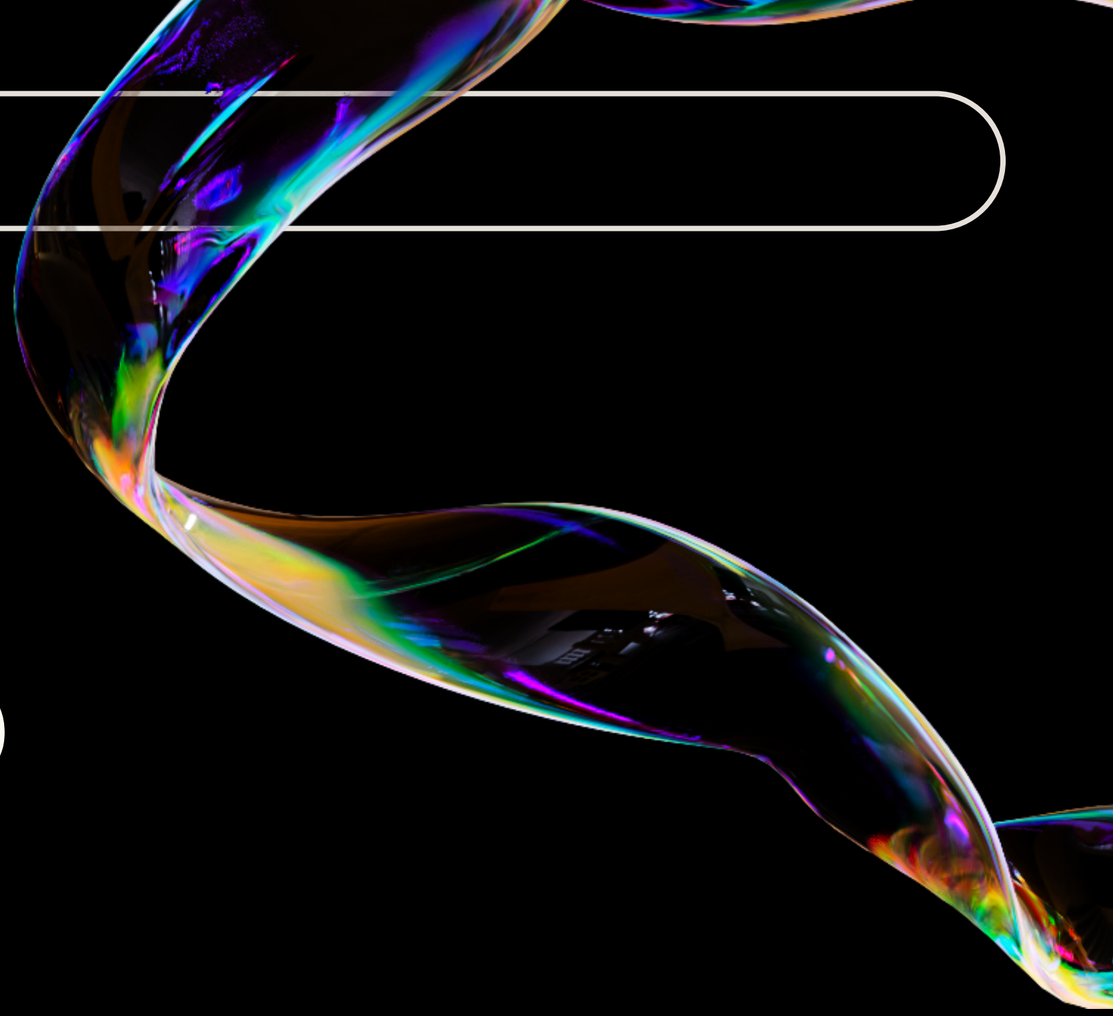
Yan (Stella) Si



HOW DO I RUN LLMS?

CDS Student Seminars

https://www.cdsseminar.com/

ysib@bu.edu

TODAY'S AGENDA

1 introduction

INTRODUCTION

With the rise of large language models (LLMs), researchers across disciplines are exploring how to use artificial intelligience in scientific discovery.

In social science, researchers are interested in using LLMs as reasoning/modeling tools and even as synthetic study participants.

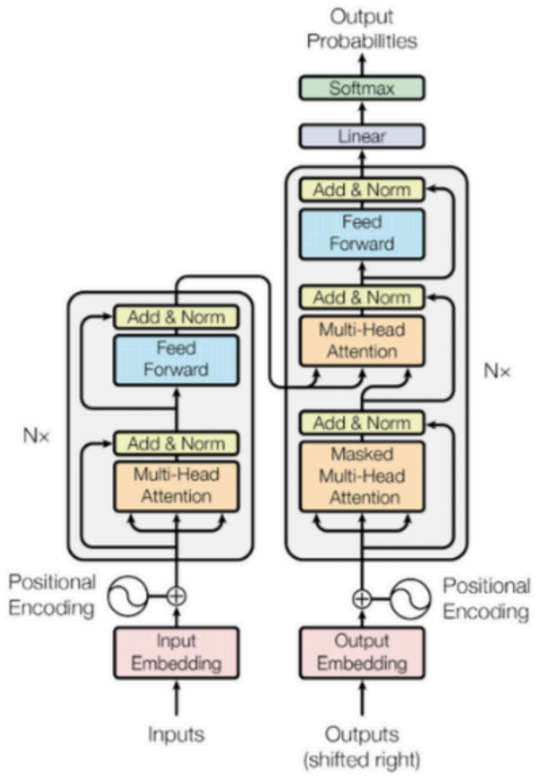
This workshop will show you how to actually use these models in practice.

The origins of LLMS

- **Rule-Based Era (1950s–1980s)**: Handcrafted linguistic rules and symbolic systems like ELIZA and SHRDLU.
- **Statistical NLP (1990s)**: Shift to data-driven methods using n-grams and Hidden Markov Models.
- **Word Embeddings (Early 2010s)**: Word2Vec and GloVe introduced dense semantic representations.
- **Neural Sequence Models (2014–2017)**: RNNs, LSTMs, and Seq2Seq with attention became standard for translation and generation.

Transformer

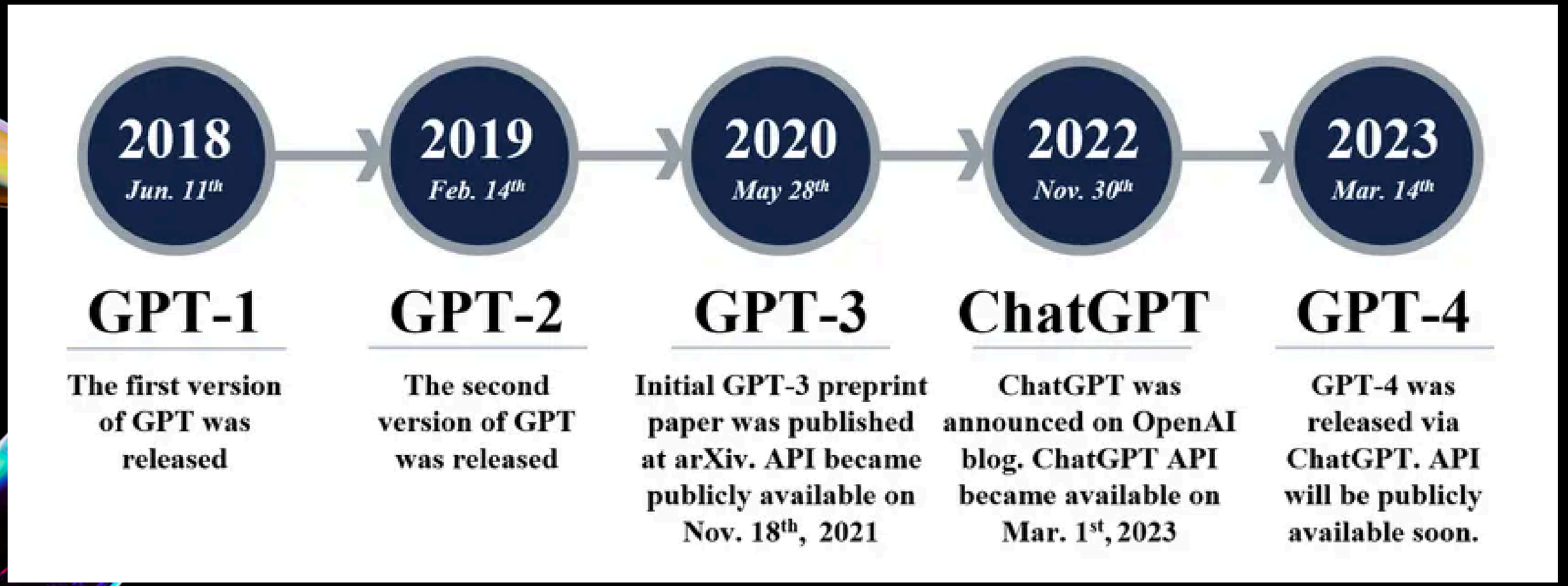
Attention Is All You Need



"Attention Is All You Need" (Vaswani et al., 2017) kickedoff the LLM revolution. In 2017, Google introduced the Transformers architecture, which made a breakthrough in how fast and how big we can train language models. It solved a key problem that RNNs struggled with: **long-range dependencies and parallelization**. This led the way to BERT, GPT, and beyond.

https://arxiv.org/pdf/1706.03762

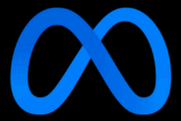
117 million | 1.5 billion | 175 billion | 175 billion | Estimated 1.8 trillion



https://medium.com/@seekmeai/the-evolution-of-openais-chatgpt-from-gpt-1-to-gpt-4-and-beyond-cdd15917961a


OpenAI


Grok


Llama


Nova


Gemini


Claude


Ministral


Phi


Deepseek


Qwen


Kimi


GLM


Sonar


Granite


Cohere


Reka

2025 International Mathematical Olympiad

Country	Team size			P1	P2	P3	P4	P5	P6	Total	Rank
	All	M	F								
People's Republic of China	6	6	0	42	42	42	42	42	21	231	1
United States of America	6	5	1	42	42	40	41	42	9	216	2
Republic of Korea	6	5	1	42	42	42	42	31	4	203	3
Japan	6	6	0	41	36	31	42	36	10	196	4
Poland	6	5	1	42	42	28	41	42	1	196	4
Israel	6	6	0	42	42	34	42	32	2	194	6
India	6	6	0	42	42	28	42	36	3	193	7
Singapore	6	6	0	42	42	26	42	38	1	191	8
Vietnam	6	5	1	40	42	27	42	34	3	188	9
Türkiye	6	6	0	42	30	28	42	42	2	186	10

DeepMind: 35/42, OpenAI: 35/42

https://medium.com/nice-math-problems/imo-2025-results-humans-still-rule-0fb25da2ead7

2025 International Mathematical Olympiad

Model	Acc	Cost	1	2	3	4	5	6
gemini-2.5-pro	31.55%	$431.97	14%	0%	71%	46%	57%	0%
o3 (high)	16.67%	$223.33	0%	0%	7%	36%	57%	0%
o4-mini (high)	14.29%	$103.34	16%	0%	5%	46%	18%	0%
Grok 4	11.90%	$527.85	13%	4%	18%	13%	25%	0%
DeepSeek-R1-0528	6.85%	$59.50	4%	0%	5%	0%	32%	0%

However, current models do not reach the same level.

https://medium.com/nice-math-problems/imo-2025-results-humans-still-rule-0fb25da2ead7

The trends suggests a HUGE potential in innovation.

How can intelligient LLMs advance scientific research?

Published as a conference paper at ICLR 2025

LLM-SR: Scientific Equation Discovery via Programming with Large Language Models

Parshin Shojaee[1*] **Kazem Meidani**[2*] **Shashank Gupta**[3]
Amir Barati Farimani[2] **Chandan K. Reddy**[1]

[1]Virginia Tech [2]Carnegie Mellon University [3]Allen Institute for AI

Abstract

Mathematical equations have been unreasonably effective in describing complex natural phenomena across various scientific disciplines. However, discovering such insightful equations from data presents significant challenges due to the necessity of navigating extremely large combinatorial hypothesis spaces. Current methods of equation discovery, commonly known as symbolic regression techniques, largely focus on extracting equations from data alone, often neglecting the domain-specific prior knowledge that scientists typically depend on. They also employ limited representations such as expression trees, constraining the search space and expressiveness of equations. To bridge this gap, we introduce LLM-SR, a novel approach that leverages the extensive scientific knowledge and robust code generation capabilities of Large Language Models (LLMs) to discover scientific equations from data. Specifically, LLM-SR treats equations as programs with mathematical operators and combines LLMs' scientific priors with evolutionary search over equation programs. The LLM iteratively proposes new equation skeleton hypotheses, drawing from its domain knowledge, which are then optimized against data to estimate parameters. We evaluate LLM-SR on four benchmark problems across diverse scientific domains (e.g., physics, biology), which we carefully designed to simulate the discovery process and prevent LLM recitation. Our results demonstrate that LLM-SR discovers physically accurate equations that significantly outperform state-of-the-art symbolic regression baselines, particularly in out-of-domain test settings. We also show that LLM-SR's incorporation of scientific priors enables more efficient equation space exploration than the baselines[1].

1 Introduction

The emergence of Large Language Models (LLMs) has marked a significant milestone in artificial intelligence, showcasing remarkable capabilities across various domains (Achiam et al., 2023). As LLMs continue to evolve, researchers are exploring innovative ways to harness their potential for solving complex problems such as scientific discovery (Wang et al., 2023a; AI4Science & Quantum, 2023). Their ability to process and comprehend vast amounts of scientific literature, extract relevant information, and generate coherent hypotheses has recently opened up new avenues for accelerating scientific progress (Zheng et al., 2023b; Ji et al., 2024). Additionally, by leveraging their ability to

INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Abdel-Rehim A *et al.* 2025 Scientific hypothesis generation by large language models: laboratory validation in breast cancer treatment. *J. R. Soc. Interface* **22**: 20240674.
https://doi.org/10.1098/rsif.2024.0674

Received: 27 September 2024
Accepted: 28 April 2025

Subject Category:
Life Sciences–Mathematics interface

Subject Areas:
biotechnology

Scientific hypothesis generation by large language models: laboratory validation in breast cancer treatment

Abbi Abdel-Rehim[1], Hector Zenil[1,2,3,4,5], Oghenejokpeme Orhobor[1], Marie Fisher[6], Ross J. Collins[6], Elizabeth Bourne[6], Gareth W. Fearnley[6], Emma Tate[6], Holly X. Smith[6], Larisa N. Soldatova[7] and Ross King[1,8]

[1]Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, UK
[2]Algorithmic Dynamics Lab, Research Departments of Biomedical Computing and Digital Twins, School of Biomedical Engineering and Imaging Sciences, King's Institute for AI, King's College London, London, England, UK
[3]Oxford Immune Algorithmics, Oxford University Innovation and London Institute for Healthcare Engineering, London, England, UK
[4]Cancer Interest Group, The Francis Crick Institute, London, England, UK
[5]Defence and National Security, The Alan Turing Institute, British Library, London, England, UK
[6]Arctoris Ltd, Oxford, UK
[7]Computer Science, Goldsmiths University of London, London, UK
[8]Department of Computer Science and Engineering, Chalmers University, Gothenburg, Sweden

HZ, 0000-0003-0634-4384; RK, 0000-0001-7208-4387

Large language models (LLMs) have transformed artificial intelligence (AI) and achieved breakthrough performance on a wide range of tasks. In science, the most interesting application of LLMs is for hypothesis formation. A feature of LLMs, which results from their probabilistic structure, is that the output text is not necessarily a valid inference from the training text. These are termed 'hallucinations', and are harmful in many applications. In science, some hallucinations may be useful: novel hypotheses whose validity may be tested by laboratory experiments. Here, we experimentally test the application of LLMs as a source of scientific hypotheses using the domain of breast cancer treatment. We applied the LLM GPT4 to hypothesize novel synergistic pairs of US Food and Drug

HAI Stanford University Human-Centered Artificial Intelligence

Policy Brief
HAI Policy & Society
May 2025

Simulating Human Behavior with AI Agents

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw,
Benjamin Mako Hill, Carrie J. Cai, Meredith Ringel Morris,
Robb Willer, Percy Liang, Michael S. Bernstein

AI agents have been gaining widespread attention among the general public as AI systems that can pursue complex goals and directly take actions in both virtual and real-world environments. Today, people can use AI agents to make payments, reserve flights, and place grocery orders for them, and there is great excitement about the potential for AI agents to manage even more sophisticated tasks.

However, a different type of AI agent—a simulation of human behaviors and attitudes—is also on the rise. These simulation AI agents aim to be useful at asking "what if" questions about how people might respond to a range of social, political, or informational contexts. If these agents achieve high accuracy, they could enable researchers to test a broad set of interventions and theories, such as how people would react to new public health messages, product launches, or major economic or political shocks. Across economics, sociology, organizations, and political science, new ways of simulating individual behavior—and the behavior of groups of individuals—could help expand our understanding of social interactions, institutions, and networks. While work on these kinds of agents is progressing, current architectures must cover some distance before their use is reliable.

Key Takeaways

Simulating human attitudes and behaviors could enable researchers to test interventions and theories and gain real-world insights.

We built an AI agent architecture that can simulate real people in ways far more complex than traditional approaches. Using this architecture, we created generative agents that simulate 1,000 individuals, each using an LLM paired with an in-depth interview transcript of the individual.

To test these generative agents, we evaluated the agents' responses against the corresponding person's responses to major social science surveys and experiments. We found that the agents replicated real participants' responses 85% as accurately as the individuals replicated their own answers two weeks later on the General Social Survey.

Because these generative agents hold sensitive data and can mimic individual behavior, policymakers and researchers must work together to ensure that appropriate monitoring and consent mechanisms are used to help mitigate risks while also harnessing potential benefits.

TODAY'S AGENDA

1 introduction

2 let’s build: chat interface and APIs

CHAT INTERFACE

Simpliest proof-of-concept?

Just use chat interfaces!
Examples: Gemini Pro, ChatGPT (GPT-5), Claude, Windsurf, Cursor, GitHub Copilot

Pros:

- simple and easy to use
- Strong context management and stability
- Generous free tiers (e.g., free Gemini Pro and GitHub Copilot for students)

Cons:

- Limited workflow customization and automation capabilities

LLM APIs

**Tip: Gemini does have a good free tier. Don’t add your billing info or you will get charged once you are past the free limit.

Simpliest proof-of-concept?

Just use chat interfaces!

Examples: Gemini Pro, ChatGPT (GPT-5), Claude, Windsurf, Cursor, GitHub Copilot

What are good fits for API tool calling?

- Multi-agent environments
- Multi-turn reasoning pipelines
- Custom evaluation of output quality
- Decision-making sequences

Pros:

- High control over model behavior and reasoning
- Enables complex, tailored applications and research workflows

Cons:

- Requires significant development effort and upfront coding costs

LLM APIs

**Tip: Gemini does have a good free tier. Don’t add your billing info or you will get charged once you are past the free limit.

How do you call LLM APIs?

Platforms with multiple models:

- **OpenRouter** – A platform that lets users access multiple AI models through a single interface. https://openrouter.ai/
- **Groq** – A hardware and inference platform designed to run AI models extremely fast. https://groq.com/

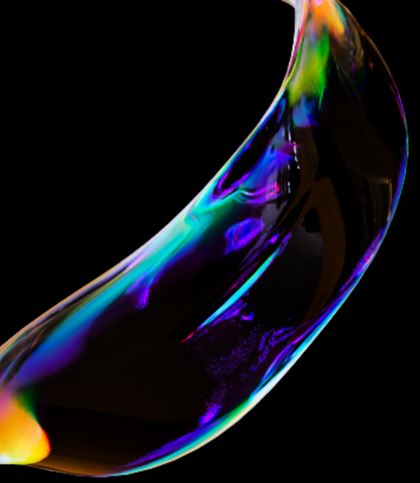
Direct Model Call:

- **OpenAI** – Access to GPT models such as GPT-4 and GPT-4o. https://platform.openai.com/docs/api-reference
- **Anthropic** – Claude models, including Claude 3 Opus, Sonnet, and Haiku. https://claude.com/platform/api
- **Google Gemini** – Gemini models available via Google AI Studio or Vertex AI. https://aistudio.google.com/
- **Cohere** – Command and Embed models for generation and embeddings. https://docs.cohere.com/
- **Mistral AI** – Mistral and Mixtral models offered through their native API. https://mistral.ai/

MODEL PRICING

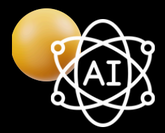
Model	Input $/1M tokens	Output $/1M tokens
GPT-5	$1.25	$10
Gemini 2.5 Flash	$0.30	$2.50
★ Gemini 2.5 Pro	$1.25	$10
GLM 4.5 (via Z.ai/OpenRouter)	$0.60	$2.20
Claude Sonnet 4	$3	$15
Claude Opus 4.1	$15	$75

https://velab.dev/blog/what-model-should-you-be-using/



DEMO

THE ART OF PROMPTING



Have AI edit your prompt

Use AI to refine your prompts. Draft what you want to say in one chat, then have the AI structure it into a clear prompt for another model.

Be clear on data structure

Pasting a full CSV without context can confuse an LLM. As with a Kaggle dataset or working with a colleague, first explain the dataset and each column. You can also give explicit instructions, such as a function, for how the model should process the data.

Seeding with starter model

To avoid wasting credits on weak models, start with a seed model for the LLM to refine. Some argue this risks originality, but like any scientific work, using prior literature is a valid foundation for discovery.

Be clear about output format

Be explicit if you need a Python function or a specific function name for the next step. State it clearly: “Write a Python function named X.” Expect to adjust the prompt continuously to correct errors or omissions.

TODAY'S AGENDA

1. introduction
2. let’s build: chat interface and APIs
3. demo
4. the art of prompting
5. open source models

OPEN SOURCE MODELS

What are good fits for using open source model?

- Fine-tune for specific tasks
- Need data privacy
- Cheap model
- Experimentation and develop new capabilities

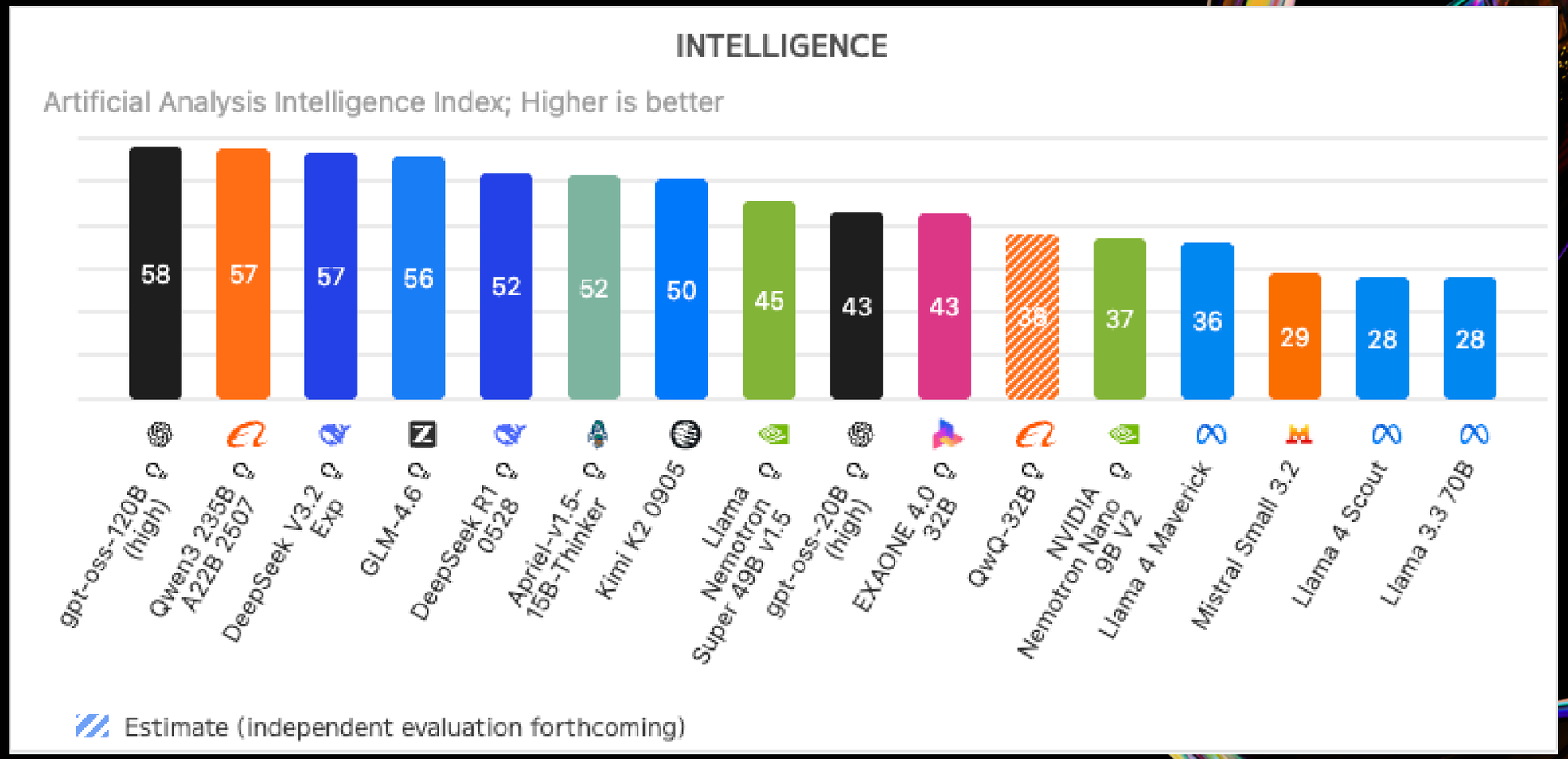
Pros:

- High control and customization
- Community and transparency
- Innovation

Cons:

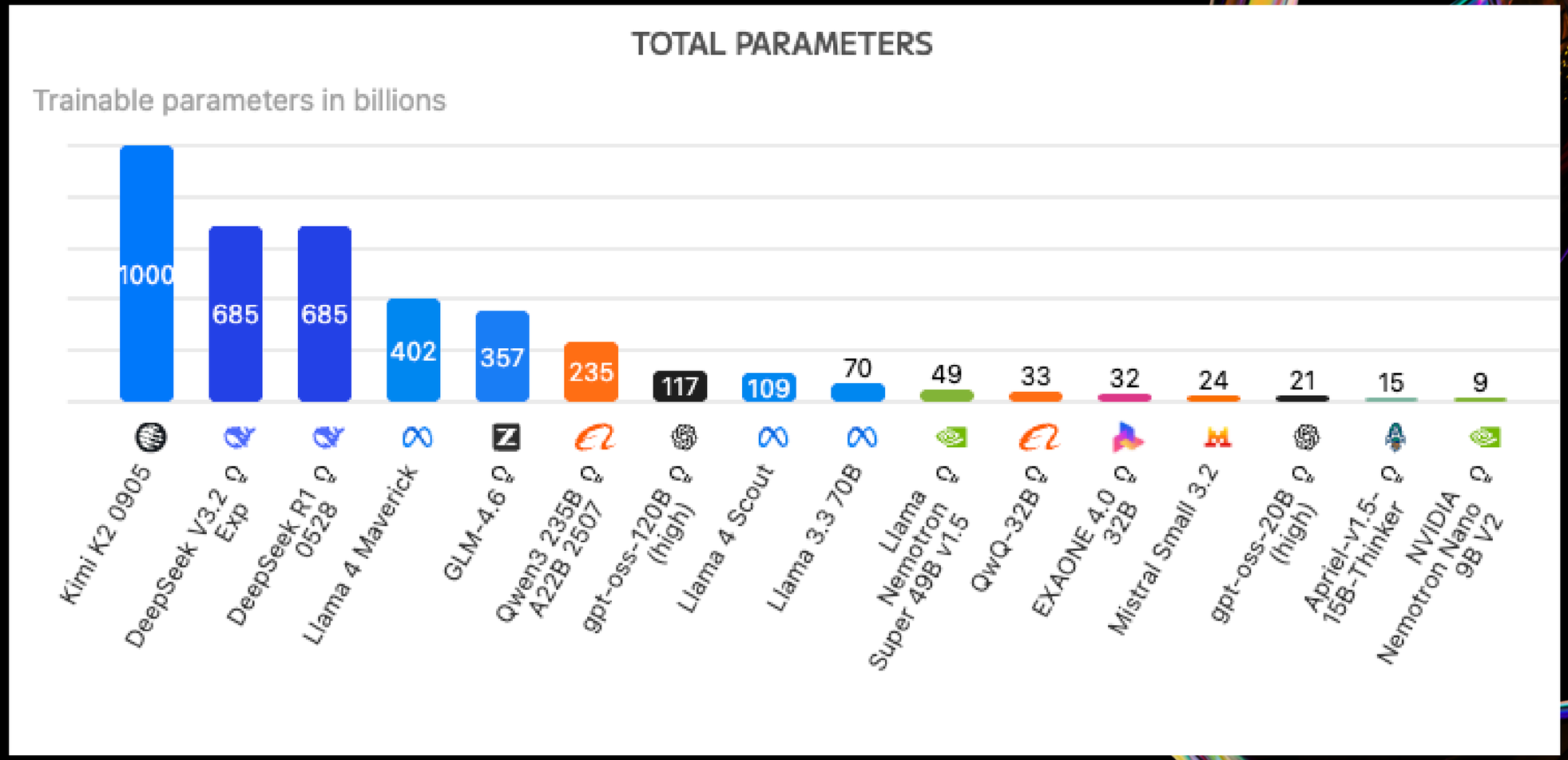
- Requires setup and Infra management
- Will not be as smart as proprietary models
- Often needs fine-tuning to perform well

OPEN SOURCE MODELS



https://artificialanalysis.ai/models/open-source

OPEN SOURCE MODELS



https://artificialanalysis.ai/models/open-source

OPEN SOURCE MODELS

Unsloth: Provides notebooks that simplify inference and fine-tuning pipelines (YC S24).

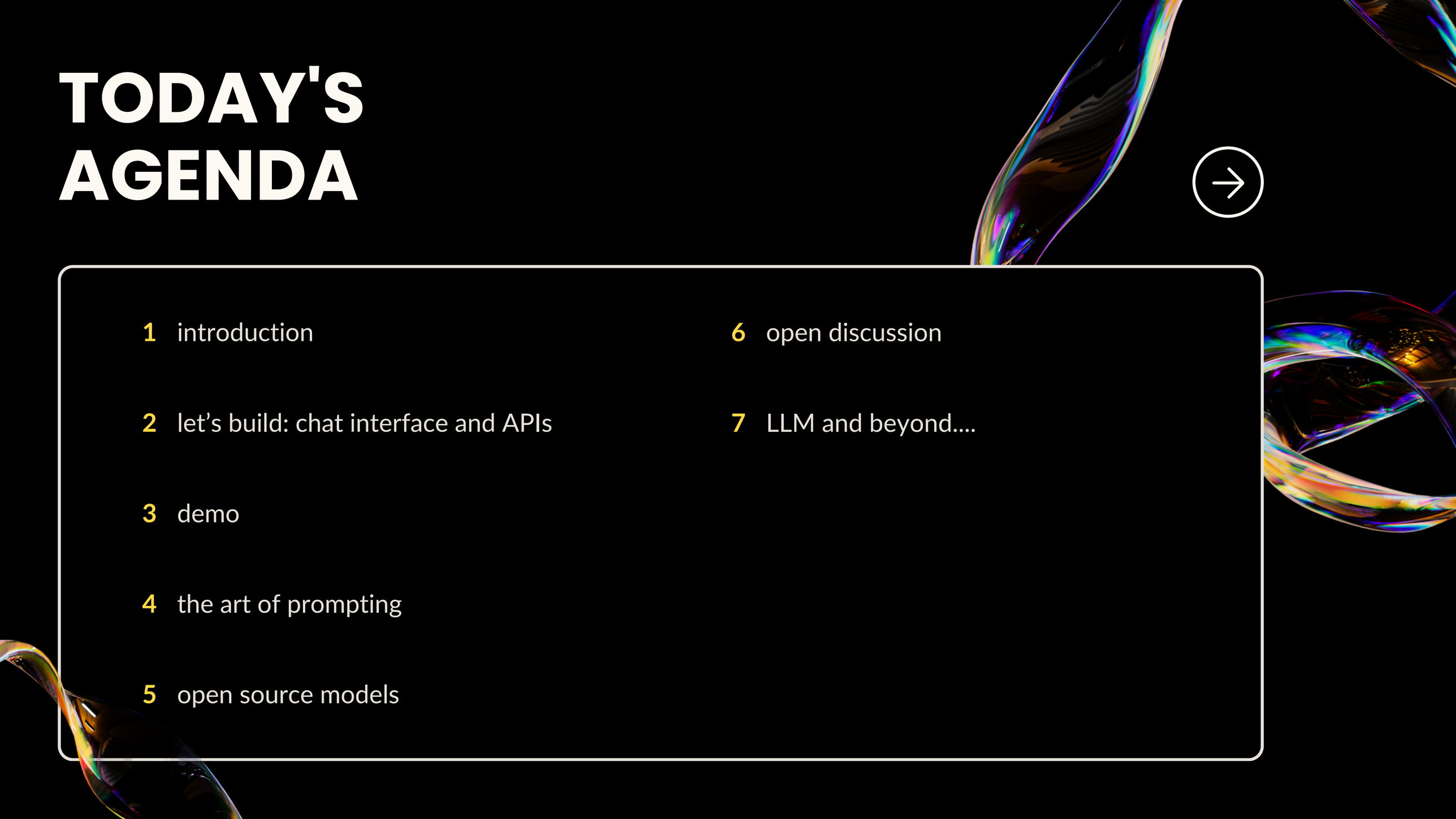
https://docs.unsloth.ai/get-started/unsloth-notebooks

vLLM: LLM inference engine optimized for GPU scaling, streaming, and batch processing (requires GPU).

Ollama: Local LLM platform for easy setup and experimentation with built-in model management.

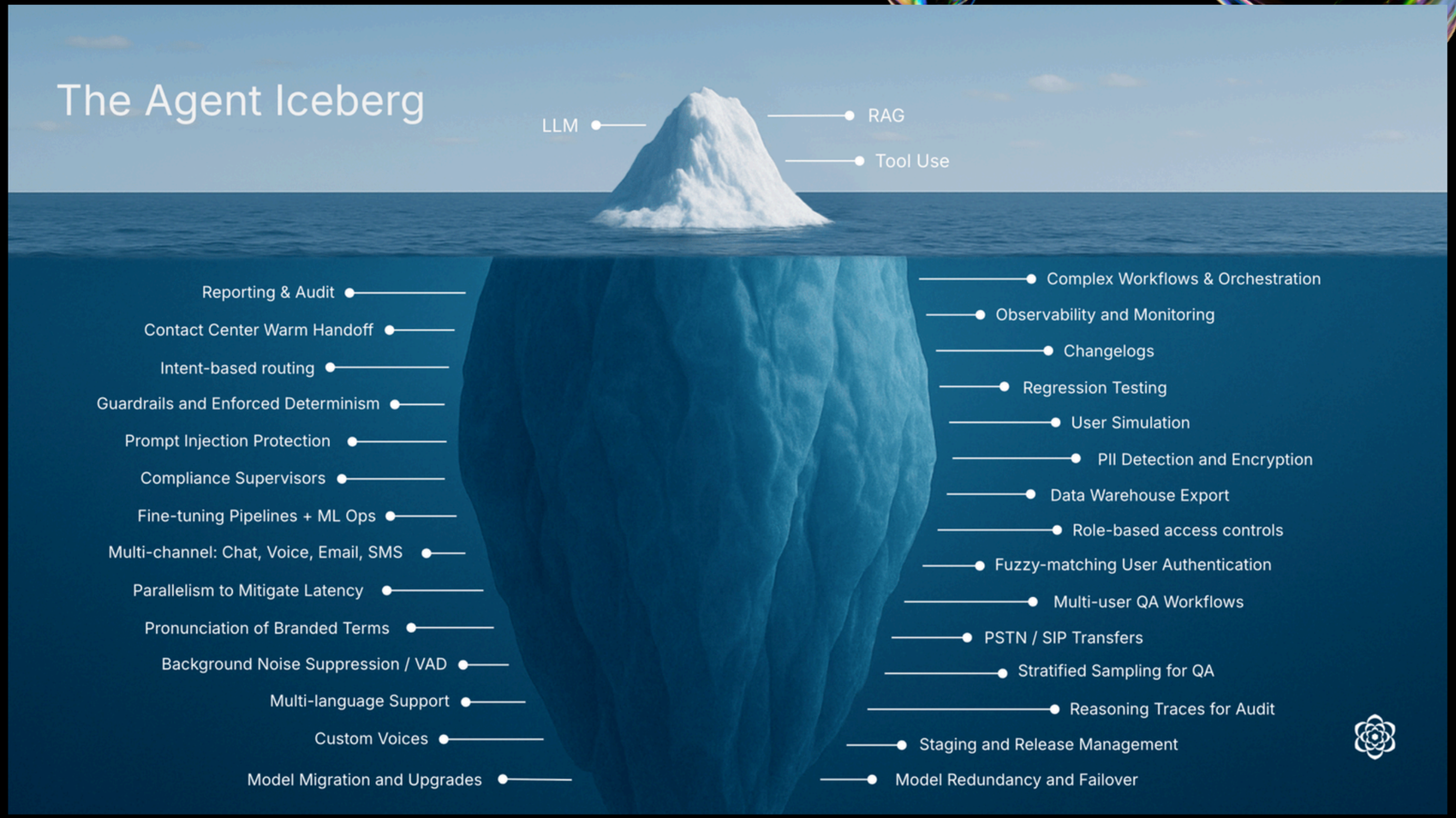
Article that explains the difference between vLLM and Ollama: https://robert-mcdermott.medium.com/performance-vs-practicality-a-comparison-of-vllm-and-ollama-104acad250fd

TODAY'S AGENDA



1. introduction
2. let’s build: chat interface and APIs
3. demo
4. the art of prompting
5. open source models
6. open discussion
7. LLM and beyond....

LLM and beyond....



https://sierra.ai/blog/the-challenge-with-rolling-your-own-agent

Resources and continued learning

Weekly news on latest LLM trends: https://velab.dev/

Deep comparitive analysis of current LLM models: https://artificialanalysis.ai/

Reddit: https://www.reddit.com/r/LLMDevs/

Discord channels

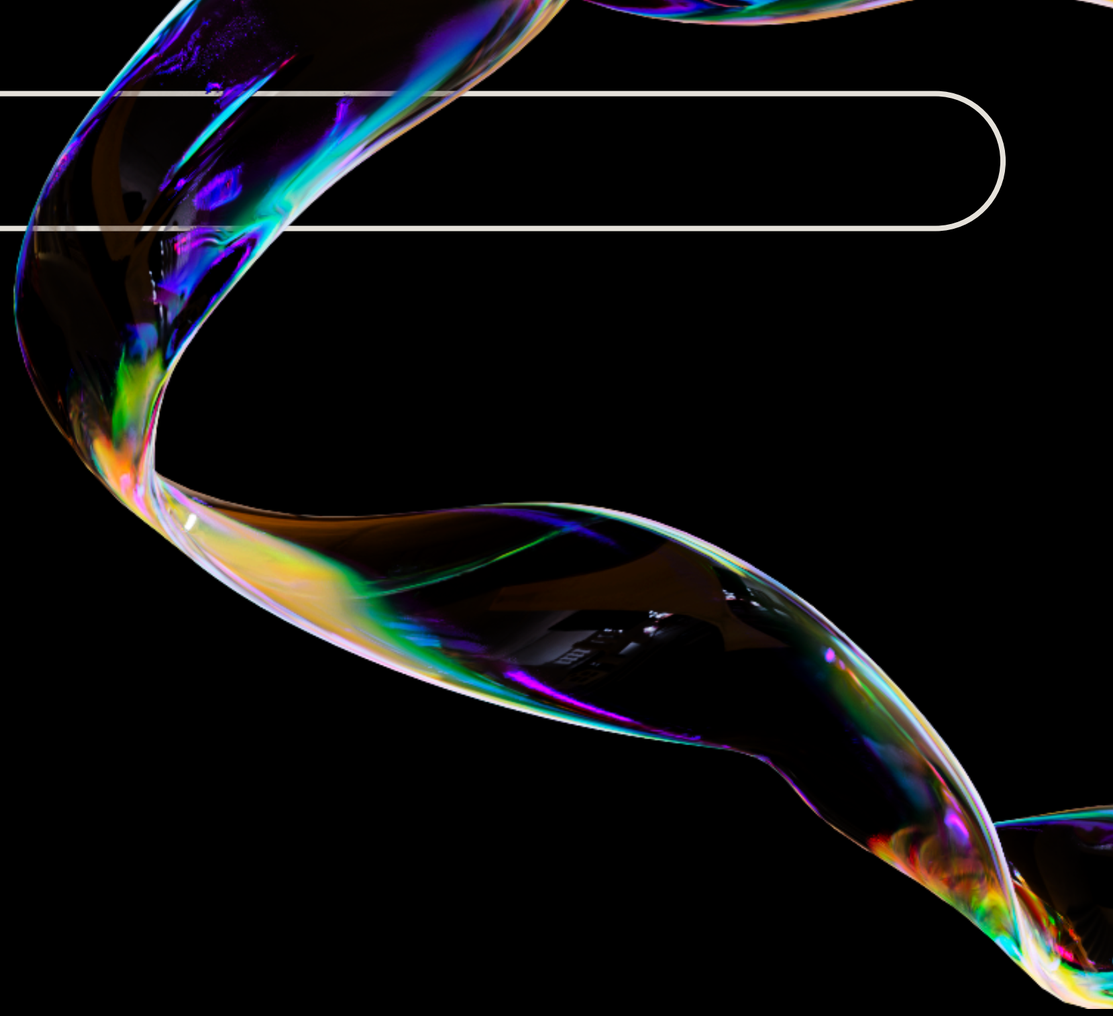
Cool papers:

OG agent framework paper: https://arxiv.org/abs/2303.11366

LLM powered foundational model of human cognition: https://arxiv.org/abs/2410.20268\

Cognitive model discovery via LLMs: https://arxiv.org/abs/2502.00879

Yan (Stella) Si



THANK YOU!!

CDS Student Seminars

https://www.cdsseminar.com/

ysib@bu.edu