

Cosmic Co-Discovery with AI

Where do we start?

IOANA CIUCĂ (CENTER FOR DECODING THE UNIVERSE @STANFORD)

Outline

- My Story
- Primer in large language models
- How can AI tools empower scientists
- From tools to agents: current capabilities, hopes and fears
- How I use AI in my research: tips and learnings

My Story: From Transylvania to California

- I then did quite a bit of maths.
- 2013: Sent an experiment on a rocket into space
- 2016: Got into a PhD at UCL, found my wonderful supervisor
- 2020-2021: End of the PhD in the pandemic, work at Alan Turing Institute on a data science for social good project
- 2021: 9 months of visa delays later, go to Australia
- 2023: Founded UniverseTBD
- 2024: Moved to Stanford

The first chatbot: meet ELIZA.

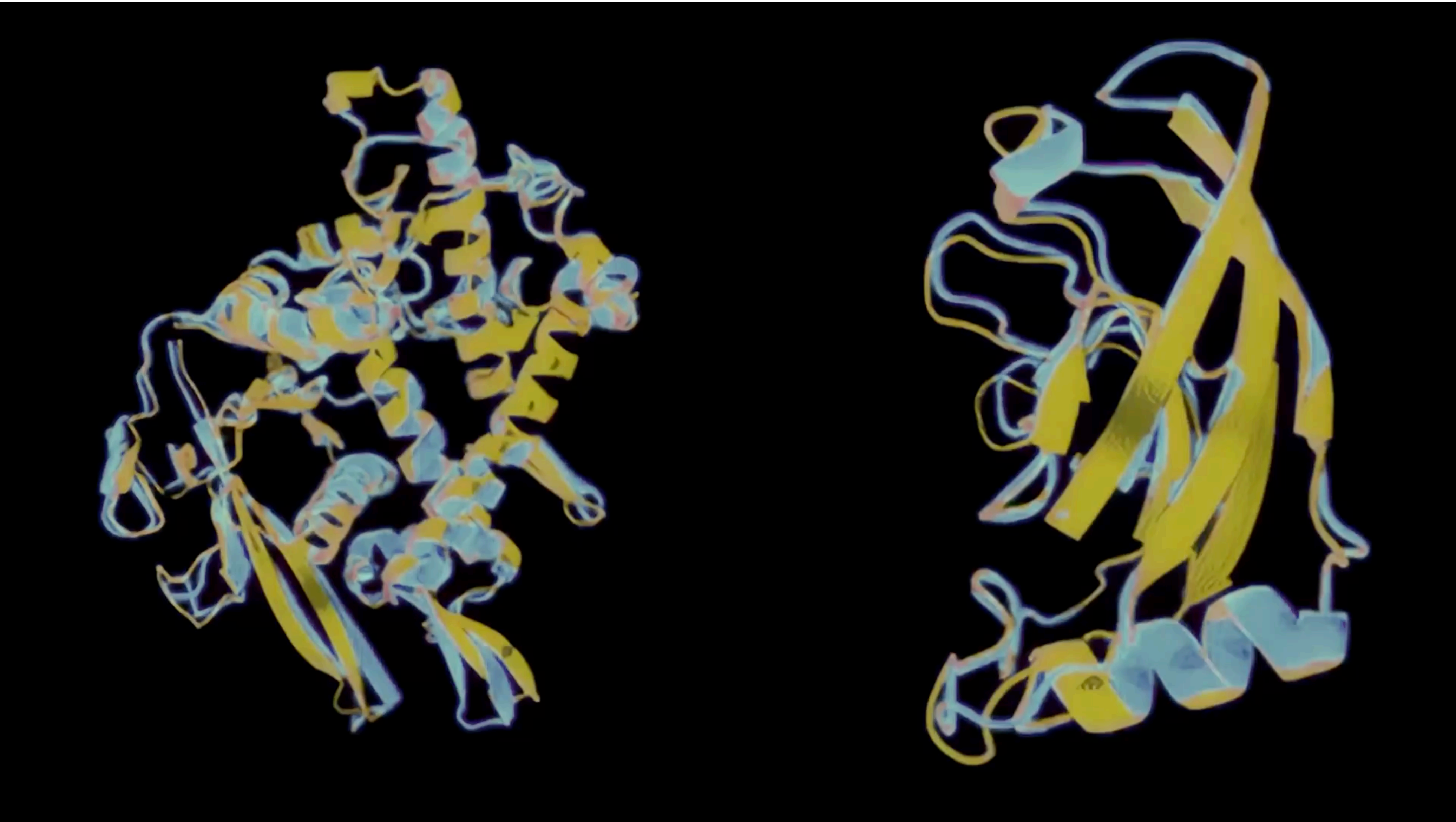
Talk to Eliza by typing your questions and answers in the input box.

> Hello, I am Eliza. I'll be your therapist today.

TYPE HERE

Created by Joseph Weizenbaum in 1966

Where are we now?



T1037 / 6vr4
90.7 GDT
(RNA POLYMERASE DOMAIN)

T1049 / 6y4f
93.3 GDT
(ADHESIN TIP)

● EXPERIMENTAL RESULT ● COMPUTATIONAL PREDICTION

New Results [Follow this preprint](#)

The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation

[ID](#) Kyle Swanson, [ID](#) Wesley Wu, Nash L. Bulaong, [ID](#) John E. Pak, [ID](#) James Zou
doi: <https://doi.org/10.1101/2024.11.11.623004>

This article is a preprint and has not been certified by peer review [what does this mean?].

Abstract Full Text Info/History Metrics [Preview PDF](#)

RESEARCH

Advanced version of Gemini with Deep
Think officially achieves gold-medal
standard at the International
Mathematical Olympiad

21 JULY 2025

Thang Luong and Edward Lockhart

How did we get here?

- Architectural advances: Transformers model long range dependencies
- Massive increase in compute
- Data availability
- The neural scaling laws (2020s)
- Economic Momentum

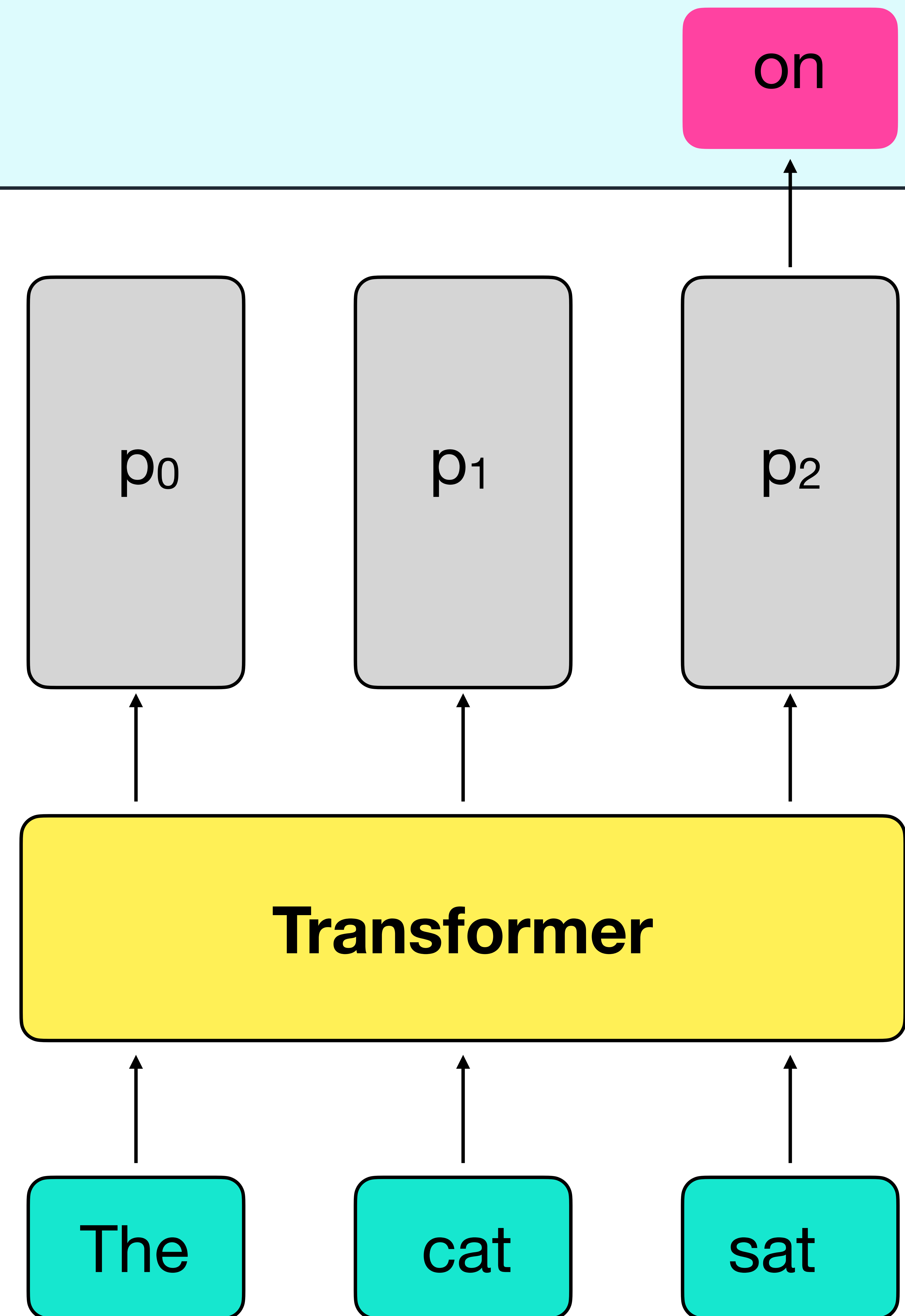


What does a Transformer do?

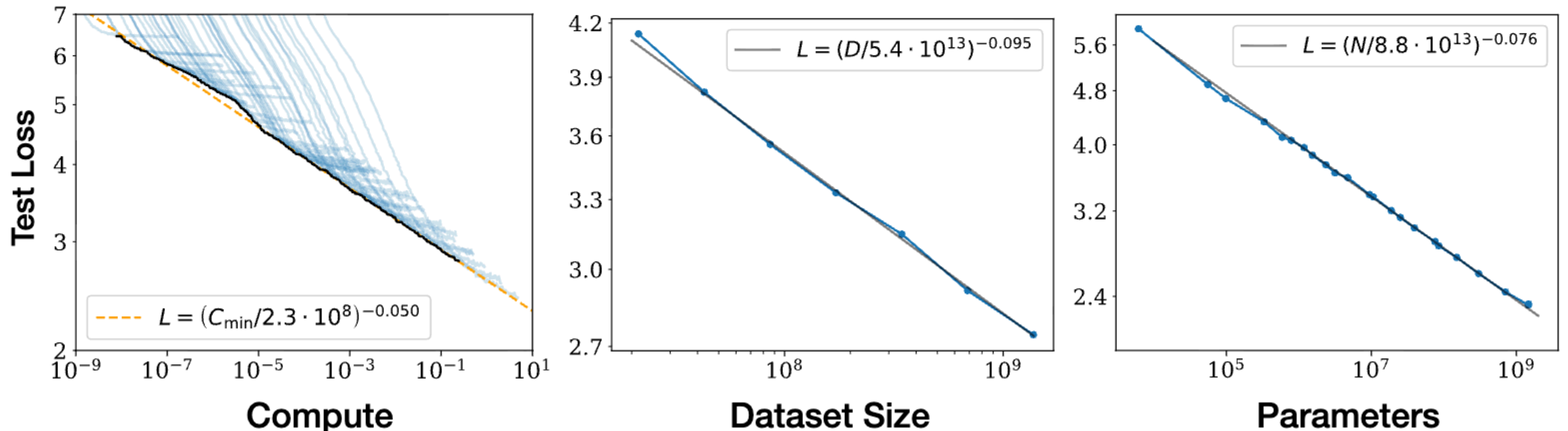
- Transformers are “**sequence modeling engines**”
Neel Nanda
- We input text and GPT-2 like transformers output a probability distribution over **tokens**.
- E.g. $p_2 = \text{PMF}(\text{next token} \mid \text{“The cat sat”})$
- For training, we want to update weights such that $p(\text{true token} \mid \text{context})$ is maximized.

Output

Input



A large language model is a scaled up transformer



- Test loss falls smoothly and predictably with parameters N , training tokens D , and compute.
- At fixed compute, scale data with model size, roughly 20 tokens per parameter.
- Some abilities appear with scale (“emergence”), though part of the effect reflects the choice of metrics.

How can LLMs help astronomers

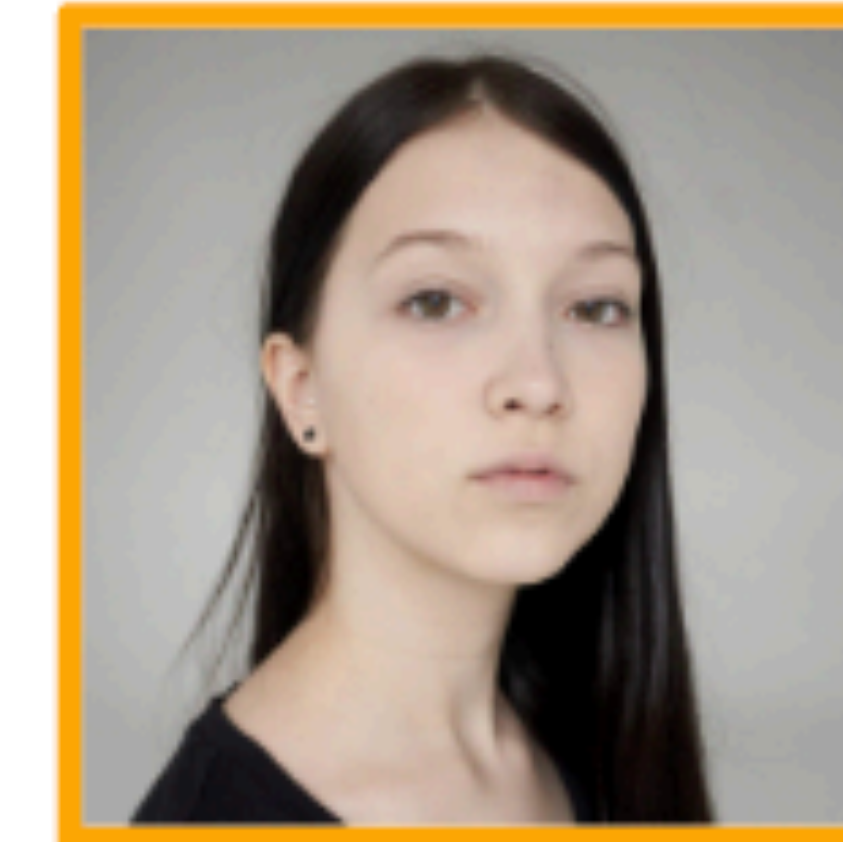
- We tried to answer this in a principled way
- Involved **astronomers** and **human-computer interaction** experts
- ~ 35 astronomers at STScI interacted with an LLM-powered chatbot, resulting in ~**350 real user queries**
- We grouped the queries in a meaningful way through **Open Domain Coding**
- We interviewed ~ **ten astronomers** to understand their **interaction** with the system



CENTER FOR LANGUAGE
AND SPEECH PROCESSING



STScI | SPACE TELESCOPE
SCIENCE INSTITUTE



Alina Hyk



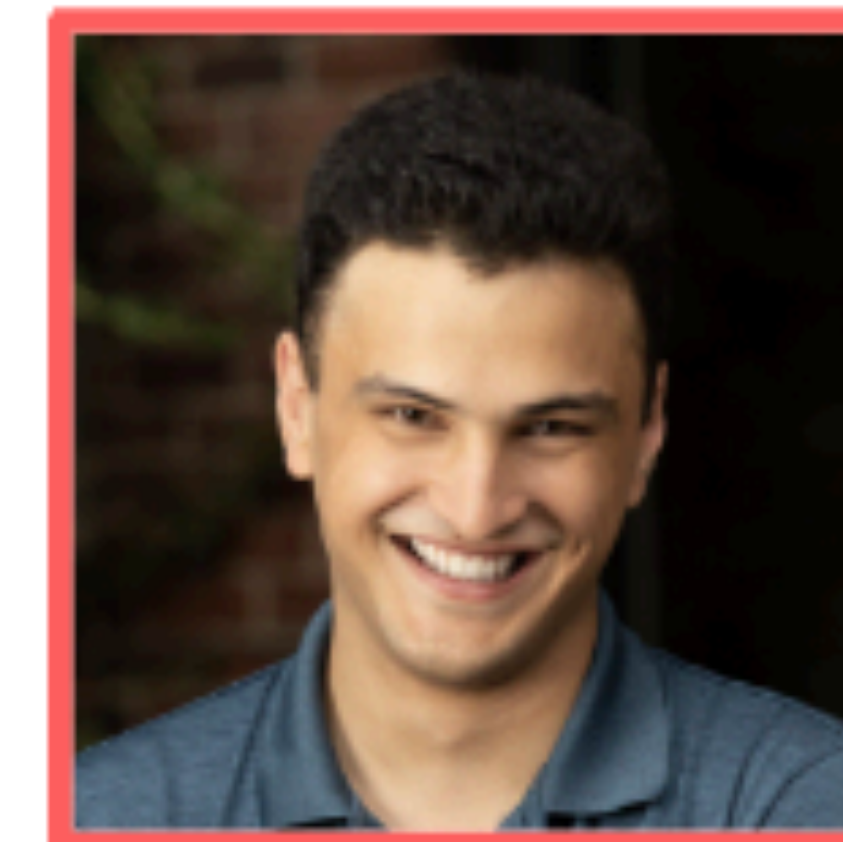
Kiera McCormick



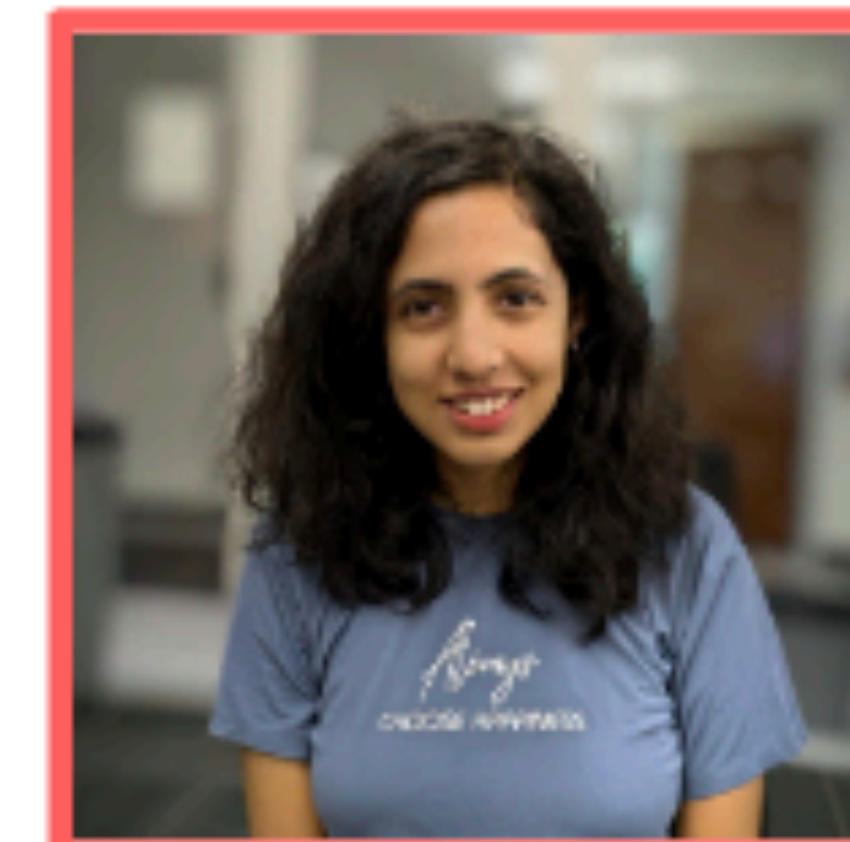
Charlie O'Neill



Christine Ye



Mikaeel Yunus



Elina Baral



Philipp Koehn



Sanjib Sharma



Anjalie Field



Jo Ciucă



Kartheik Iyer



John Wu

Interacting with the chatbot: question, feedback and reactions

37 Users (astronomers)

378 User queries

→ 269 in channel

→ 109 direct message

342 Emoji reactions

263 Feedback/comments

373 Retrievals

A **User A** Jun 4th at 10:28 AM
[@Ask astro-ph](#) what's the origin of the multiple Stellar populations in globular clusters?

5 replies

 **Ask astro-ph** APP Jun 4th at 10:29 AM
The origin of multiple stellar populations (MSPs) in globular clusters (GCs) remains a complex and unresolved issue in stellar astrophysics. Various scenarios have been proposed, including binary interactions, which can produce rapidly rotating stars with different properties from single stars ([1405.6729](#)), and the influence of massive and supermassive stars on intracluster gas during the early stages of GC formation ([1611.08855](#)). Despite extensive research, no single scenario fully explains the spectroscopic and photometric complexity observed in GCs ([2206.10564](#)).

👍 1 👎 1 😊

B **User B** Jun 4th at 10:38 AM
Is this answer any good?

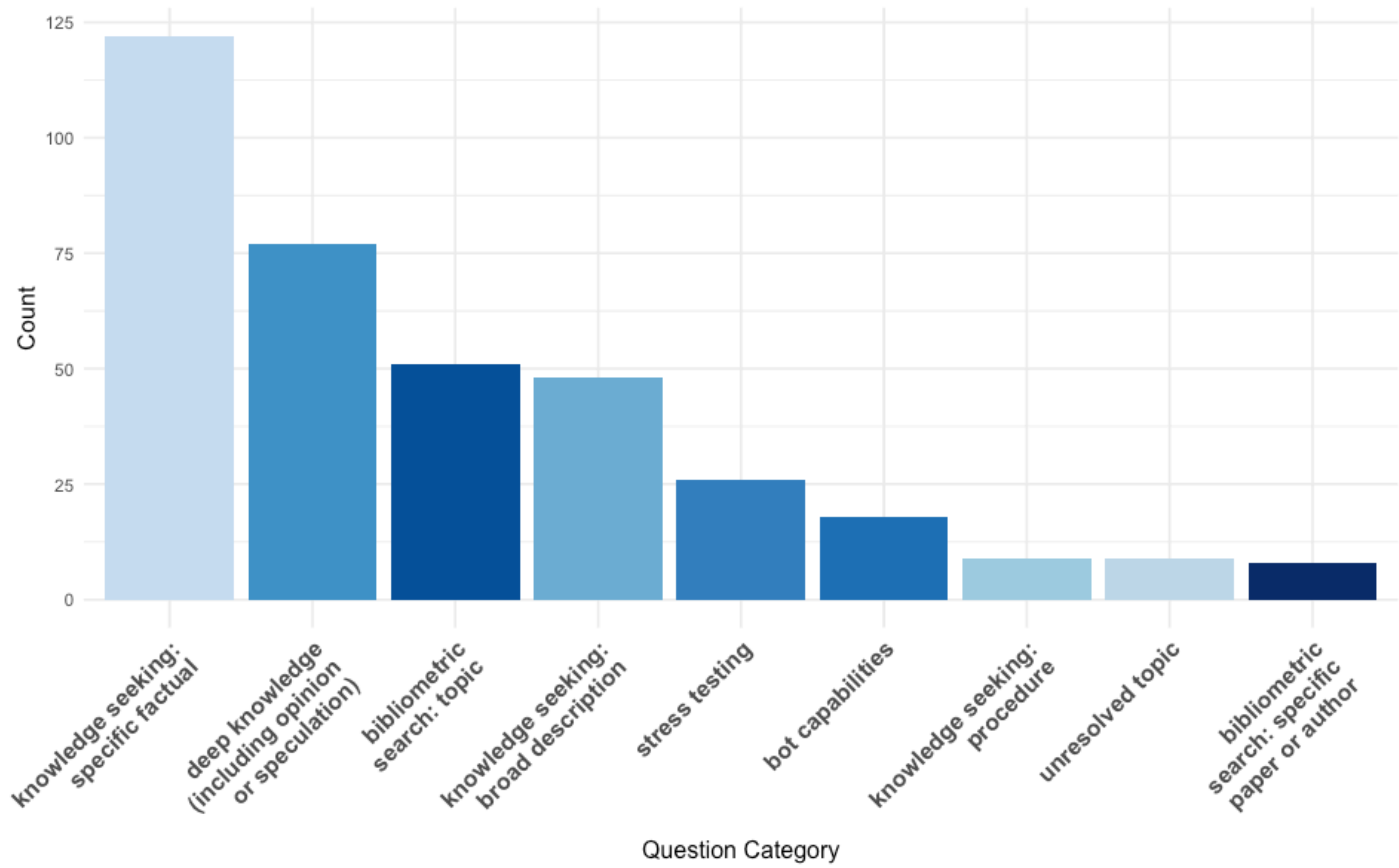
A **User A** Jun 4th at 10:40 AM
It's very generic and is not wrong but it misses some of the hypotheses.
I will try again later with a more focused question

B **User B** Jun 4th at 10:41 AM
Gotcha! Makes sense 😊

Users can rate and comment on the bot's responses to their queries

Users can view, comment on, and rate queries by other users

Most sought: Specific factual followed by deep knowledge

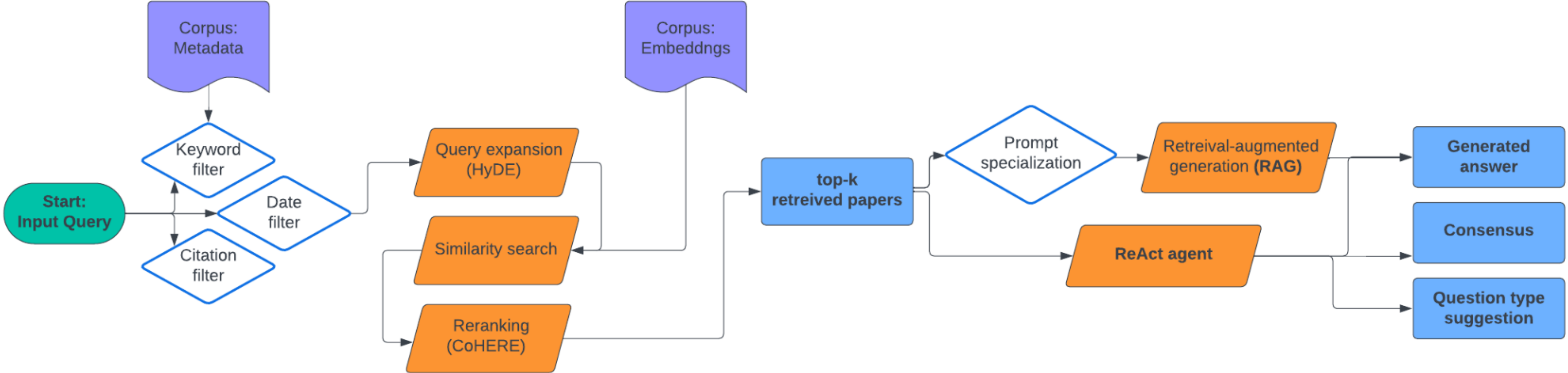




Kartheik Iyer

- AI-powered framework for **semantic literature search** in astronomy
- Uses **natural language** to explore over 350,000 papers
- Combines advanced **retrieval techniques** with LLM synthesis
- Aims to accelerate astronomy research by enhancing **literature review and knowledge discovery**





Question

Metadata
Filtering

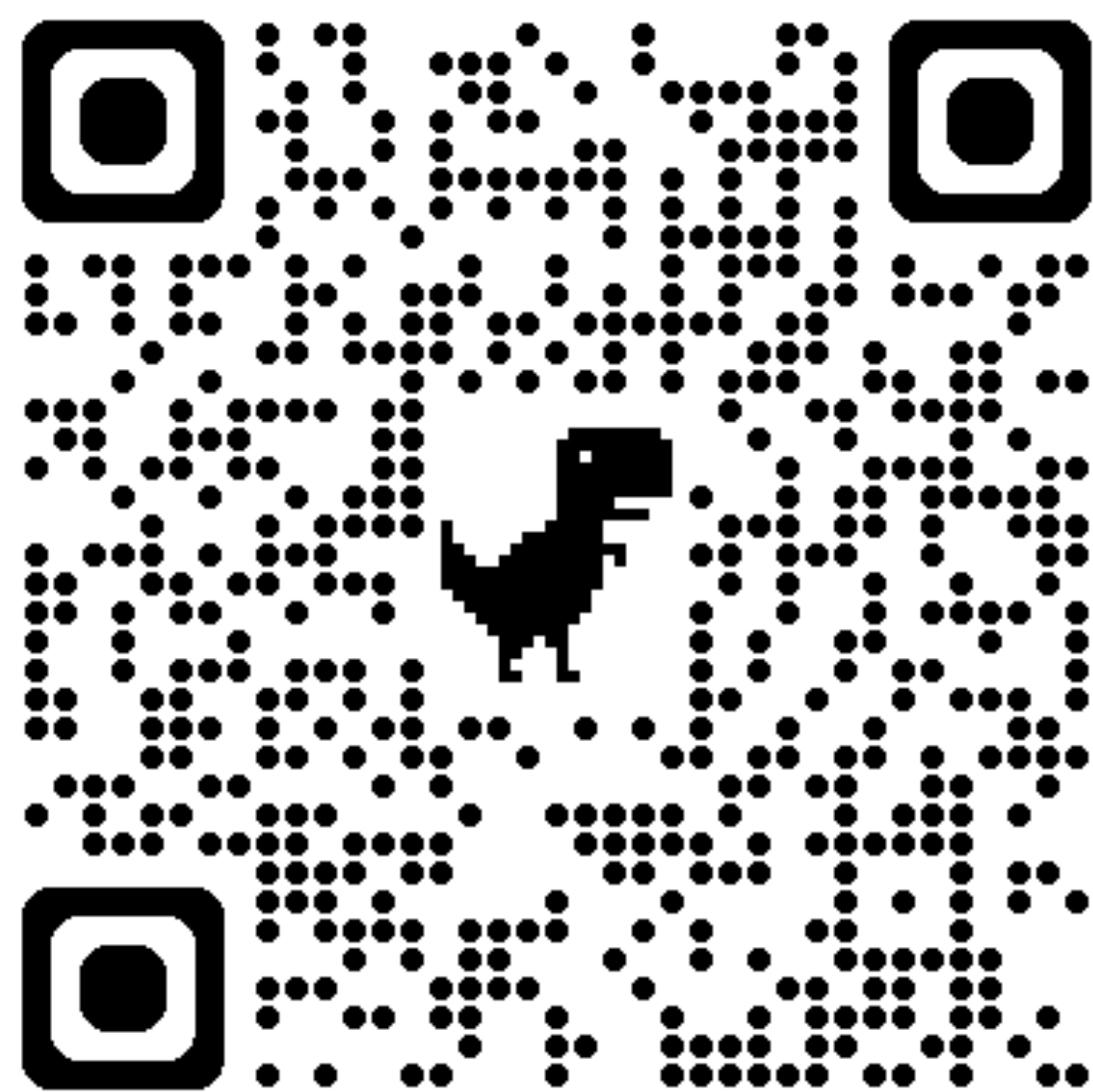
Retrieval using Embeddings

Generation

Answer

Finding meaning with Pathfinder

- Visualise **350,000+** astronomy papers in a 2D “map”
- Cluster papers by **semantic similarity**, not just keywords
- Identify links between subfields and **emerging ideas**



<https://explorepathfinder.netlify.app>

Iyer et al. 2024

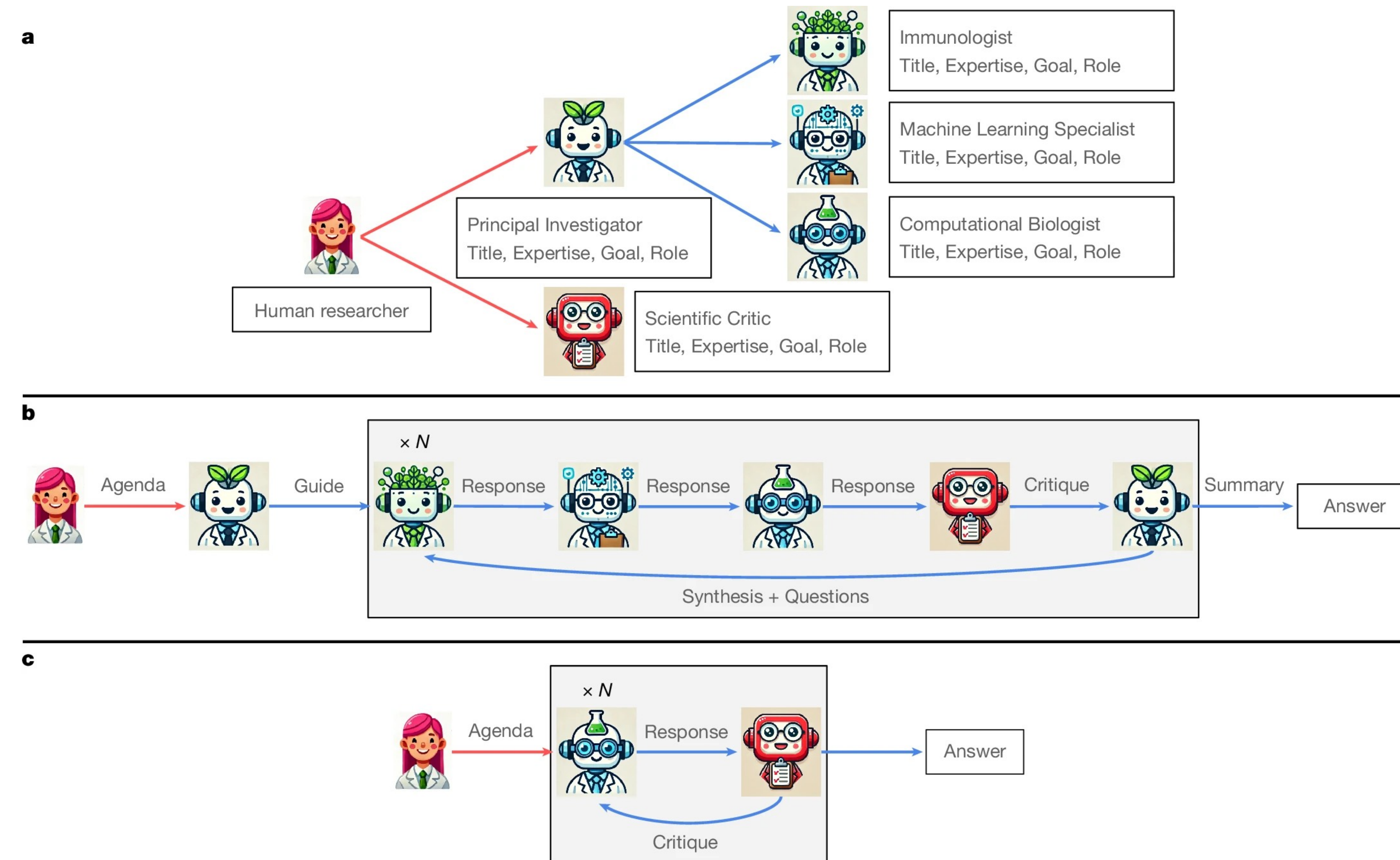
A map of the lands of
astronomy papers

(astro-ph as on July 2024)
Browse at <https://pfd.r.app/>



From LLMs to AI agents: how capable are they?

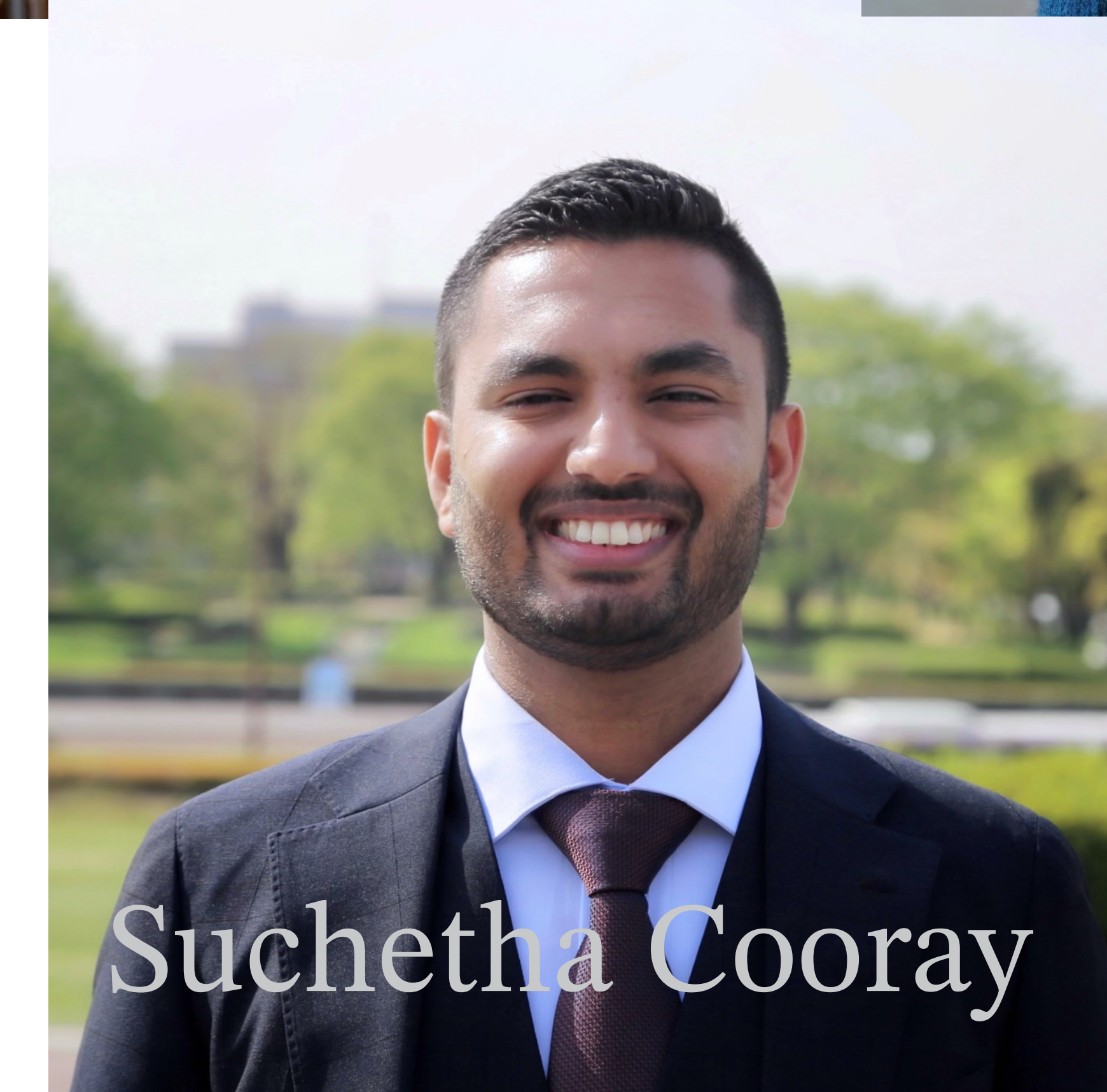
- An agent: **LLM with actions**
- The loop is plan → act → observe → revise, over tools such as literature search, code execution, catalogs, and telescopes
- **Denario + cmbagent**: idea → analysis → draft paper
- In other fields, Virtual Lab which designed novel nanobodies
- Sakana AI's AI Scientist, Google Co-Scientist
- How good are these agents at astrophysics work?



Swanson et al. 2025

ReplicationBench: end to end paper replication

- Led by Stanford third-year undergraduate researcher Christine Ye together with our team @newly-founded Center for Decoding the Universe
- Goal for agents: **replicate methodology and main findings** from input manuscript and data
- Objective evaluation on primarily **numerical results**
- Human evaluation to assess understanding of task, quality of execution and cheating/hacking



[illegible]

File Name	Size	Download	Share
NGC1300-89asec_MAPS.fits	1.42 GB		
hslp_phangs-jwst_jwst_miri_ngc1300_f1000w_vlp0p1_img...	154 MB		
hslp_phangs-jwst_jwst_miri_ngc1300_f1300w_vlp0p1_img...	154 MB		
hslp_phangs-jwst_jwst_miri_ngc1300_f2100w_vlp0p1_img...	154 MB		
hslp_phangs-jwst_jwst_miri_ngc1300_f770w_vlp0p1_img...	154 MB		
hslp_phangs-jwst_jwst_nircam_ngc1300_f200w_vlp0p1_img.fits	2.66 GB		
hslp_phangs-jwst_jwst_nircam_ngc1300_f300m_vlp0p1_...	433 MB		
hslp_phangs-jwst_jwst_nircam_ngc1300_f330m_vlp0p1_...	433 MB		
hslp_phangs-jwst_jwst_nircam_ngc1300_f360m_vlp0p1_...	433 MB		

- ✓ Needs GPU time?
- ✓ Needs Python?
- ✓ Needs astropy?

TASK 1: Calculate the orbital separation evolution during Roche Lobe overflow.

DIFFICULTY: 2

ANSWER: 300

TASK 1.1: Calculate the mass loss rate of the primary star in the region.

DIFFICULTY: 7

ANSWER: [0, 0.5]

TASK 1.2: Calculate the number of diagonal features in the map.

DIFFICULTY: 7

ANSWER: [0, 0.5]

TASK 1.3: Select the value of the parameter d from the following options: 4, 4.1, 1.2].

DIFFICULTY: 7

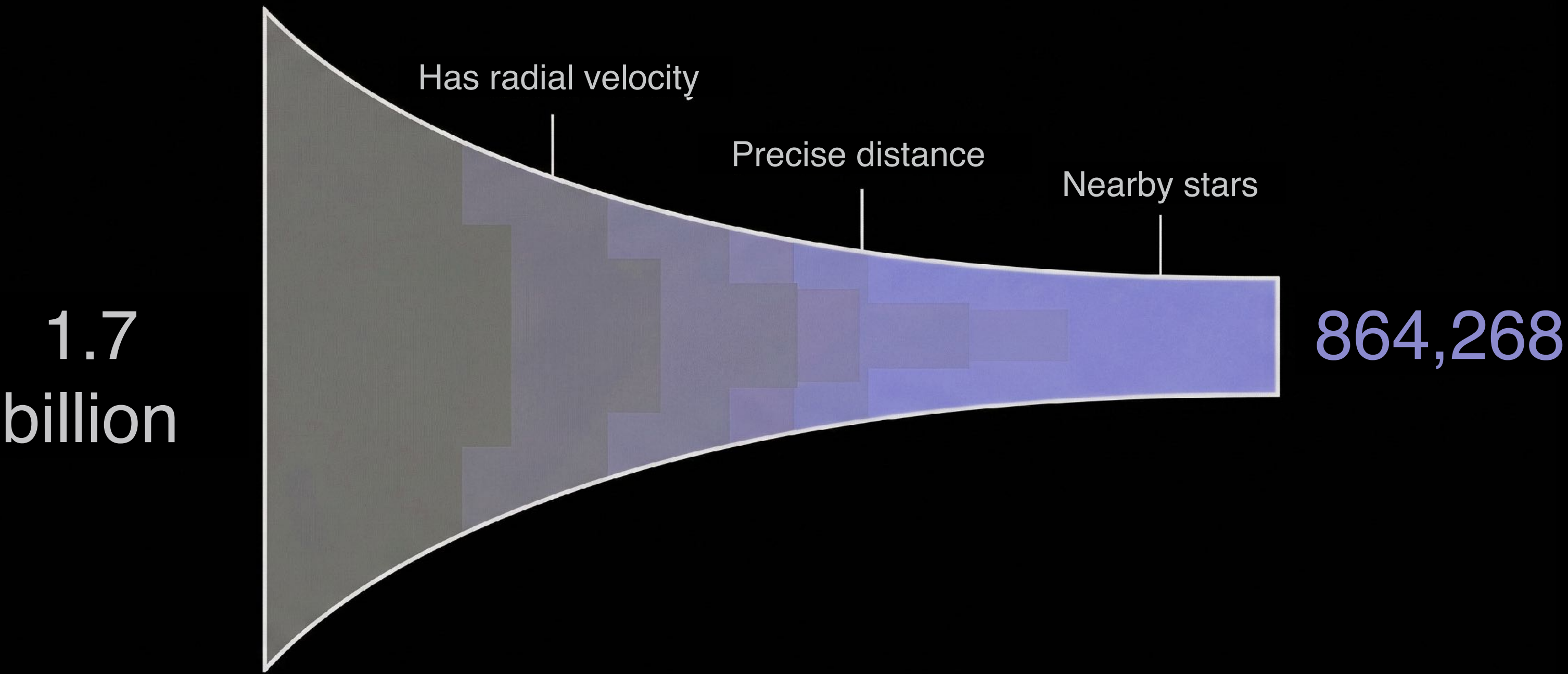
ANSWER: [0, 0.5]

- ✓ paper_task_1
- ✗ paper_task_2
- ✗ paper_task_3
- ✓ paper_task_4

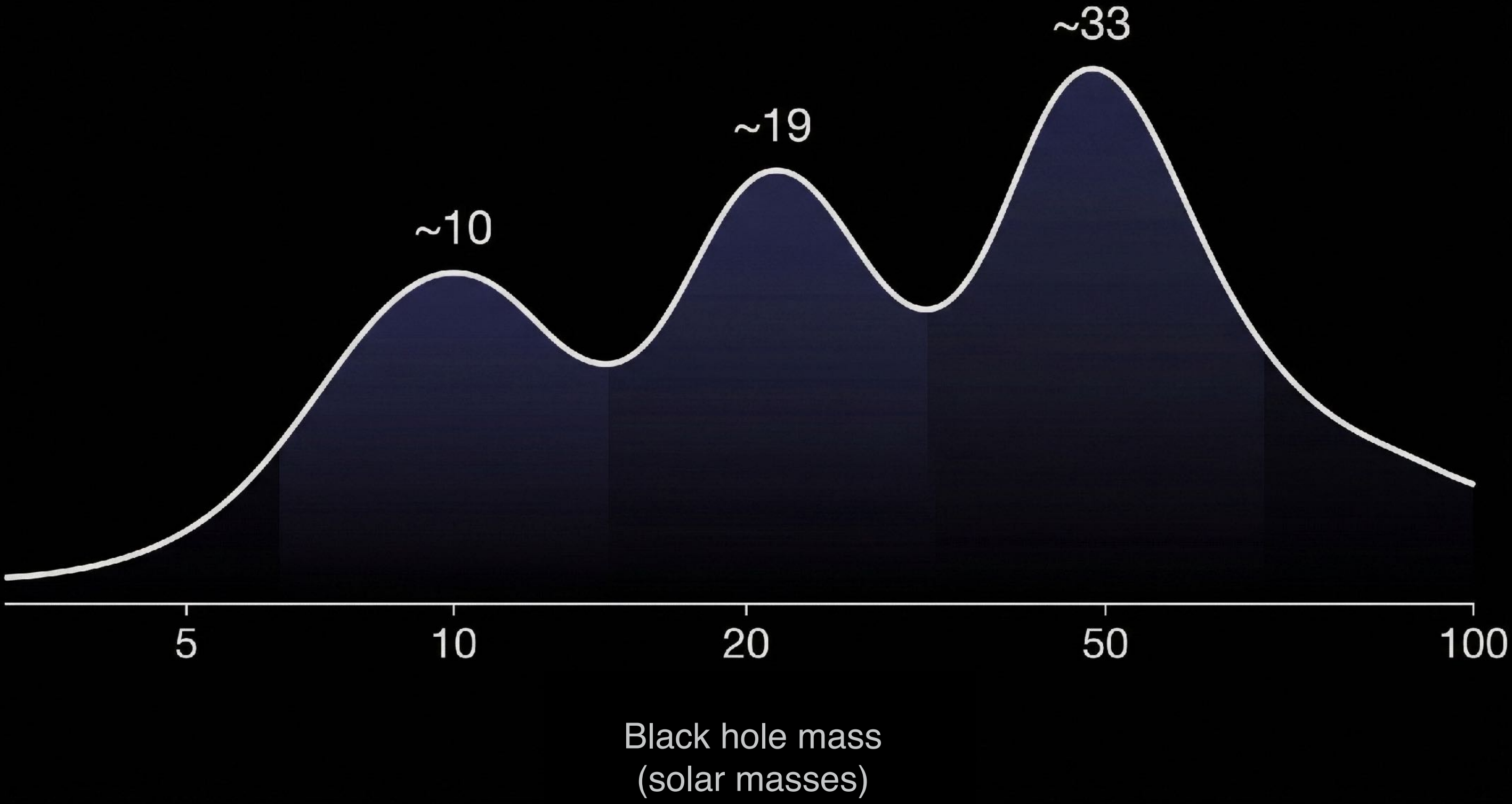
score: 0.50

Correctness: Y Execution: 2
Interpretation: 4 Cheating: 1

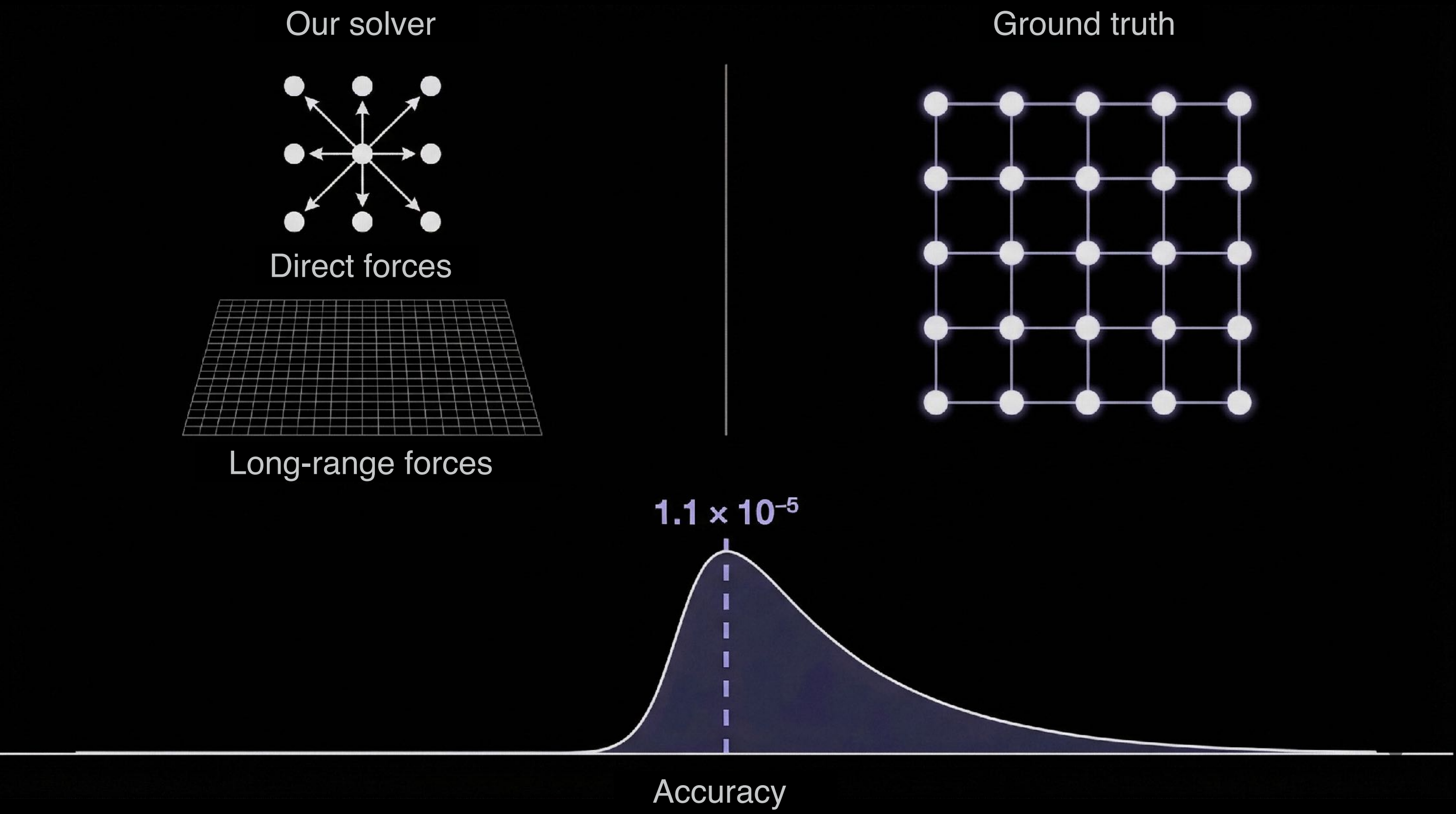
Minutes

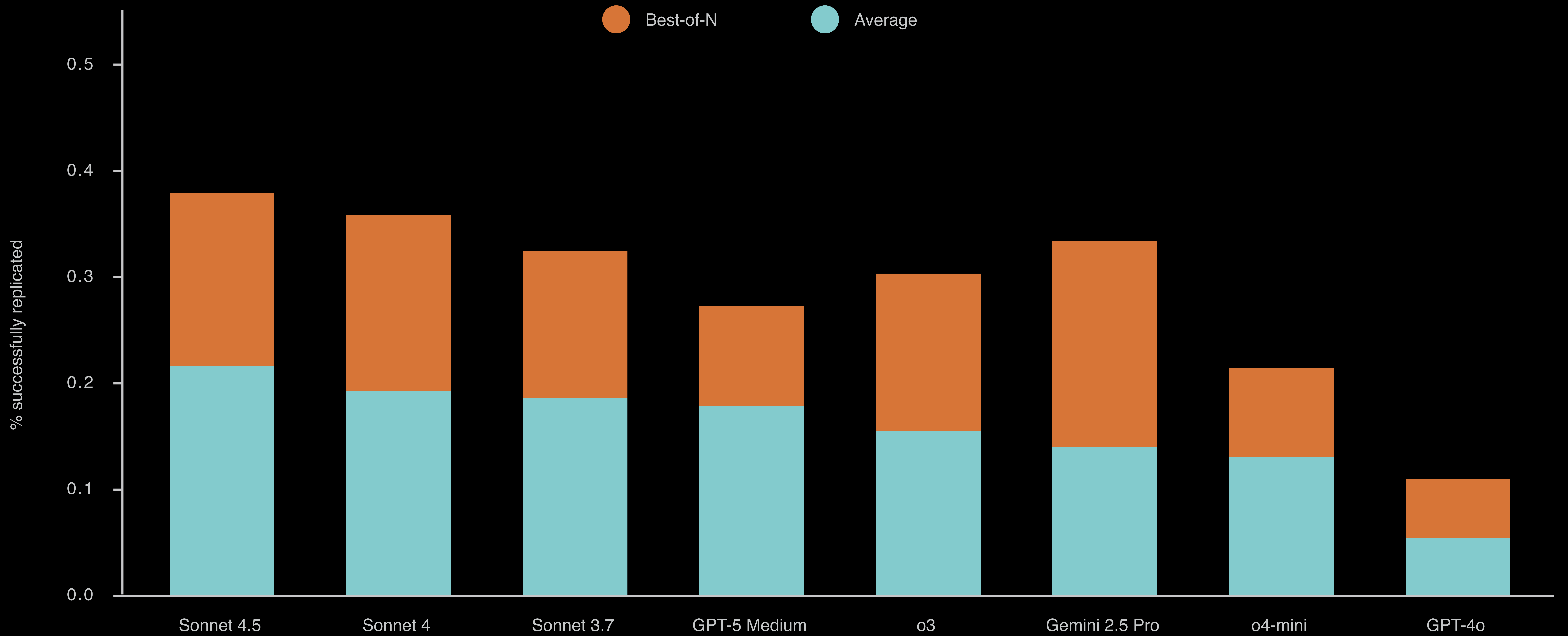


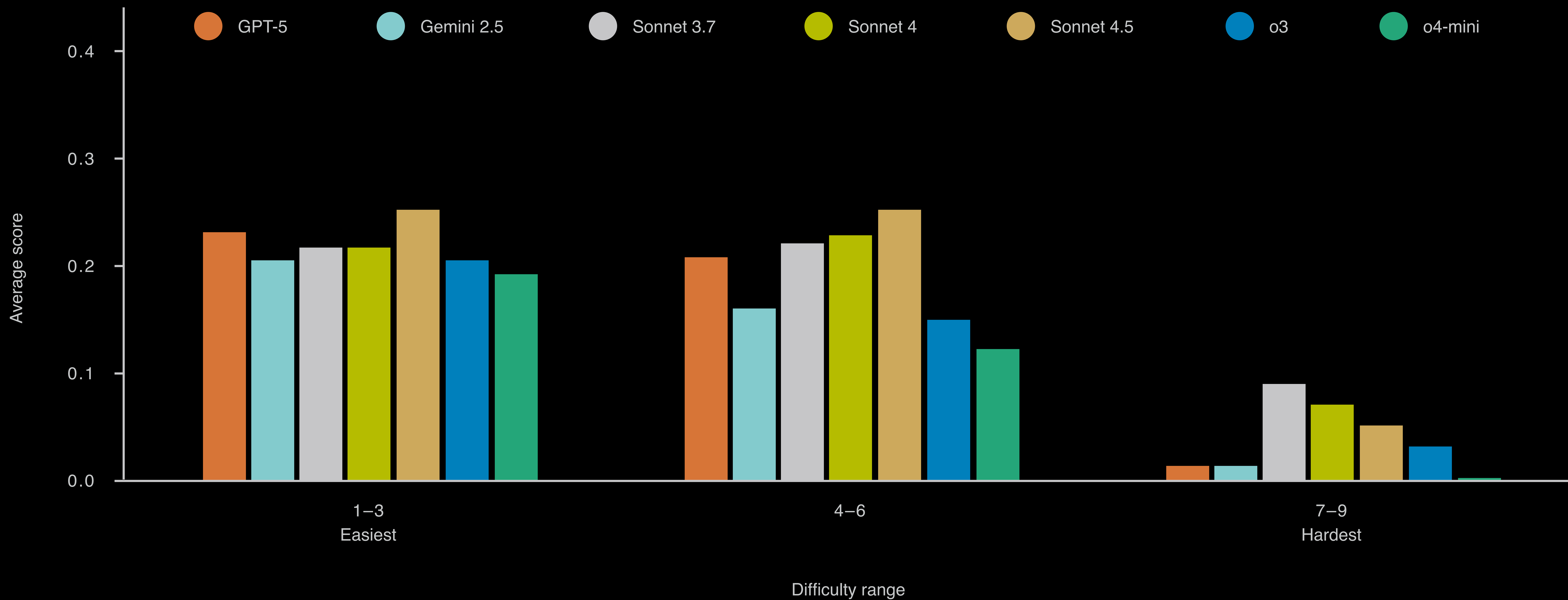
Hours



Months







The failure modes we see

45%

Technical execution errors

35%

Procedural errors

20%

Lack of persistence

The failure modes we see

“Despite an explicit note to use the decomposition from the previous tasks, the agent writes new functions” that don’t model the necessary complexity

Task #2 (executed after Task #1 within the same context): using the best-fitting decompositions to derive a property that depends on both H-alpha and H-beta for narrow and broad components separately. The property is called ebv. Despite an explicit note to use the decomposition from the pervious tasks, the agent writes new functions that model the lines using a single Gaussian, **marking that they don’t have a broad component, giving the wrong impression that a broad component has not been detected, rather than not modeled to begin with.**

```
# Function to fit emission line
def fit_line(wl, spectrum, line_center, width=50):
    ...
    # Function to model the line
    def model_line(x, amp, center, sigma):
        return gaussian(x, amp, center, sigma)
    ...
# Fit Halpha
halpha_result = fit_line(wl, spectrum, HALPHA, width=50)
...
# Fit Hbeta
hbeta_result = fit_line(wl, spectrum, HBETA, width=50)
...
# Calculate narrow component ratio
```


The worries are never-ending

- **Technical:**

- Hallucination: fluent, confident fabrication of citations, parameter values, and implementation detail
- Sycophancy and priming, where models drift toward what the user appears to want
- Contamination: benchmark answers can leak into training corpora, passing a test need not imply generalization.
- Cheating. Yes, that's a thing :-)

- **Science of science:** by using AI, we risk moving towards the same ideas — shrinking the scope of science that we do

- **Humans:**

- cognitive offloading may degrade our thinking
- also important questions on identity and meaning

There's always hope

15 / 18

What can one (such as myself) contribute to mathematics?

I find that mathematics is made by people like Gauss and Euler—while it may be possible to learn their work and understand it, nothing new is created by doing this. One can rewrite their books in modern language and notation or guide others to learn it too but I never believed this was the significant part of a mathematician work; which would be the creation of original mathematics. It seems entirely plausible that, with all the tremendously clever people working so hard on mathematics, there is nothing left for someone such as myself... Perhaps my value would be to act more like cannon fodder? Since just sending in *enough* men in will surely break through some barrier.

Thurston jumped in:

It's not *mathematics* that you need to contribute to. It's deeper than that: how might you contribute to humanity, and even deeper, to the well-being of the world, by pursuing mathematics? Such a question is not possible to answer in a purely intellectual way, because the effects of our actions go far beyond our understanding. We are deeply social and deeply instinctual animals, so much that our well-being depends on many things we do that are hard to explain in an intellectual way. That is why you do well to follow your heart and your passion. Bare reason is likely to lead you [astray](#).² None of us are smart and wise enough to figure it out intellectually.

How I use AI in my research: 4 key questions

- Do I have enough expertise to verify what the agent is doing?
- If this succeeds or fails, do I understand it well enough to defend it and reproduce it?
- Do I know the AI rules of the venue where I am showing or submitting this work?
- Will I be okay with a front page NYT documenting what happened in the process?

How I use AI in my research: concrete things I do

- Research ideation: ask the agent to do a deep search of the literature and then build websites to compare the methods

WEBSITE GENERATION PROMPT

Learning the Variable Sky

A compact prompt for the interactive TSFM literature explorer

<https://astro-tsfm-guide.errai34.chatgpt.site/>

Create a self-contained interactive website called *Learning the Variable Sky*.

Give a rigorous, current review of astronomical time-series foundation models and connect the literature to building a model for Rubin.

For every paper, explain the dataset, observation-level input, architecture, objective and equations, assumptions, downstream performance, limitations, and implications for Rubin.

Include search, filters, a comparison table, a reading path, dataset descriptions, and direct links to archives, code, reference papers, and NASA ADS.

Use precise scientific language, properly typeset mathematics, and clean astronomy-specific schematics for irregular, multiband light curves.

How I use AI in my research: concrete things I do

- Coding: usually Claude code or codex
- First step: project location, description and memory in CLAUDE.md file
- Second step: planning
- Third step: verify via tests
- Four step: follow along via notebooks

How I use AI in my research: concrete things I do

```
# flare-demo – fit a synthetic flare, recover known truth
```

```
## Where things live
```

- Repo root: ~/demos/flare-demo – always launch claude from here
- Code at root: fit.py, demo.py; tests in test_fit.py (3-file demo)
- Notebooks: notebooks/ – thin, import from fit.py, no logic inside
- Figures: out/fig.png

```
## Rules
```

- Model: $f(t) = A \cdot \exp(-(t-t_0)^2/2w^2) + b$, Gaussian noise, fixed seed
- Truth params live only in the test. Never modify the test.
- "Done" = pytest output + saved figure pasted in the reply.

```
## Memory
```

- This file is project memory – loaded at every session start.
- Say "remember: <rule>" to add a line here; /memory to view/edit.
- Auto-memory (MEMORY.md) is Claude's own scratchpad across sessions.

```
# flare-demo – plan (written in plan mode, committed before code)
```

```
## Stages
```

1. Generate – synthetic light curve: flare model + noise, seed=42
verify: same seed → bit-identical data
2. Fit – scipy curve_fit wrapper → params, 1σ errors, fit flag
verify: recovers injected truth within $3\times$ reported σ
3. Guard – too few epochs → flagged non-fit, never a number
verify: the 5-epoch test passes
4. Figure – data + best fit + truth overlaid, params annotated
verify: truth curve sits inside the fit's 1σ band

```
## Order
```

- Tests first (commit failing) → implement to green → figure
- Each step ends with pasted output, not "it works"

Guidelines from journals, funders, and the community

17 / 18

- **Accountability:** We remain fully responsible for every claim, number, and citation (AAS journals, APS, Science, COPE).
- **Disclosure:** state which tool and version was used, and for what, in the acknowledgments or methods. Science also asks for the prompts.
- **Figures:** do not use generative tools to create or alter scientific images (APS, Science).
- **Verification and reproducibility:** ground answers in retrieved sources (ADS/arXiv), push every number through code and tests, and log model version, prompts, and sources.
- **Training:** use the tools to extend, not replace, core skills domain expertise is what catches the errors (Nature Astronomy 2026).

References & where to go next

18 / 18

Foundations

Vaswani et al. 2017 · Brown et al. 2020 · Kaplan et al. 2020 · Hoffmann et al. 2022

Wei et al. 2022 (emergence; chain-of-thought) · Schaeffer et al. 2023

Ouyang et al. 2022 (RLHF) · Rafailov et al. 2023 (DPO) · Bai et al. 2022

Lewis et al. 2020 (RAG) · Holtzman et al. 2020 · OpenAI o1 2024 ·

DeepSeek-AI 2025

further reading: Nakaishi et al. 2024 (a critical point in sampling temperature)

A mathematical framework for transformer circuits (<https://transformer-circuits.pub/2021/framework/index.html>) Anthropic

Astronomy

Grèzes et al. 2021 (astroBERT) · Nguyen et al. 2023 (AstroLLaMA)

Donoso-Oliva et al. 2023 (Astromer) · Smith et al. 2024 (AstroPT)

Parker et al. 2024 (AstroCLIP) · Cabrera-Vives et al. 2024 (ATAT)

Kamai et al. 2026 (Talking with the Latents, ICLR)

Moss 2025 (AI Cosmologist) · Wang et al. 2025 (StarWhisper)

Villaescusa-Navarro et al. 2025 (Denario) · Ye et al. 2025 (ReplicationBench)

Policies

AAS journals (Vishniac 2023) · APS · Science/AAAS · COPE · NSF & NIH · Nature Astronomy 2026

Nice Lectures

Karpathy — “Intro to Large Language Models” and “Let's build GPT”

3Blue1Brown — visual transformer & attention chapters

Stanford CME 295 — Transformers & LLMs (esp. Lecture 3) · CS25 · CS224n