

>
```

Qwen just released their 0.8B model.
So naturally, I had to try running it on my 7-year-old Samsung S10E.

https://www.reddit.com/r/LocalLLaMA/comments/1r1zody/running_qwen_35_08b_locally_in_the_browser_on/

Qwen 3.5 Small Models

Fine-tune Qwen3.5

Train Qwen3.5 locally on just 5GB VRAM.

You can now fine-tune [Qwen3.5](#) model family (0.8B, 2B, 4B, 9B, 27B, 35B-A3B, 122B-A10B) with [Unsloth](#). Support includes [vision](#) and text. **Qwen3.5-35B-A3B** - bf16 LoRA works on **74GB VRAM**.

- Unsloth makes Qwen3.5 train **1.5x faster** and uses **50% less VRAM** than FA2 setups.
- Qwen3.5 BF16 LoRA VRAM requirements: **0.8B**: 3GB • **2B**: 5GB • **4B**: 10GB • **27B**: 56GB
- Fine-tune **0.8B, 2B, 4B** and **9B** via our **free Google Colab notebooks**:

[Qwen3.5-0.8B](#)

[Qwen3.5-2B](#)

[Qwen3.5-4B](#)



- If you want to **preserve reasoning** ability, you can mix reasoning-style examples with direct answers (keep a minimum of 75% reasoning). Otherwise you can emit it fully.
- After fine-tuning, you can export to [GGUF](#) (for llama.cpp/Ollama/LM Studio/etc.) or [vLLM](#)
- [Reinforcement Learning](#) (RL) for Qwen3.5 [VLM RL](#) also works via Unsloth inference.
- We have **A100** Colab notebooks for [Qwen3.5-27B](#) and [Qwen3.5-35B-A3B](#).

If you're on an older version (or fine-tuning locally), update first:

```
pip install --upgrade --force-reinstall --no-cache-dir unsloth unsloth_zoo
```


⚠ If training seems **slower than usual**, it's because Qwen3.5 use custom Mamba Triton kernels. Compiling those kernels can take longer than normal, especially on T4 GPUs.





Below is a minimal SFT recipe (works for "text-only" fine-tuning). See also [vision fine-tuning](#) section.





```
from unsloth import FastLanguageModel
import torch
from datasets import load_dataset
from trl import SFTTrainer, SFTConfig


max_seq_length = 2048 # start small; scale up after it works





# Example dataset (replace with yours). Needs a "text" column.
url = "https://huggingface.co/datasets/laion/OIG/resolve/main/unified_chip2.jsonl"
dataset = load_dataset("json", data_files={"train": url}, split="train")
```





 **Qwen/Qwen3.5-0.8B** NEW


 Image-Text-to-Text  Safetensors  PyTorch 





 Qwen |  2026.03.03 |  15.2k |  10





 **Qwen/Qwen3.5-2B** NEW


 Image-Text-to-Text  Safetensors  PyTorch 





 Qwen |  2026.03.03 |  4.2k |  3





 **Qwen/Qwen3.5-4B** NEW

 Image-Text-to-Text  Safetensors  PyTorch 

 Qwen |  2026.03.03 |  19.2k |  11

 **Qwen/Qwen3.5-9B** NEW

 Image-Text-to-Text  Safetensors  PyTorch 

 Qwen |  2026.03.03 |  19.2k |  15

In-context Learning

Language Models are Few-Shot Learners

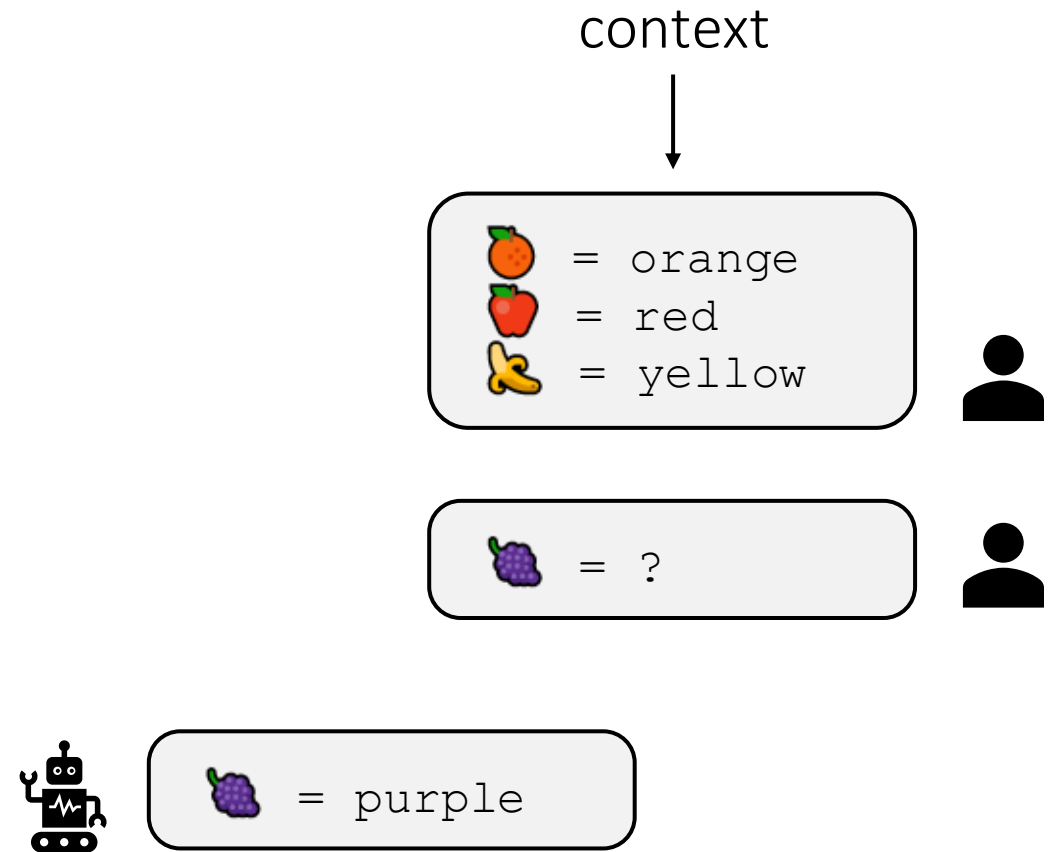
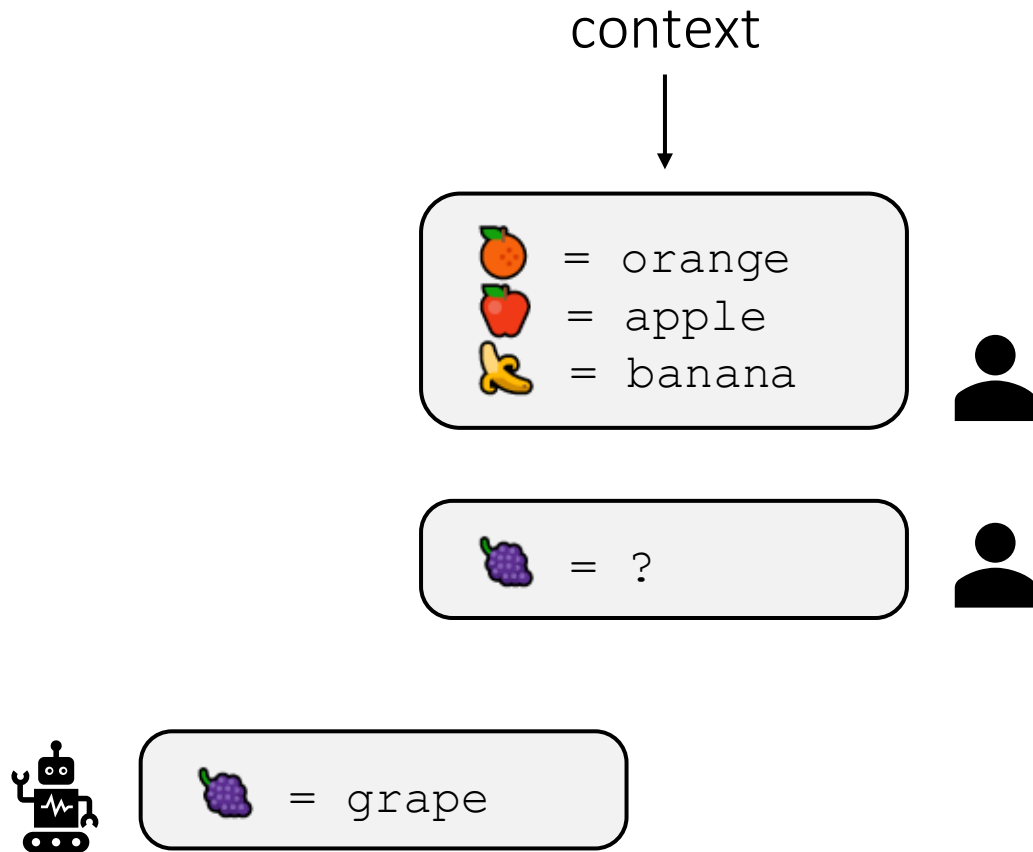
OpenAI

GPT-3 Paper

Abstract

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

In-context Learning



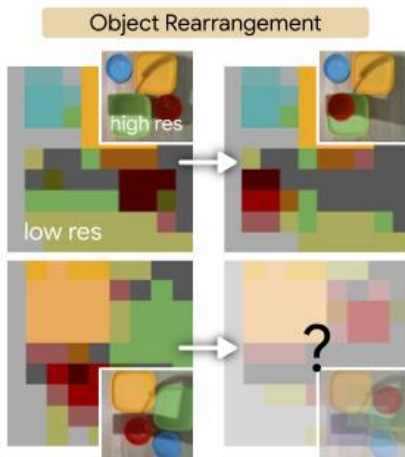
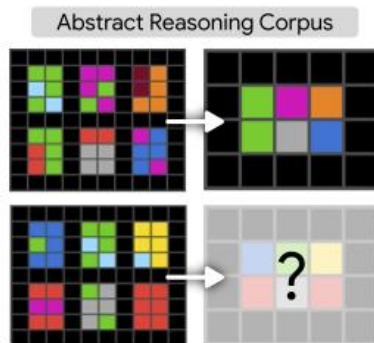
Large Language Models as General Pattern Machines

Suvir Mirchandani Fei Xia Pete Florence Brian Ichter Danny Driess Montserrat Gonzalez Arenas
Kanishka Rao Dorsa Sadigh Andy Zeng



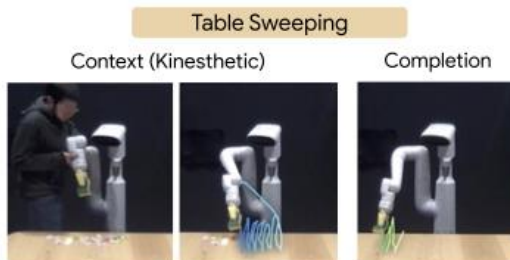
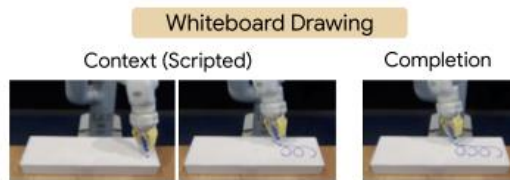
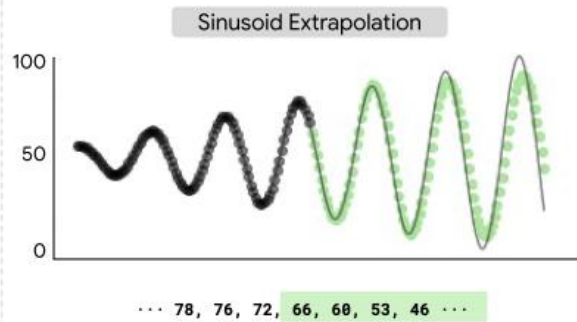
Sequence Transformation

Pattern transformations (symbolic)



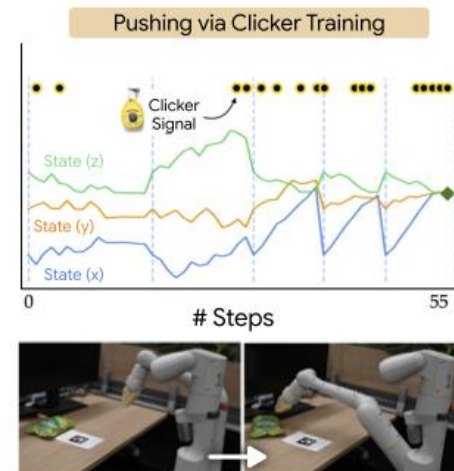
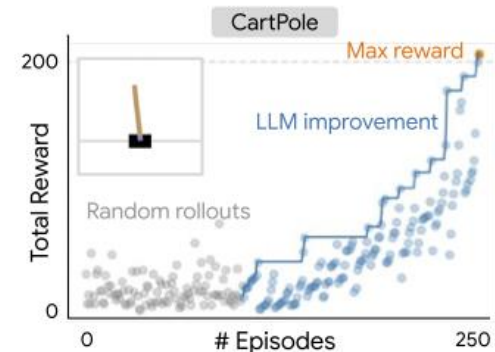
Sequence Completion

Simple function classes (numeric)



Sequence Improvement

Online policies (numeric & symbolic)



An Explanation of In-context Learning as Implicit Bayesian Inference

Sang Michael Xie
Stanford University
xie@cs.stanford.edu

Aditi Raghunathan
Stanford University
aditir@stanford.edu

Percy Liang
Stanford University
pliang@cs.stanford.edu

Tengyu Ma
Stanford University
tengyuma@cs.stanford.edu

Abstract

Large language models (LMs) such as GPT-3 have the surprising ability to do in-context learning, where the model learns to do a downstream task simply by conditioning on a prompt consisting of input-output examples. The LM learns from these examples *without being explicitly pretrained to learn*. Thus, it is unclear what enables in-context learning. In this paper, we study how in-context learning can emerge when pretraining documents have long-range coherence. Here, the LM must infer a latent document-level concept to generate coherent next tokens during pretraining. At test time, in-context learning occurs when the LM also infers a shared latent concept between examples in a prompt. We prove when this occurs despite a distribution mismatch between prompts and pretraining data in a setting where the pretraining distribution is a mixture of HMMs. In contrast to messy large-scale datasets used to train LMs capable of in-context learning, we generate a small-scale synthetic dataset (GINC) where Transformers and LSTMs both exhibit in-context learning¹. Beyond the theory, experiments on GINC exhibit large-scale real-world phenomena including improved in-context performance with model scaling (despite the same pretraining loss), sensitivity to example order, and instances where zero-shot is better than few-shot in-context learning.

Agentic Context Engineering: Evolving Contexts for Self-Improving Language Models

Qizheng Zhang^{1*} Changran Hu^{2*} Shubhangi Upasani² Boyuan Ma² Fenglu Hong²
Vamsidhar Kamanuru² Jay Rainton² Chen Wu² Mengmeng Ji² Hanchen Li³
Urmish Thakker² James Zou¹ Kunle Olukotun¹

¹ Stanford University ² SambaNova Systems, Inc. ³ UC Berkeley * equal contribution

✉ qizhengz@cs.stanford.edu, changran_hu@berkeley.edu, kunle@cs.stanford.edu

🌐 <https://github.com/ace-agent/ace> 🌐 <https://ace-agent.github.io>

Abstract

Large language model (LLM) applications such as agents and domain-specific reasoning increasingly rely on *context adaptation*—modifying inputs with instructions, strategies, or evidence, rather than weight updates. Prior approaches improve usability but often suffer from brevity bias, which drops domain insights for concise summaries, and from context collapse, where iterative rewriting erodes details over time. Building on the adaptive memory introduced by Dynamic Cheatsheet, we introduce ACE (Agentic Context Engineering), a framework that treats contexts as evolving playbooks that accumulate, refine, and organize strategies through a modular process of generation, reflection, and curation. ACE prevents collapse with structured, incremental updates that preserve detailed knowledge and scale with long-context models. Across agent and domain-specific benchmarks, ACE optimizes contexts both offline (*e.g.*, system prompts) and online (*e.g.*, agent memory), consistently outperforming strong baselines: +10.6% on agents and +8.6% on finance, while significantly reducing adaptation latency and rollout cost. Notably, ACE could adapt effectively without labeled supervision and instead by leveraging natural execution feedback. On the AppWorld leaderboard, ACE matches the top-ranked production-level agent on the overall average and surpasses it on the harder test-challenge split, despite using a smaller open-source model. These results show that comprehensive, evolving contexts enable scalable, efficient, and self-improving LLM systems with low overhead.

Context engineering as opposed to fine-tuning in the age of large models?

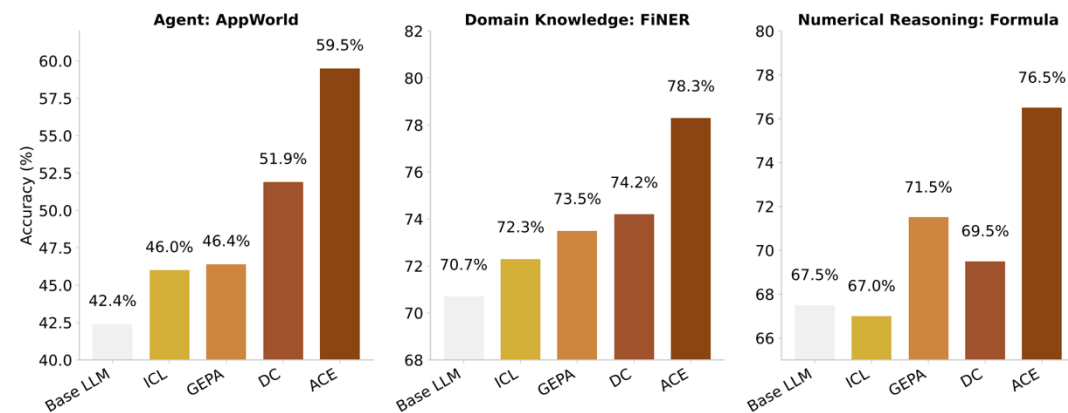
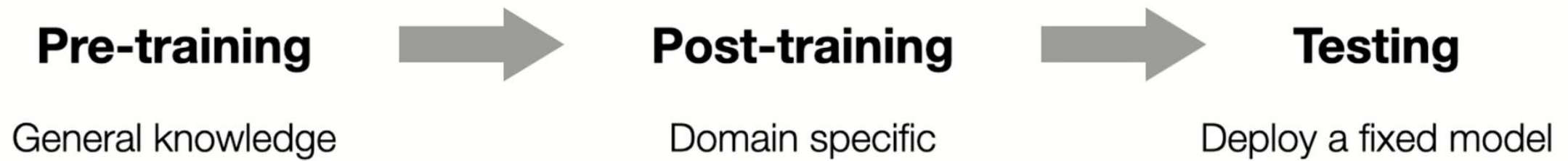


Figure 1: **Overall Performance Results.** Our proposed framework, ACE, consistently outperforms strong baselines across agent and domain-specific reasoning tasks.

ML in 2019



ML today



ML today

The test instance arrives

ML in the future

Pre-training



Post-training



Test-time training

Sun et al. 2019

Training **before** you see the test instance

Training **after** you see the test instance

Visual Generation



DALL·E 2. OpenAI. 2022.

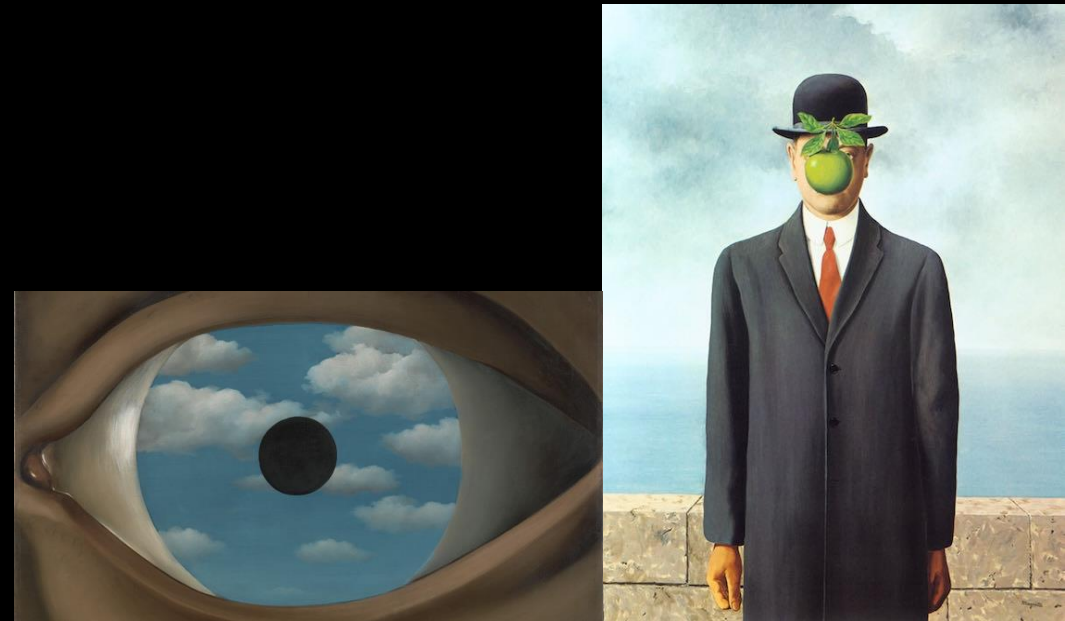
Demonstrated the power of diffusion models



“An IT-guy trying to fix hardware of a PC tower is being tangled by the PC cables like Laokoon. Marble, copy after Hellenistic original from ca. 200 BC. Found in the Baths of Trajan, 1506.”



“About 99% of the time, the right time is right now in René Magritte style.”



“The False Mirror” (1928)

“The Son of Man” (1964)

2023

2024

2025

2026

Video Generation (2023-2026)



Infinite Interactive Worlds

Project Genie. Google DeepMind. 2026.



Sora. OpenAI. 2024.
Demonstrated the power of Diffusion Transformer (DiT)



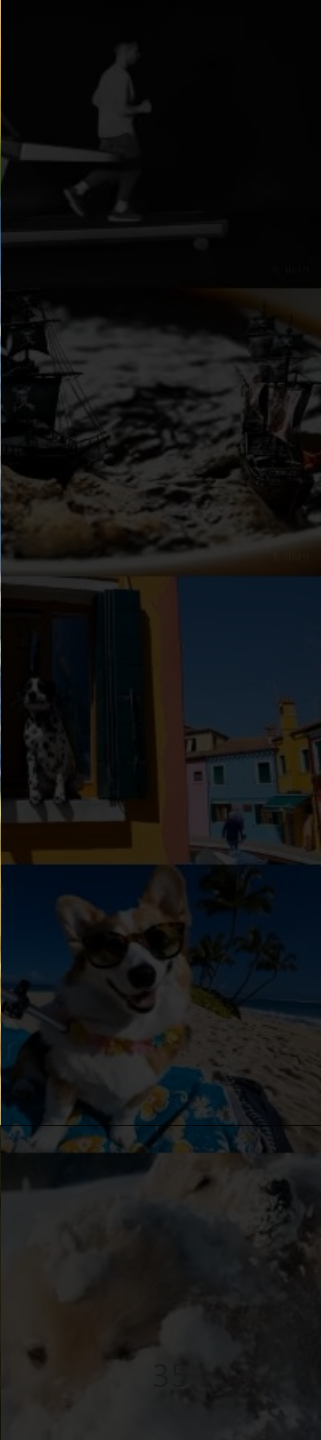
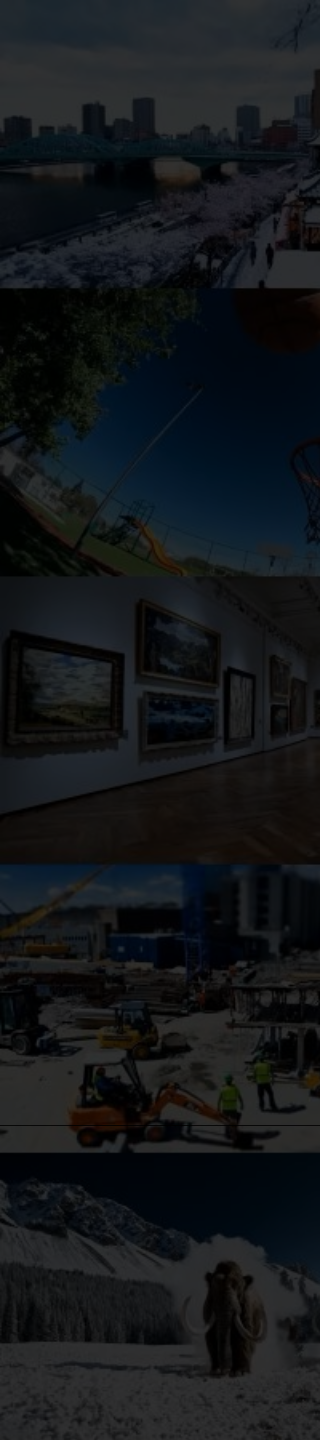
Sora: Video Generation Models as World Simulators. OpenAI. 2024.



Pixels Defying Physics

Sora: Video Generation Models as World Simulators. OpenAI. 2024.





Pixels Defying Physics

Sora: Video Generation Models as World Simulators. OpenAI. 2024.



Reality Check

First frame from Physics IQ Benchmark, video generated by Veo 3.1 from Google DeepMind with the prompt: “what will happen if the spinning wheel keeps turning and the stick hits the first block on the right.” Adapted from VDAWorld (O’Mahony et al. 2025).



WorldScore

A Unified Evaluation Benchmark for World Generation

Models	WorldScore-Static	WorldScore-Dynamic	Camera Ctrl	Object Ctrl	Content Align	3D Consist	Photo Consist	Style Consist	Subjective Qual	Motion Acc	Motion Mag	Motion Smooth
Gen-3	60.71	<u>57.58</u>	29.47	62.92	50.49	68.31	87.09	62.82	63.85	54.53	27.48	68.87
Hailuo	57.55	56.36	22.39	69.56	<u>73.53</u>	67.18	62.82	54.91	52.44	63.46	27.20	70.07
DynamiCrafter	52.09	47.19	25.15	47.36	25.00	72.90	60.95	78.85	54.40	41.11	39.25	26.92
VideoCrafter1-T2V	47.10	43.54	21.61	50.44	60.78	64.86	51.36	38.05	42.63	11.76	75.00	18.87
VideoCrafter1-I2V	50.47	47.64	25.46	24.25	35.27	74.42	73.89	65.17	54.85	55.63	25.00	42.49
VideoCrafter2	52.57	47.49	28.92	39.07	72.46	65.14	61.85	43.79	56.74	47.12	30.40	29.39
T2V-Turbo	45.65	40.20	27.80	30.68	69.14	38.72	34.84	49.65	<u>68.74</u>	34.87	40.09	7.48
EasyAnimate	52.85	51.65	26.72	54.50	50.76	67.29	47.35	73.05	50.31	<u>75.00</u>	31.16	40.32
CogVideoX-T2V	54.18	48.79	40.22	51.05	68.12	68.81	64.20	42.19	44.67	25.00	<u>47.31</u>	36.28
CogVideoX-I2V	62.15	59.12	38.27	40.07	36.73	86.21	88.12	<u>83.22</u>	62.44	69.56	26.42	60.15
Allegro	55.31	51.97	24.84	57.47	51.48	70.50	69.89	65.60	47.41	54.39	40.28	37.81
Vchitect-2.0	42.28	38.47	26.55	49.54	65.75	41.53	42.30	25.69	44.58	33.59	33.81	21.31
SceneScape	50.73	35.51	84.99	47.44	28.64	76.54	62.88	21.85	32.75	0.00	0.00	0.00
Text2Room	62.10	43.47	94.01	38.93	50.79	<u>88.71</u>	88.36	37.23	36.69	0.00	0.00	0.00
LucidDreamer	70.40	49.28	88.93	41.18	75.00	90.37	90.20	48.10	58.99	0.00	0.00	0.00
WonderJourney	63.75	44.63	84.60	37.10	35.54	80.60	79.03	62.82	66.56	0.00	0.00	0.00
InvisibleStitch	61.12	42.78	<u>93.20</u>	36.51	29.53	88.51	<u>89.19</u>	32.37	58.50	0.00	0.00	0.00
WonderWorld	<u>72.69</u>	50.88	92.98	51.76	71.25	86.87	85.56	70.57	49.81	0.00	0.00	0.00
4D-fy	27.98	32.10	69.92	55.09	0.85	35.47	1.59	32.04	0.89	22.22	22.88	80.06
LTX-Video	55.44	56.54	25.06	53.41	39.73	78.41	88.92	53.50	49.08	76.22	29.95	<u>71.09</u>
Voyager	77.62	54.53	85.95	<u>66.92</u>	68.92	81.56	85.99	84.89	71.09	0.00	0.00	0.00
Wan2.1	57.56	52.85	23.53	40.32	45.44	78.74	78.36	77.18	59.38	54.27	33.26	38.05

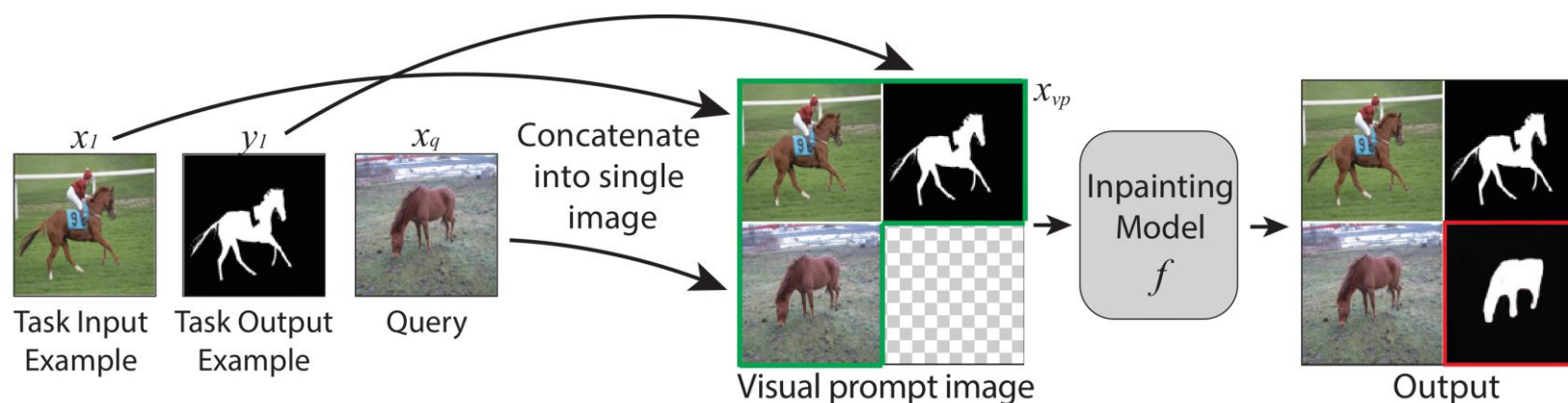
Visual Understanding via Conditional Generation

Visual Prompting via Image Inpainting

Amir Bar*^{1,2}, Yossi Gandelsman*¹, Trevor Darrell¹, Amir Globerson², Alexei A. Efros¹

UC Berkeley¹

Tel Aviv University²



Visual Understanding via Conditional Generation

Video models are zero-shot learners and reasoners

Thaddäus Wiedemer*, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini*, Robert Geirhos*

Google DeepMind

* Joint leads

Perception



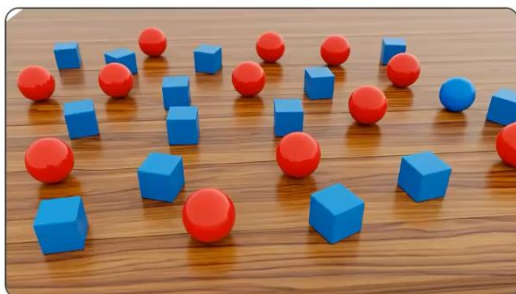
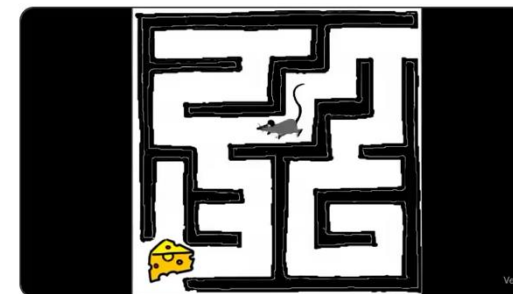
Modeling



Manipulation

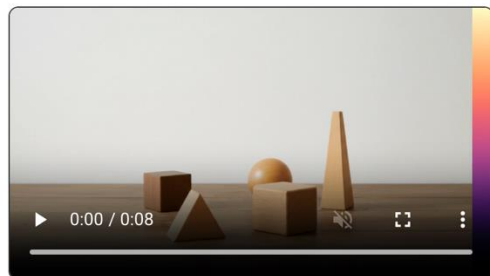


Reasoning



Visual Understanding via Conditional Generation

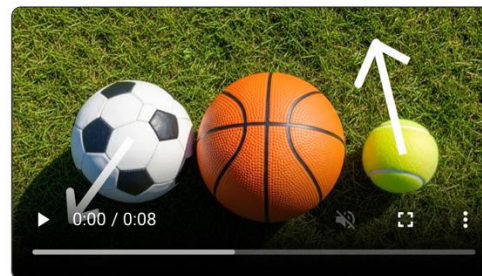
Failure cases (click to expand) ▼



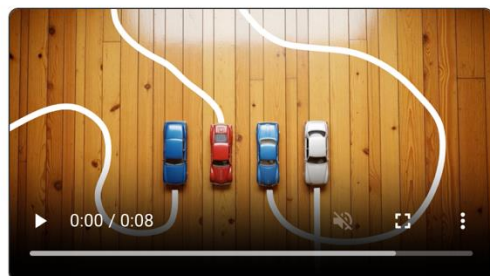
Monocular depth estimation



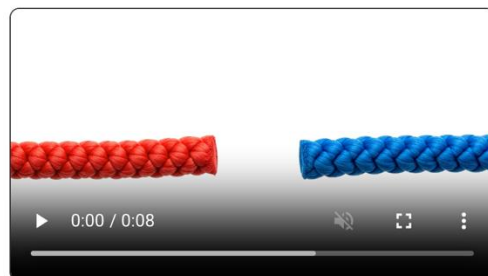
Monocular surface normal estimation



Force prompting



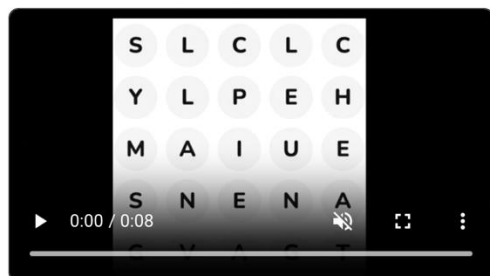
Motion trajectory prompting



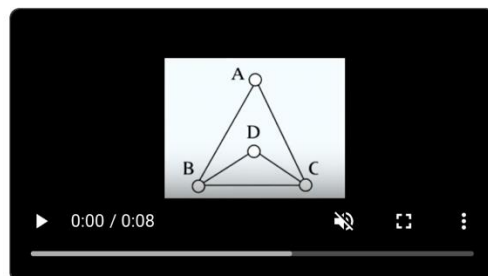
Tying the knot



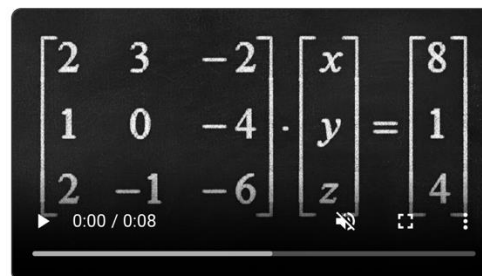
Connect the path puzzle



Letter word search



Eulerian path



Solving linear equations

Visual Understanding via Conditional Generation



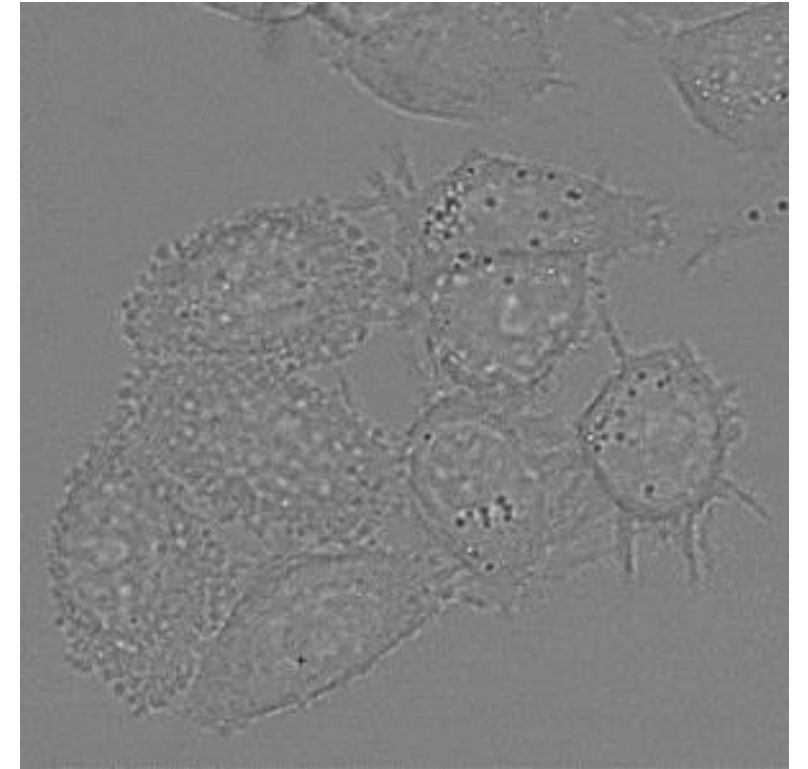
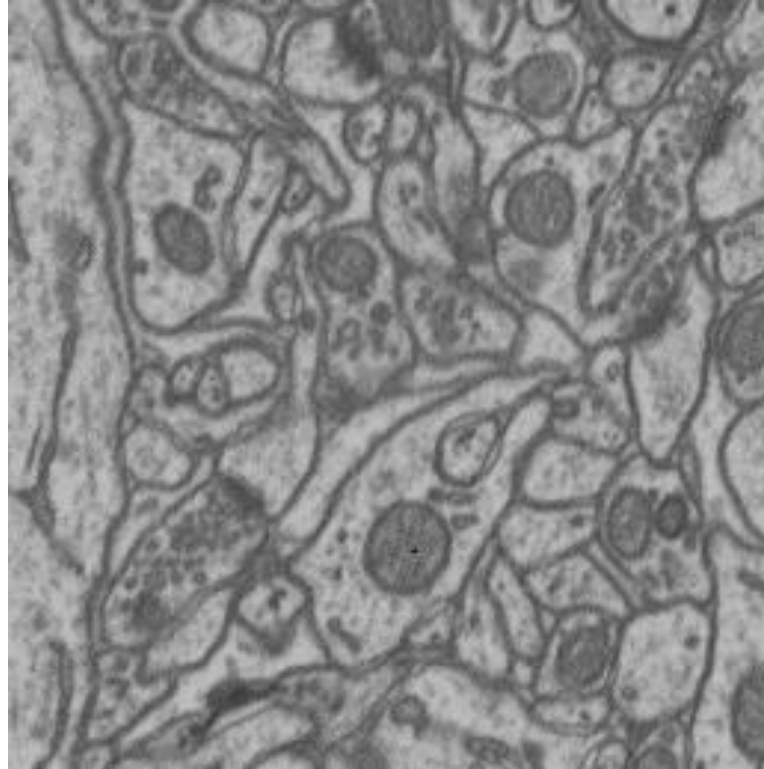
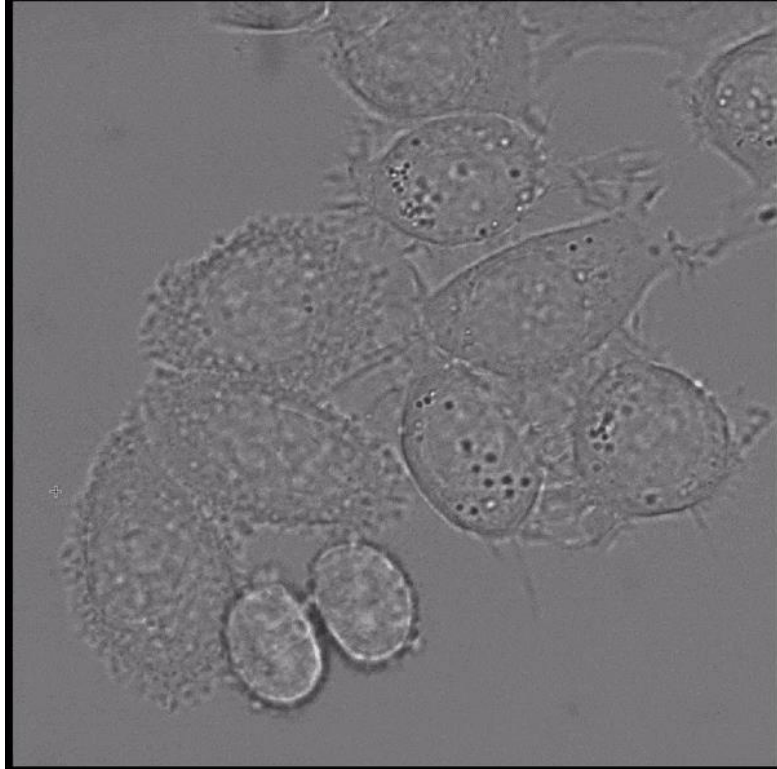
automatic 3D data labelling pipeline using Gemini (Nano Banana 2)

A photograph of three penguins on a sandy beach. The penguins are black and white with orange beaks. In the background, there are waves breaking on the shore. A semi-transparent search bar is overlaid on the image, containing a magnifying glass icon on the left and an 'X' icon on the right. The search bar is currently empty.

Introducing Meta Segment Anything Model 3 (SAM 3)

Segment Anything Model 3 (SAM 3)

Segment Anything for Microscopy



<https://github.com/computational-cell-analytics/micro-sam>

SAM 3D

GPT-4



SORA



**What remarkable instruments!
Time to learn to **conduct** them!**

Genie



**Everyone else scaling
to 10T parameters...**

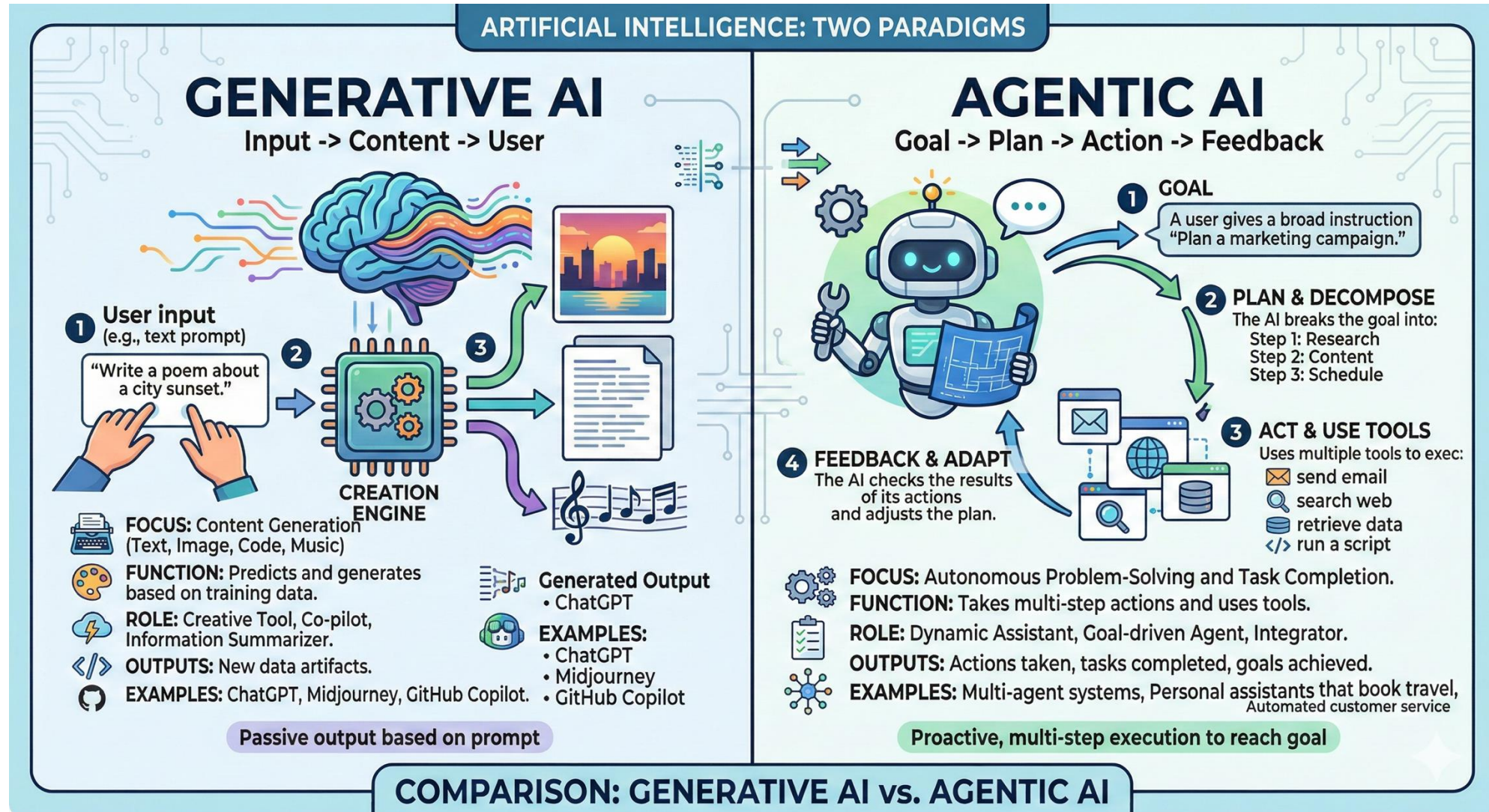


Me with Megalo-Modelo-Phobia.



Part 2 – Expanding Horizons

Agentic AI Systems (2024-now)



ChatGPT Atlas

The screenshot shows a web browser window with several tabs open. The active tab is displaying a recipe on allrecipes.com for 'Jack-o'-Lantern Stuffed Peppers'. The recipe text describes the dish and provides instructions. Below the text is a video player showing a close-up of the stuffed peppers, with a 'Getting started...' overlay. To the right of the video is an advertisement for Zepbound. At the bottom of the page, there is a chat window with the text 'Starting to gather options for ordering meat and produce'.



Codex

Introducing the Codex app. OpenAI. 2026. <https://www.youtube.com/watch?v=HFM3se4INiw>

GPT5.4 Codex Demo Prompt (from Samuel Albanie)

Build me a cinematic, self-playing animated visualization of how the 2025/26 Premier League table has evolved across the season so far.

This will be screen-recorded for a Twitter/X video demo, so optimize for that context:

****Format & Motion:****

- The animation should auto-play from matchday 1 to the present, no user interaction required
- Target 20-40 seconds of total runtime — long enough to absorb, short enough to loop
- Aspect ratio should be 16:9 (landscape, 1920x1080 ideally)
- The first 2 seconds must be visually arresting — this needs to stop a fast-scrolling Twitter feed
- Smooth, fluid motion throughout. Ease-in/ease-out on transitions. No jank, no jumpiness
- Include a subtle matchday or gameweek counter so the viewer tracks progression

****Visual & Artistic Direction:****

- Each club must use its actual brand/crest colors
- Prioritize artistic ambition over convention. No generic line charts, no plain bar chart races. Think more like a data artist than a dashboard engineer — take inspiration from the aesthetics of Giorgia Lupi's "data humanism," the flowing visual energy of a Beeple piece, or a Pentagram studio commission
- The palette, typography, spacing, and motion should feel cohesive and intentional — like a title sequence, not a BI report
- Subtle particle effects, glow, grain, bloom, or texture are encouraged if they serve the mood
- The animation should build tension — I want to *feel* title races tightening and relegation battles intensifying, not just see lines moving

****Data:****

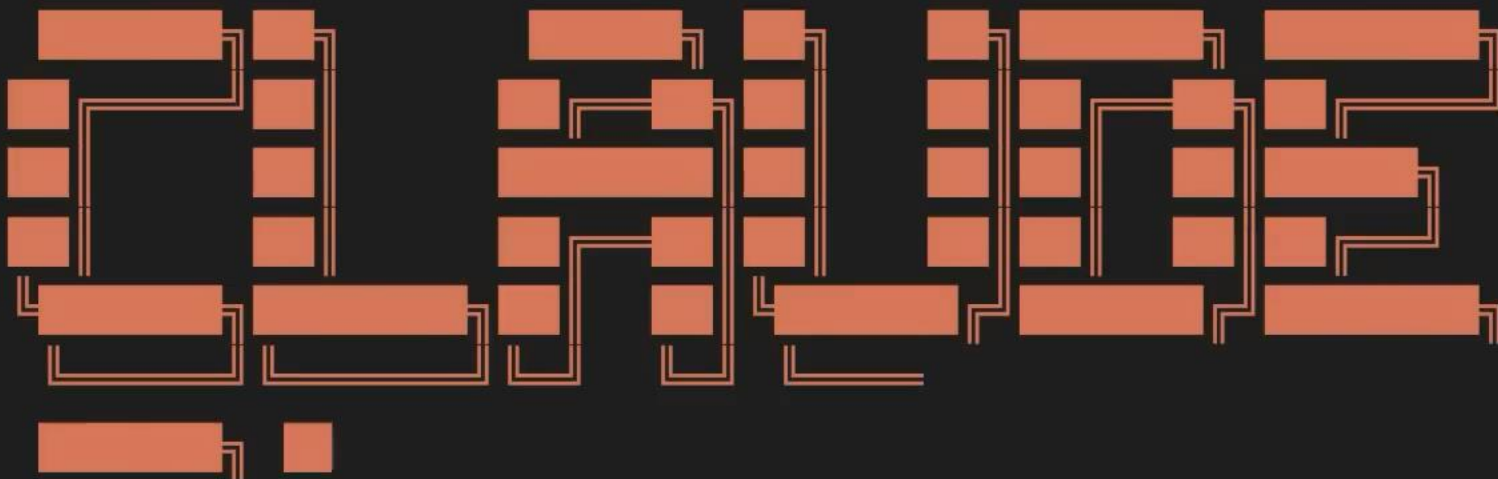
- Use web search to pull in the actual current standings and week-by-week results. The data must be real
- DO NOT use placeholder or dummy data under any circumstances. If you cannot fetch real data, stop and tell me

****Technical:****

- Single self-contained HTML file that runs in a browser
- Use canvas, WebGL, Three.js, or D3 — whichever you believe ...



* Welcome to the **Claude Code** research preview!



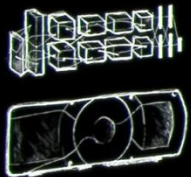
Claude Code

Cursor



What's Next?

2012 ALEXNET



PERCEPTION AI

SPEECH RECOGNITION
DEEP RECSYS
MEDICAL IMAGING

GENERATIVE AI

DIGITAL MARKETING
CONTENT CREATION

AGENTIC AI

CODING ASSISTANT
CUSTOMER SERVICE
PATIENT CARE

PHYSICAL AI

SELF-DRIVING CARS
GENERAL ROBOTICS



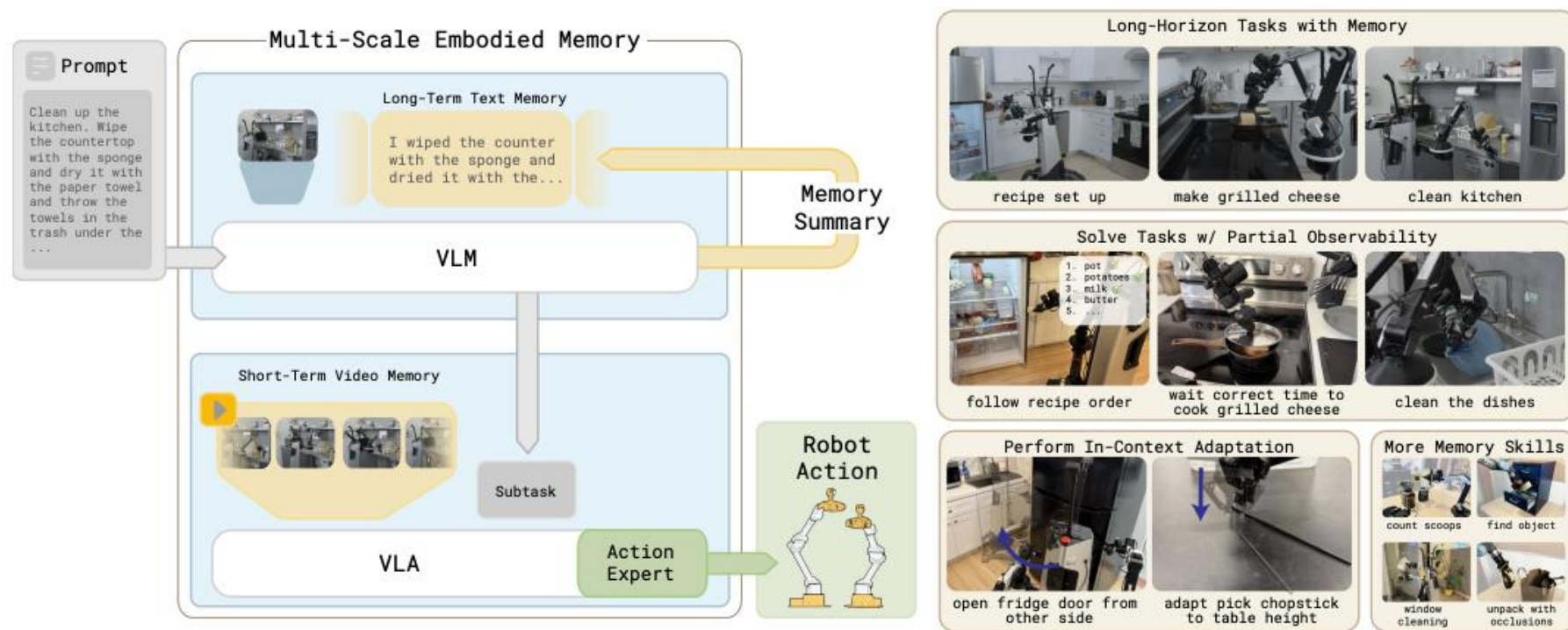
Today's Roadblocks to Physical AI

- Data collection cost
- Physical accuracy (hallucination)
- Embodied memory for long-horizon tasks
- Hardware dexterity
- Latency in real-time deployment
- Safety guardrails
- Ethics and governance
- ...

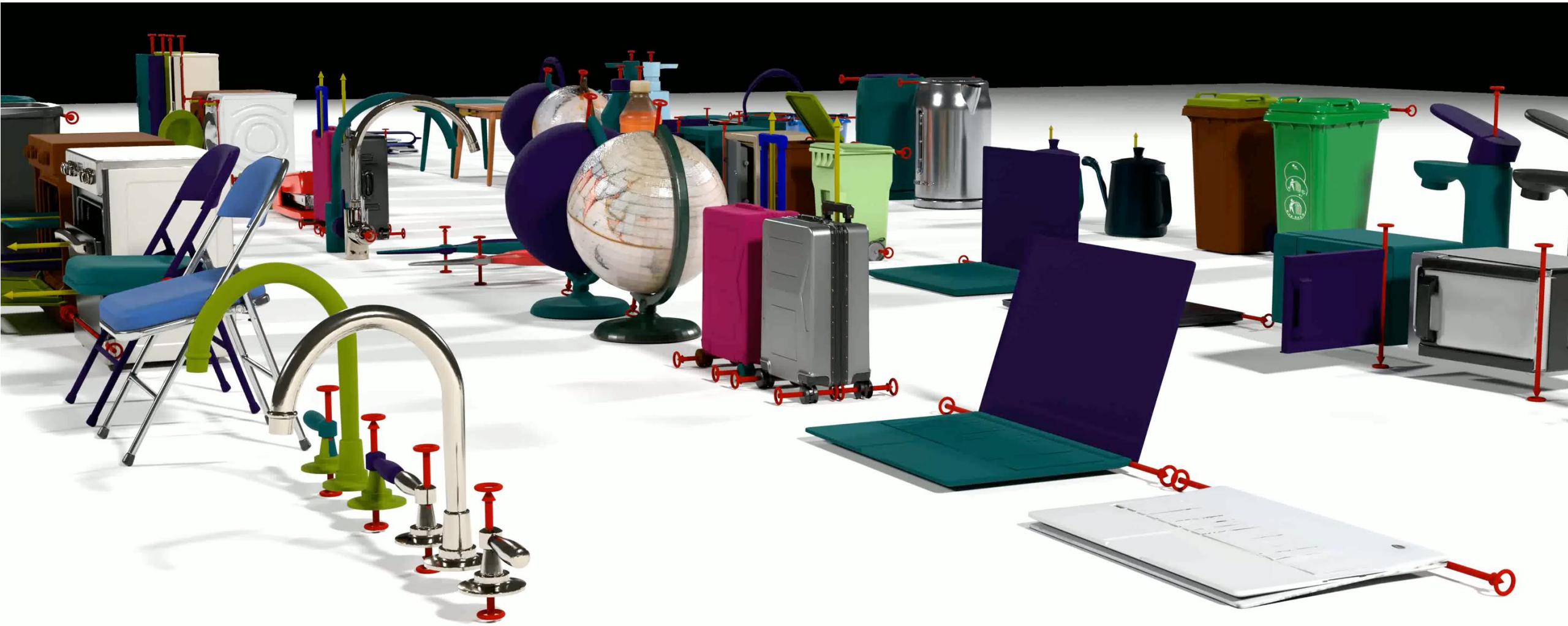
MEM: Multi-Scale Embodied Memory for Vision Language Action Models

Marcel Torne*^{1,2,†} Karl Pertsch*¹ Homer Walke^{3,†} Kyle Vedder¹ Suraj Nair¹ Brian Ichter¹
Allen Z. Ren¹ Haohuan Wang¹ Jiaming Tang^{1,4,†} Kyle Stachowicz^{1,3,†} Karan Dhabalia¹ Michael Equi¹
Quan Vuong¹ Jost Tobias Springenberg¹ Sergey Levine¹ Chelsea Finn¹ Danny Driess¹

<https://pi.website/research/memory>



Generative Simulation



Generative Simulation

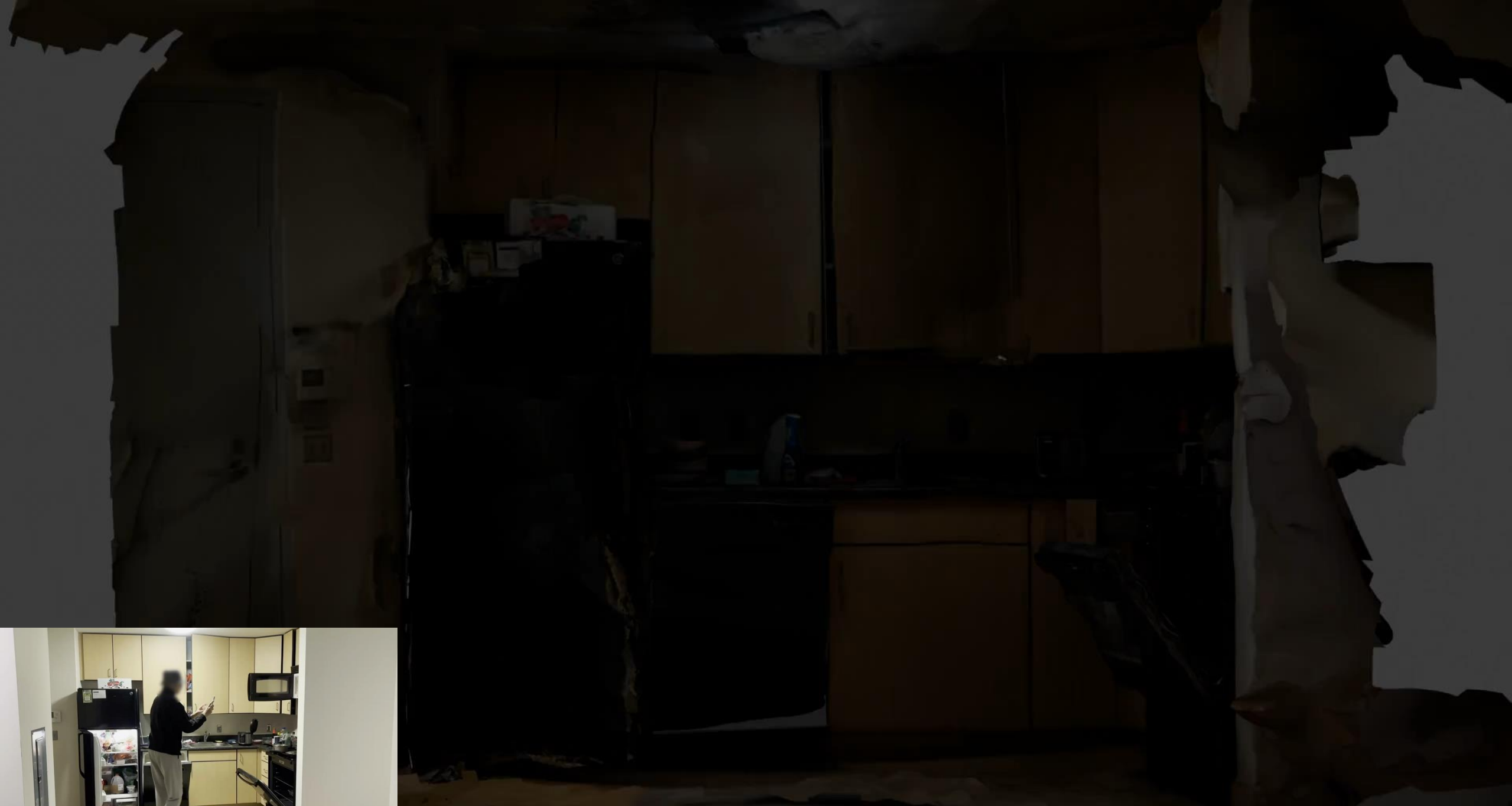




NeuROK: Generative 4D Neural Object Kinematics. Geng et al. CVPR 2026.



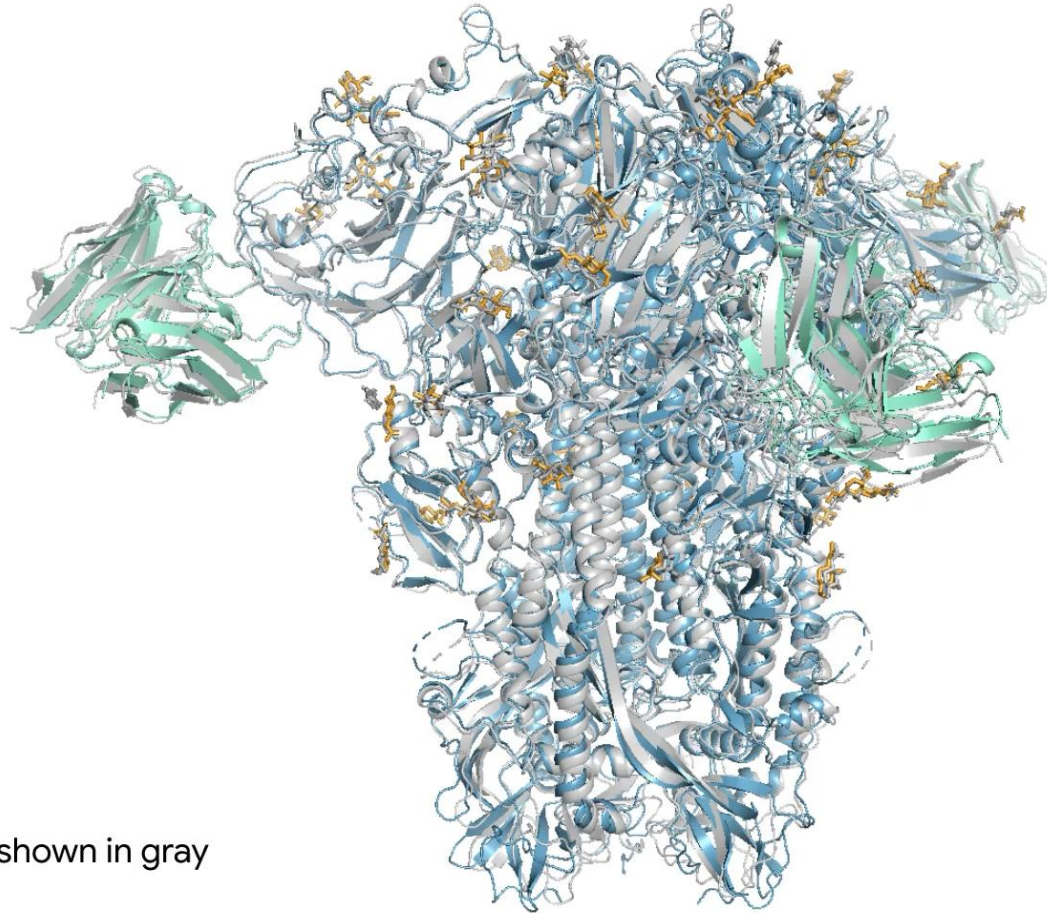
NeuROK: Generative 4D Neural Object Kinematics. Geng et al. CVPR 2026.



NeuROK: Generative 4D Neural Object Kinematics. Geng et al. CVPR 2026.

AI for Science

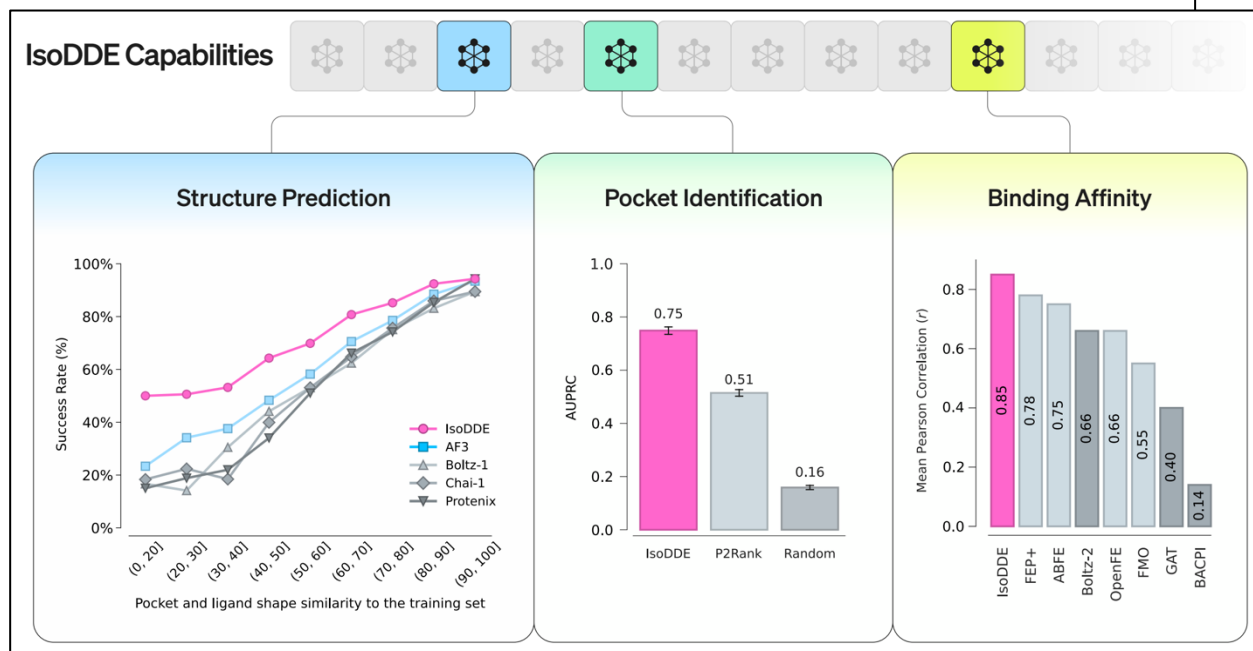
7PNM



Ground truth shown in gray

AlphaFold 3 – diffusion model predicting 3D folding structure of proteins

AI for Science

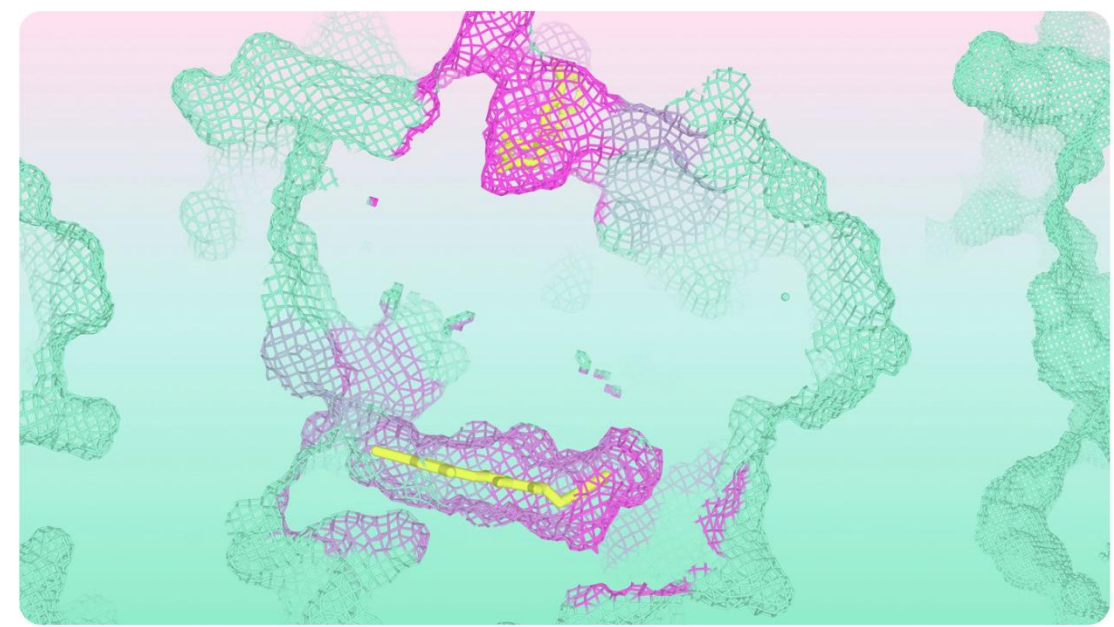


VISION

The Isomorphic Labs Drug Design Engine unlocks a new frontier beyond AlphaFold

FEBRUARY 10, 2026 | 5 MIN READ

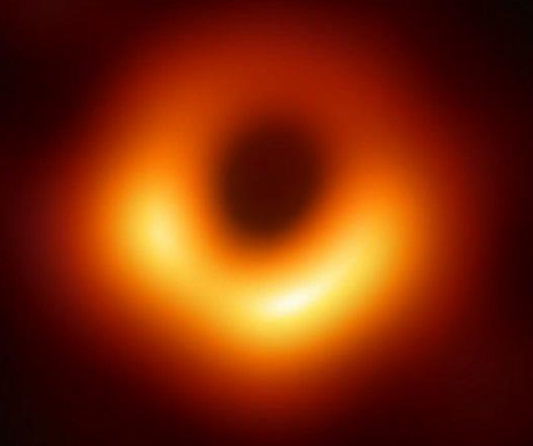
COPY URL



- Structure Prediction of Truly Novel Systems
- Opening a New Window for Complex Biologics
- A New Gold-Standard for Binding Affinity Prediction
- Expanding the Ligandable Proteome
- Advancing Drug Discovery

Today, we are excited to share an update on our progress towards a new frontier of drug design. We have unlocked a new paradigm of predictive accuracy in understanding our biomolecular world, allowing us to rationally design new medicines on a computer with unprecedented understanding and precision.

First Image of a Black Hole

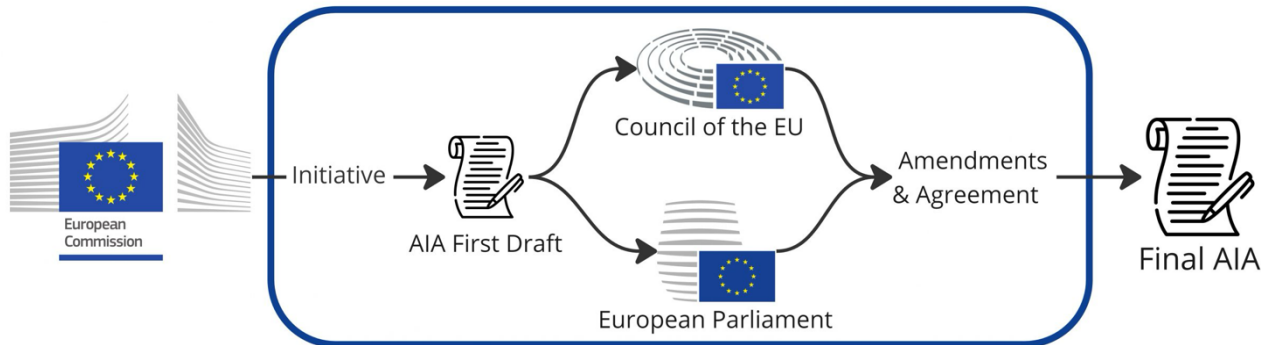


reconstructed from massive amounts of radio signals using Interferometry
Multi-view Geometry in Fourier domain (like MRI scans)

Ethics and Governance



EU Artificial Intelligence Act



Ordinary Legislative Procedure

<https://artificialintelligenceact.eu/>



AI (artificial intelligence)

Iran war heralds era of AI-powered bombing quicker than 'speed of thought'

Speed and scale of US military's AI war planning raises fears human decision-making may be sidelined



Academics say AI is collapsing the time required for military decision-making. Photograph: Majid Asgaripour/Reuters

Robert Booth and Dan Milmo

Tue 3 Mar 2026 06.00 GMT

Summary – Applications & Challenges

- The Era of Large Models
 - Multimodal Language Models (MLLMs)
 - In-context learning
 - Image and video generation, interactive world generation
 - Segment Anything (SAM), SAM 3D
- Expanding Horizons
 - Agentic AI
 - Physical AI
 - AI for Science
 - Ethics and governance

Topics We Didn't Have Time to Cover










- Latent predictive world models: [JEPA](#), [V-JEPA](#)
- Understanding x generation: [Representation Autoencoders \(RAEs\)](#)
- Vision as a first-class citizen: [Multimodal Pretraining](#)
- Spatial reasoning: [Theory of Space](#), [Spatial Mental Modeling](#)
- Autonomous driving: [UniAD](#), [Tesla FSD](#), [Waymo](#), [Wayve](#)
- Vision-Language-Action (VLA) models: [RT-X](#), [\$\pi_0\$](#) , [Gemini Robotics](#)
- ...

Visual General Intelligence

-Vision Research Toward the AGI Era-

CVPR 2026 Workshop

Invited Speakers

<p>Robert Geirhos Google DeepMind</p>  <p>View Profile</p>	<p>Aditi Raghunathan Carnegie Mellon University</p>  <p>View Profile</p>	<p>Matt Deitke Meta Superintelligence Lab</p>  <p>View Profile</p>	<p>Kristen Grauman University of Texas, Austin / Meta</p>  <p>View Profile</p>	<p>Yuki M. Asano University of Technology Nuremberg</p>  <p>View Profile</p>
<p>Alexei A. Efros University of California, Berkeley</p>  <p>View Profile</p>	<p>Jamie Shotton Wayve</p>  <p>View Profile</p>	<p>Kaiming He Massachusetts Institute of Technology / Google DeepMind</p>  <p>View Profile</p>	<p>Andrea Vedaldi University of Oxford / Meta</p>  <p>View Profile</p>	



Coursework 2

- Due at **12:00 noon on Tuesday 24 March 2026**
- Office hour: 3:45-4:45pm, Friday 13 March

Codex, Why You Are So Slow?

An AI-generated song for vibe coders
by Johnny Vibe Building

- Grok for fast & smooth video motion
- Midjourney for nice Niji 7 anime style
- Kimi for aesthetic prompts
- Suno v5: Amazing!
- OpenAI: Please make GPT 5.3 xhigh lightning fast! 🤔

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

MLM17: Advanced Computer Vision

