AstroMLab
Expediting Discoveries in Astronomy with LLM

Yuan-Sen Ting (丁源森)

The Ohio State University

NSF awarded over $200 million for *AI Research Institutes*

Biological Sciences

~ 2 centers

Physical Sciences

~ 3 centers

Environmental Sciences

~ 2 centers

NSF awarded over $200 million for *AI Research Institutes*

Biological Sciences

~ 2 centers

Physical Sciences

~ 3 centers

Environmental Sciences

~ 2 centers

6 centers x
15M ~ 100M

Hype, myth, or *real deal*?

Article | Open access | Published: 15 July 2021

Highly accurate protein structure prediction with AlphaFold

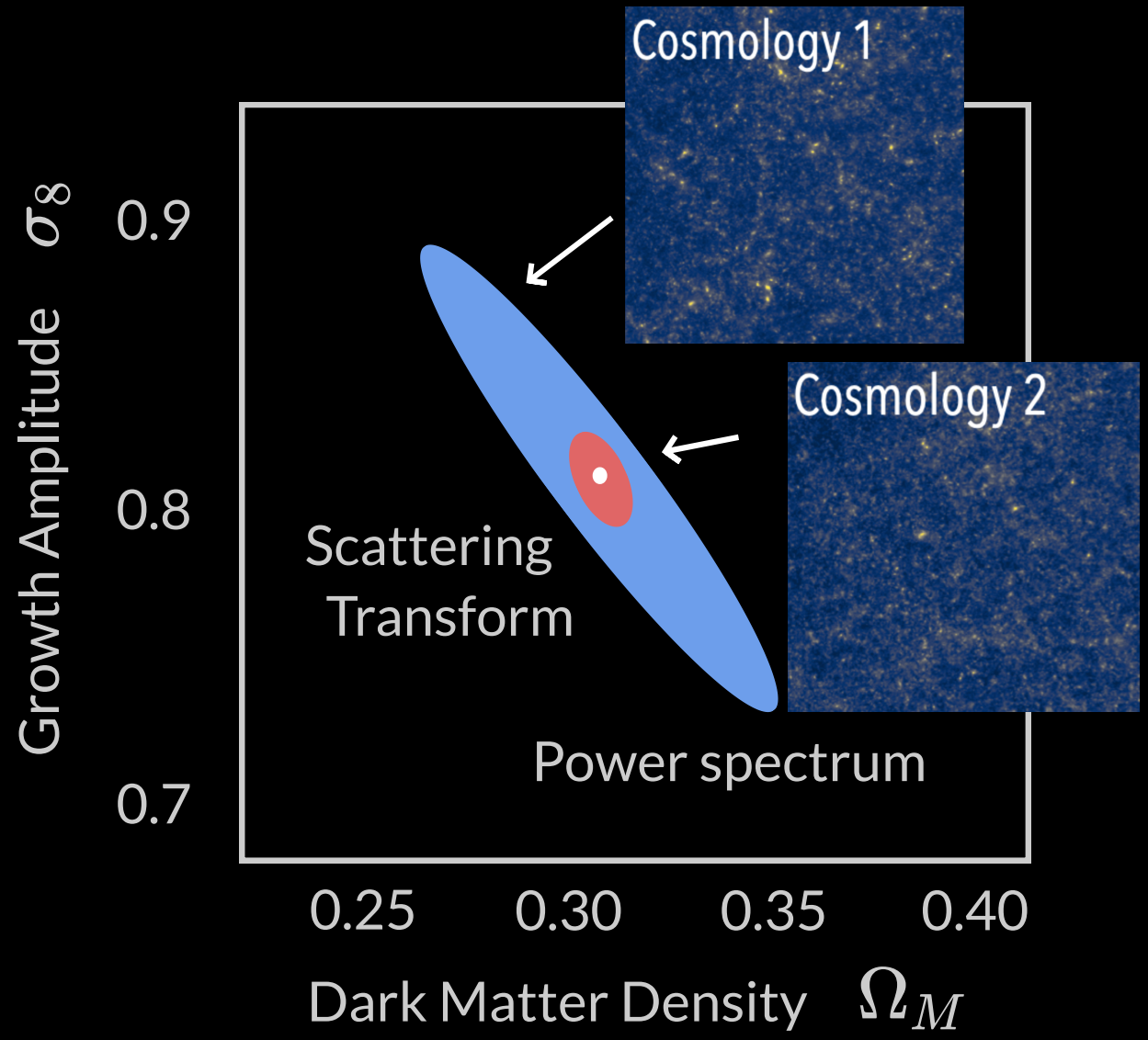
John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, ... Demis Hassabis

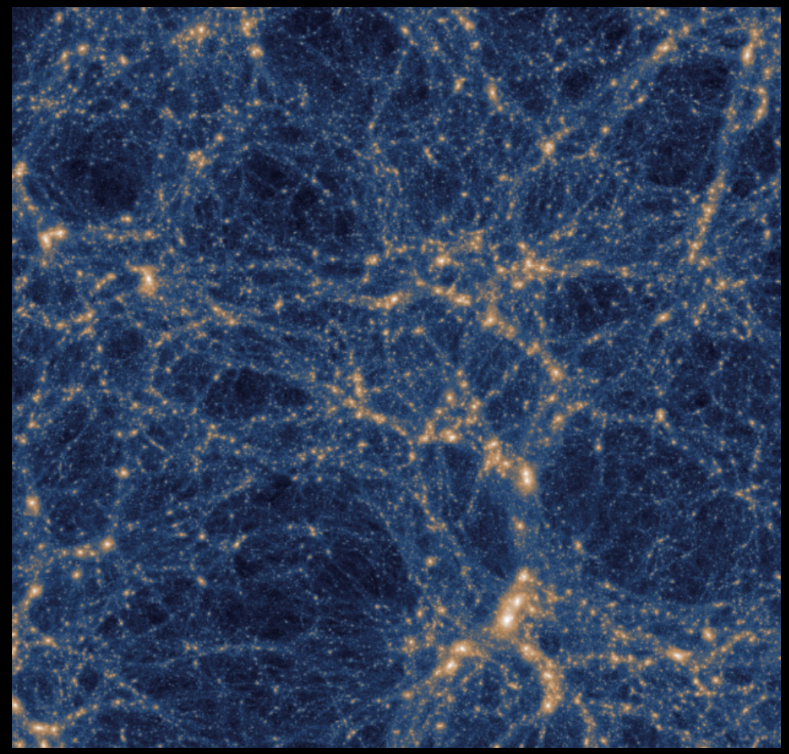
+ Show authors

Nature **596**, 583–589 (2021) | Cite this article

Why hasn't astronomy had its
"AlphaFold" moment yet?"

Most AI in Astronomy focuses on *extending* statistical methods

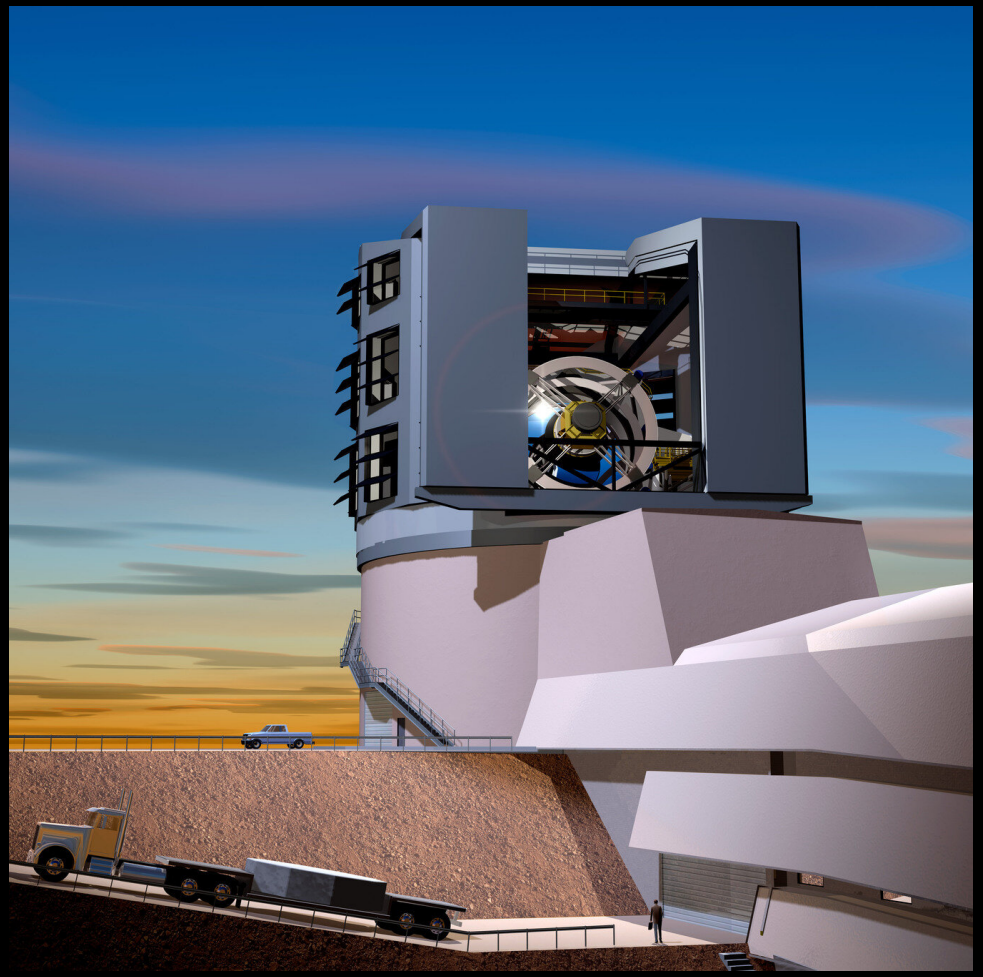




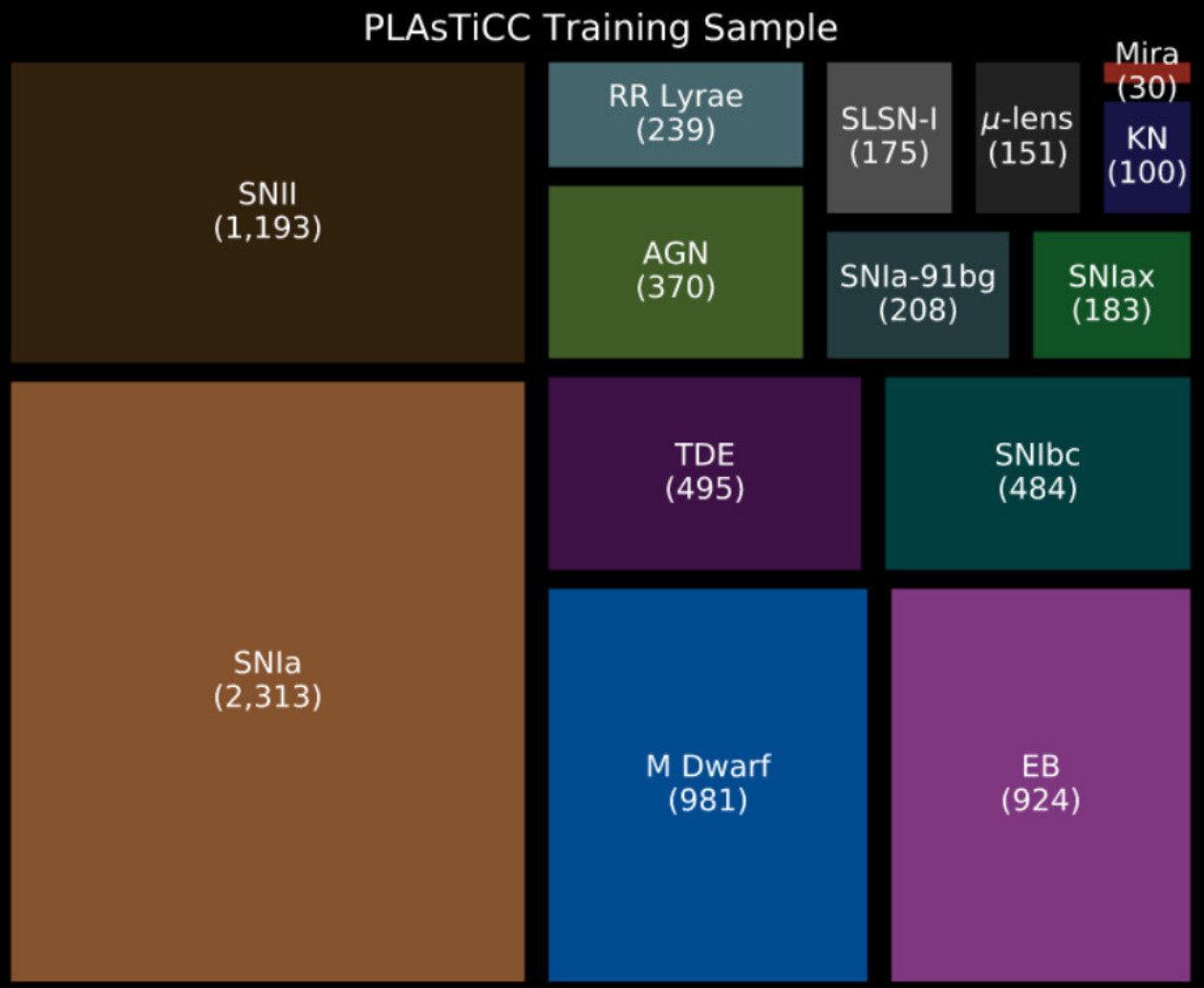
International Astrostatistics Association Award

Cheng, YST, Menard & Bruna + 20

or building effective brokers / *classifiers*



Rubin Observatory



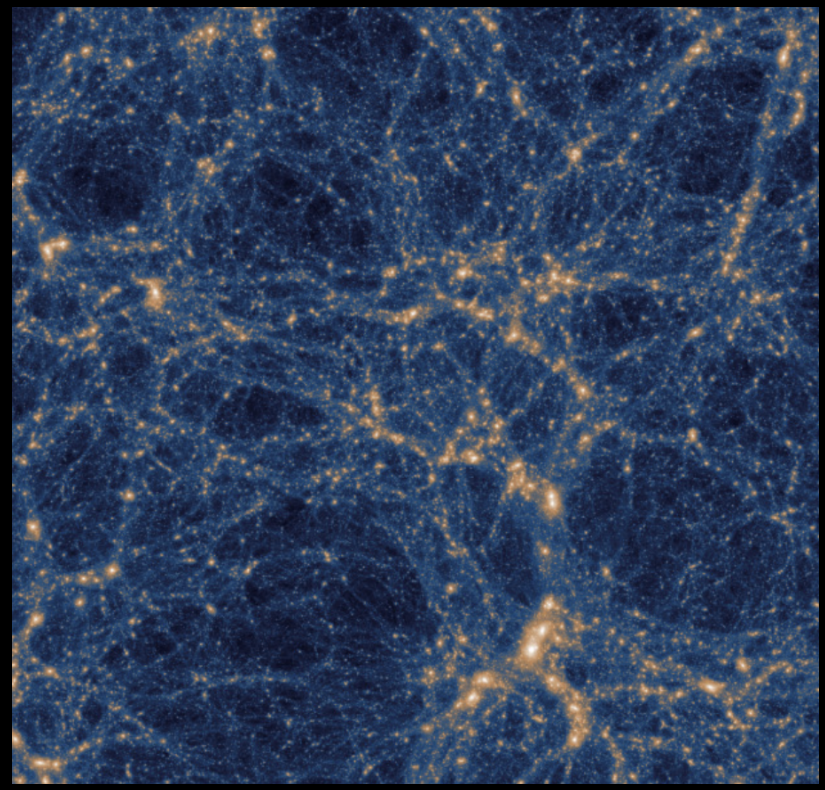
Improving *individual* downstream tasks with annotated data in a confined setting will *not* revolutionize astronomy

The *complexity* of astronomy is *too low* for AI

Highly non-Gaussian

My niece

Weakly non-Gaussian



Cosmic large-scale structure

Astronomy is *not* biology

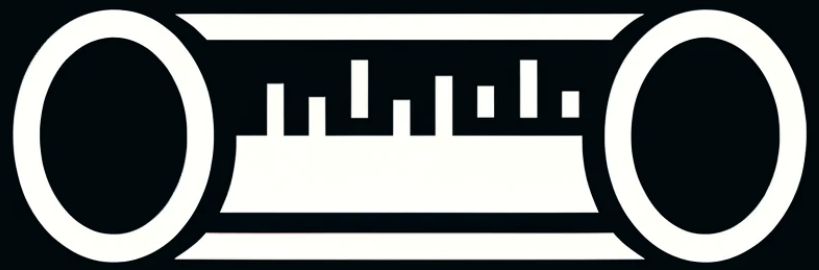
Data / Observation

Theory / Hypothesis

Astronomy is *not* biology

Data / Observation

Theory / Hypothesis

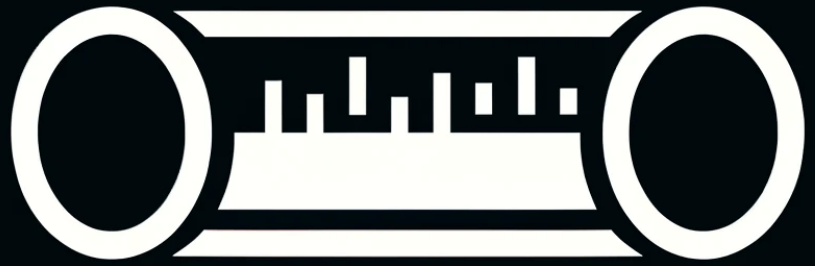


Analysis Pipelines

Astronomy is *not* biology

Data / Observation

Theory / Hypothesis



Analysis Pipelines

True

False

Biology faced a fundamental *bottleneck*

Data / Observation

Theory / Hypothesis

Analysis Pipelines

True

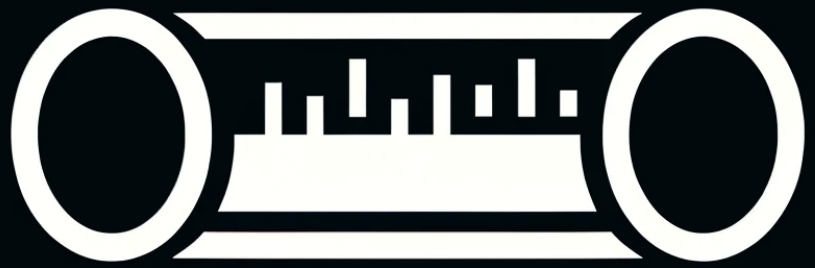
False

Biology faced a fundamental *bottleneck*

Data / Observation

Theory / Hypothesis

Alphafold



Analysis Pipelines

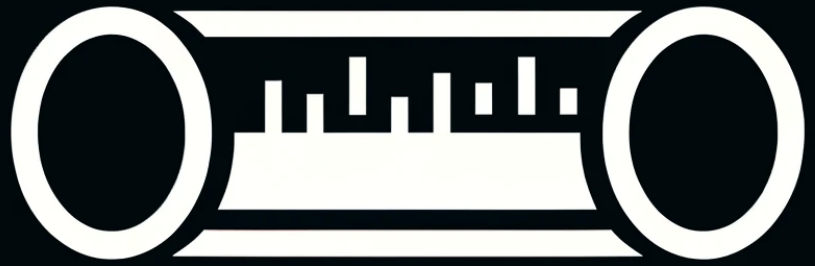
True

False

Astronomy already has a successful *standard model*

Data / Observation

Theory / Hypothesis



LamdaCDM

True

False

The Bitter Lesson - Rich Sutton, 2019

The Bitter Lesson - Rich Sutton, 2019

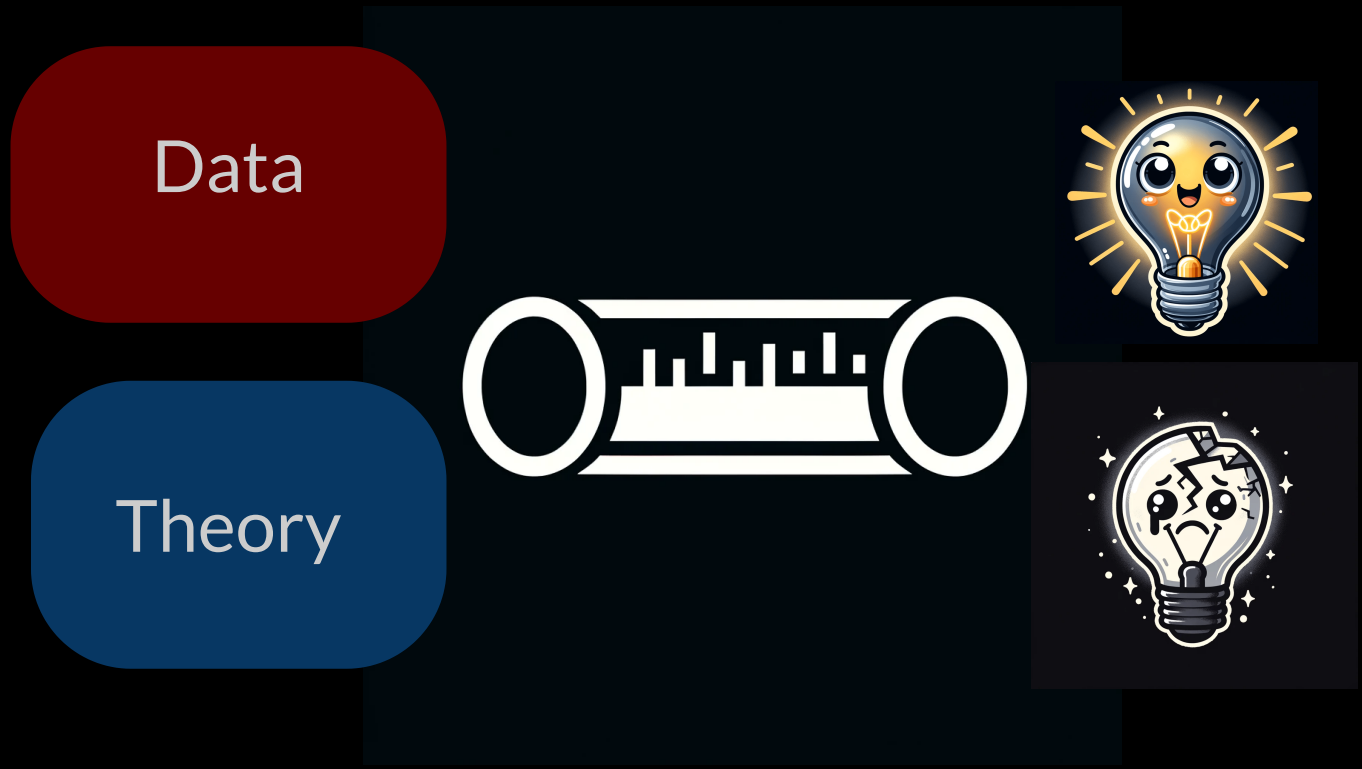
"We should build in only the *meta-methods* that can find and capture this arbitrary complexity.

Essential to these methods is that they can find good approximations, but the search for them should be *by our methods, not by us*.

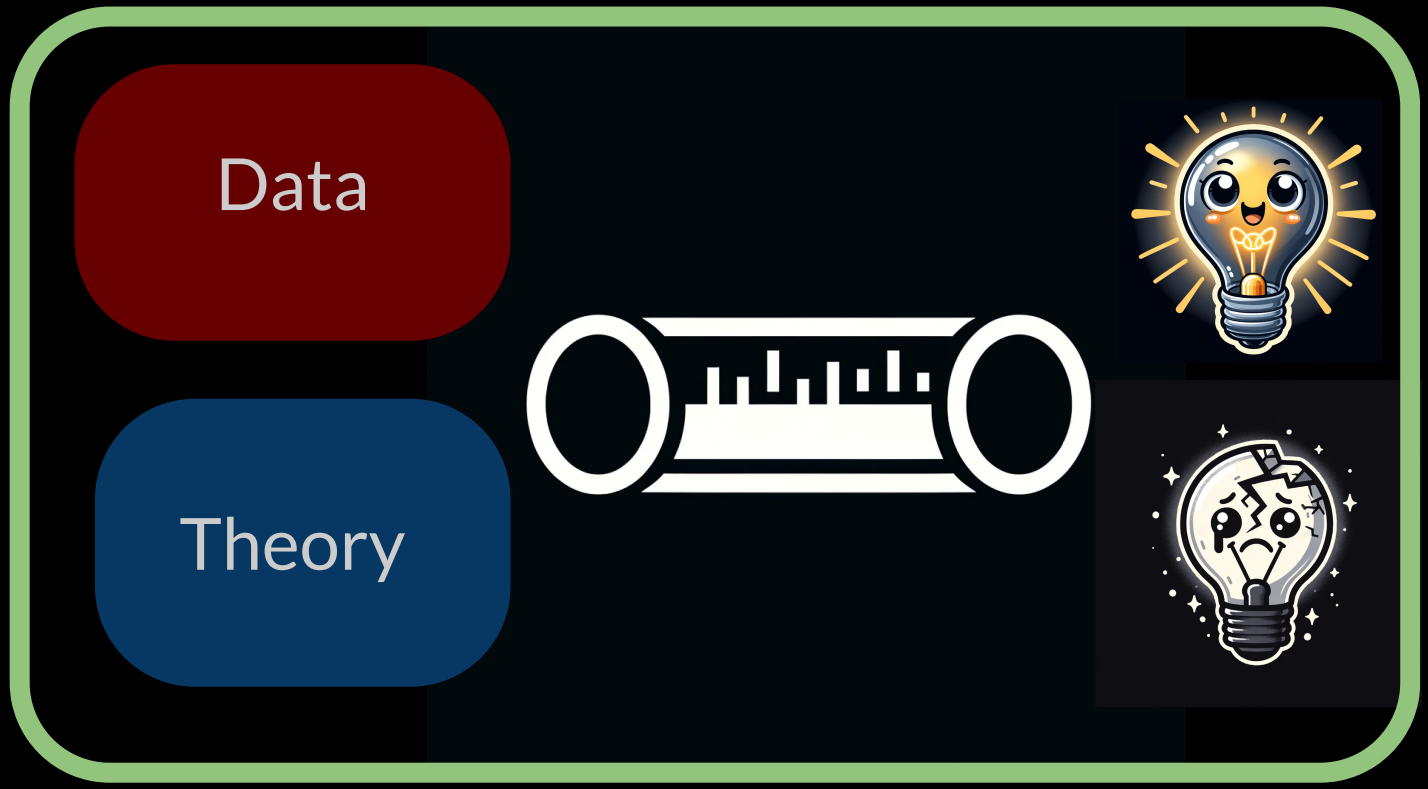
We want AI agents that can *discover like we can*, not which contain what we have discovered."

Toward an *AI Astronomer*

Research is essentially a *reinforcement learning* process



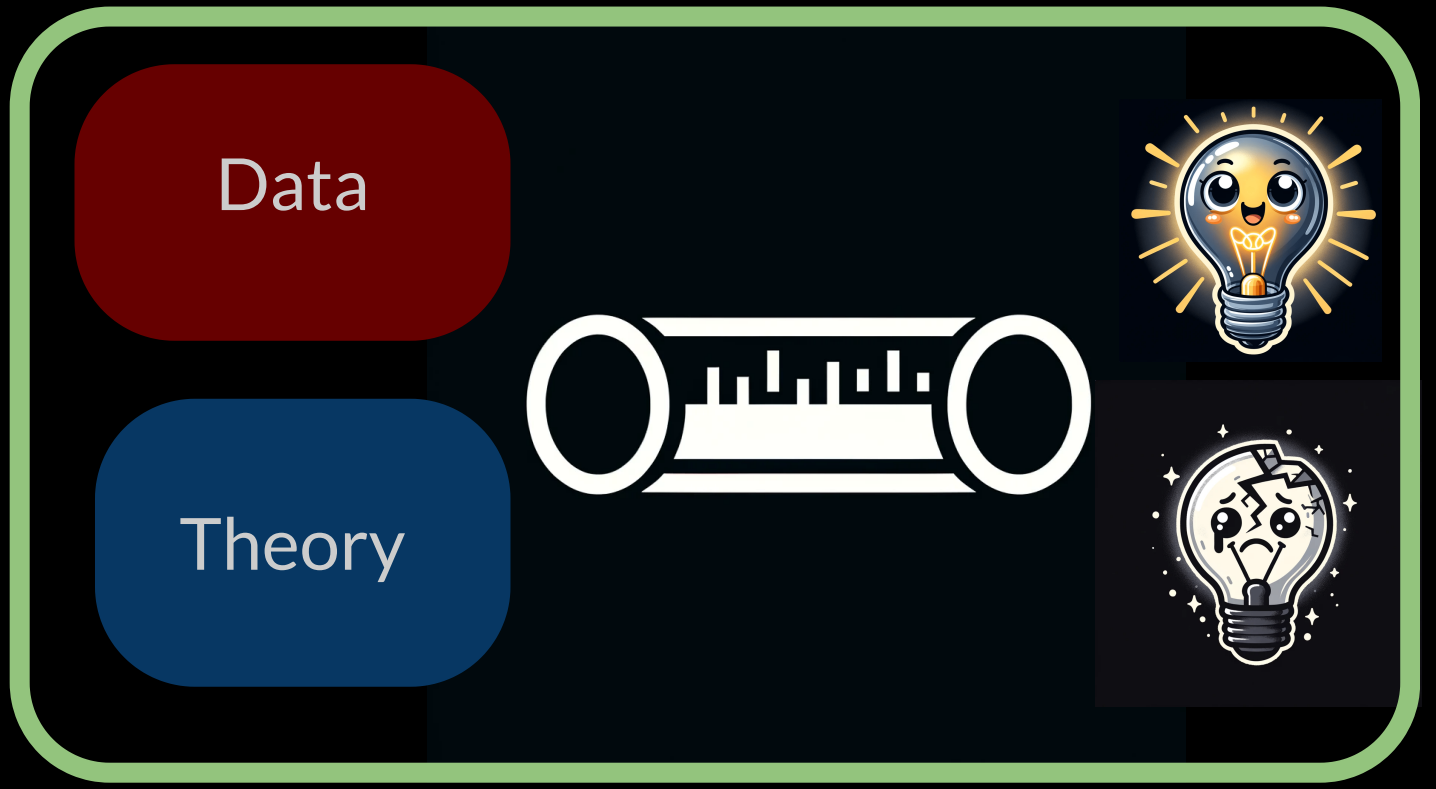
Research is essentially a *reinforcement learning* process



State of the research

Research is essentially a *reinforcement learning* process

Evaluate the state

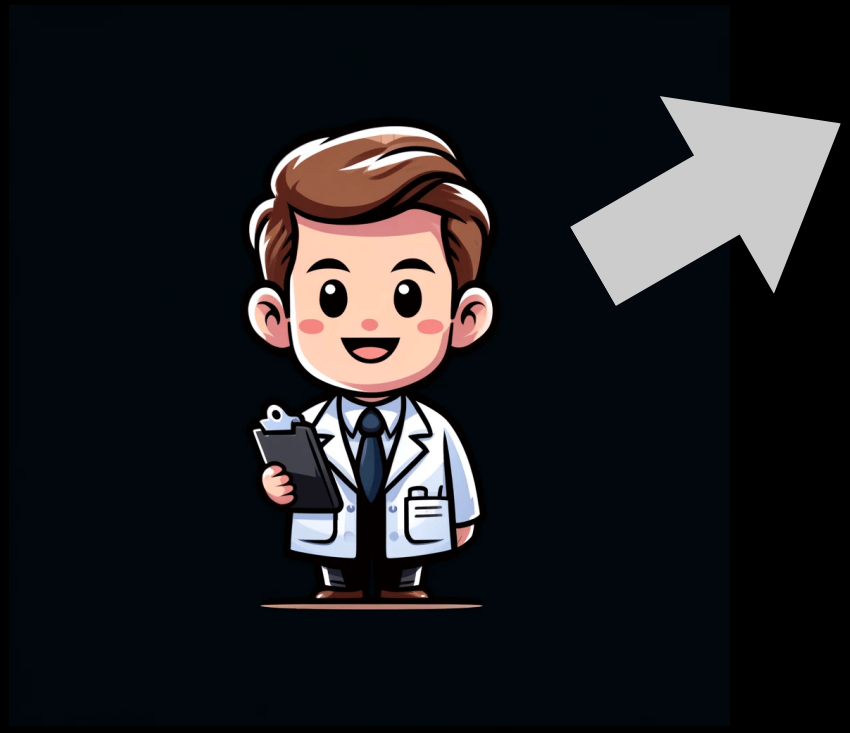


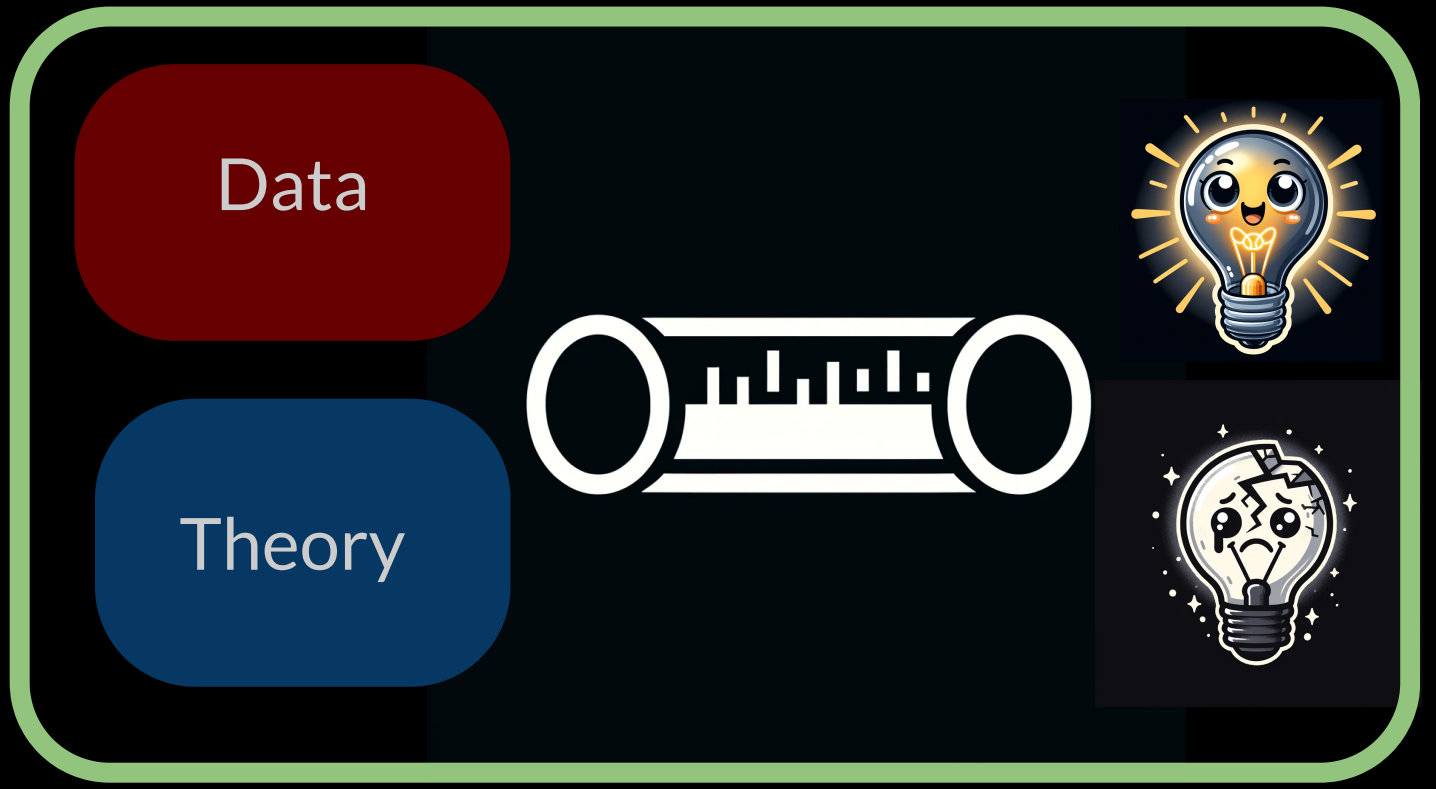
State of the research

Research is essentially a *reinforcement learning* process

Evaluate the state

Making "plan"



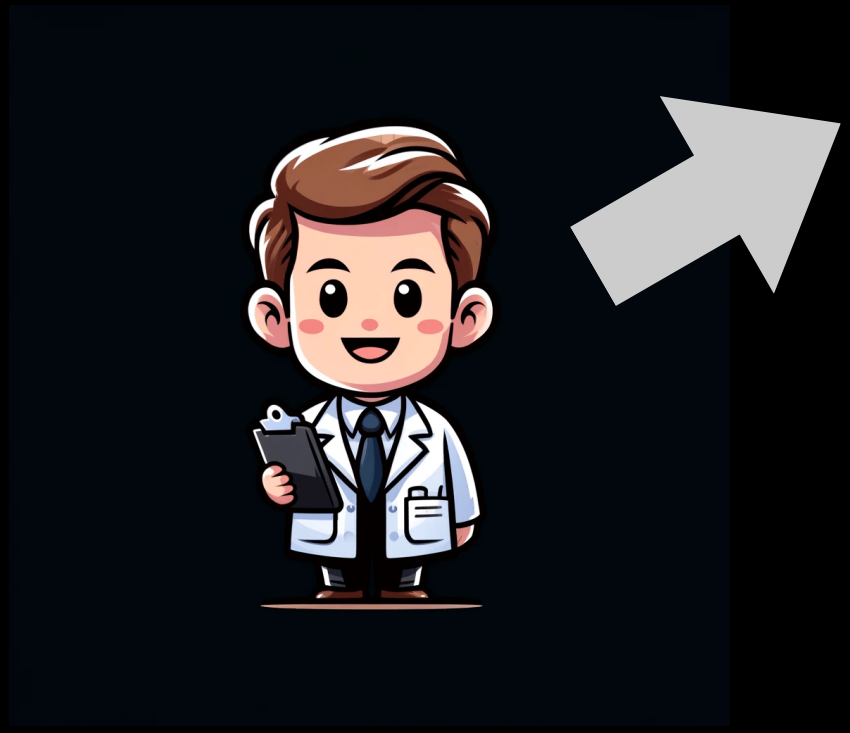


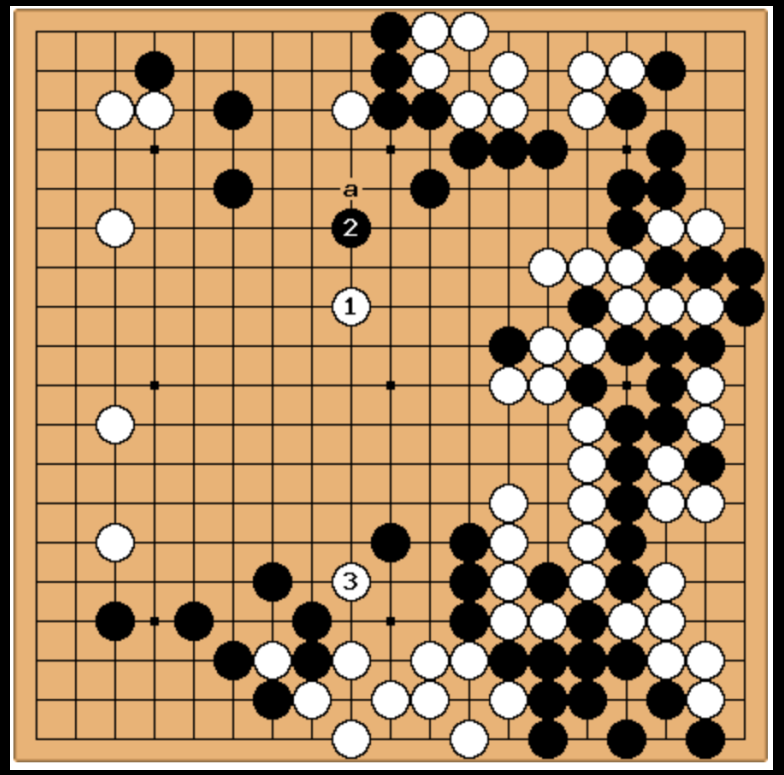
State of the research

Research is essentially a *reinforcement learning* process

Evaluate the state

Making "plan"

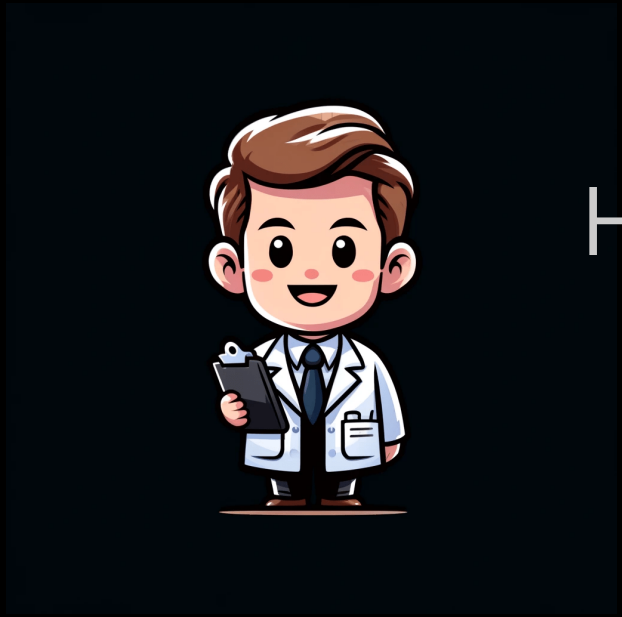




State of the research

Enabling LLM agents to
learn how to make plan
through *open world exploration*

Enabling LLM agents to
learn how to make plan
through *open world exploration*



Human *"intuition"* +
experience

Interpreting Multi-band Galaxy Observations with Large Language Model-Based Agents

Zechang Sun
Department of Astronomy
Tsinghua University
Beijing, China
szc22@mails.tsinghua.edu.cn

Yuan-Sen Ting
Department of Astronomy
The Ohio State University
Columbus, OH 43210, USA
ting.74@osu.edu

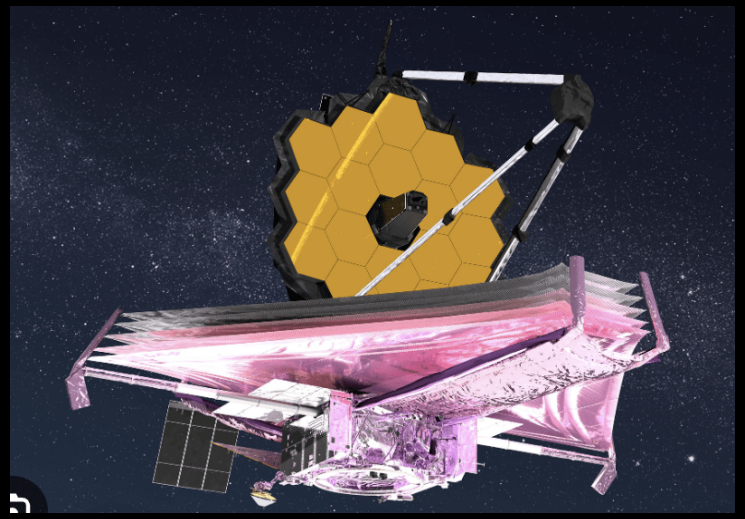
Yaobo Liang
Microsoft Research Asia
Beijing, China
yaobo.liang@microsoft.com

Nan Duan
Microsoft Research Asia
Beijing, China
nanduan@microsoft.com

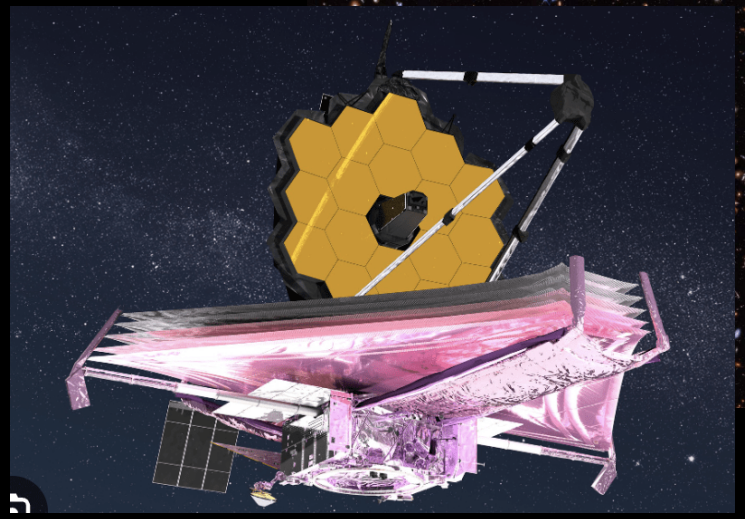
Song Huang
Department of Astronomy
Tsinghua University
Beijing, China
shuang@mail.tsinghua.edu.cn

Zheng Cai
Department of Astronomy
Tsinghua University
Beijing, China
zcai@mail.tsinghua.edu.c

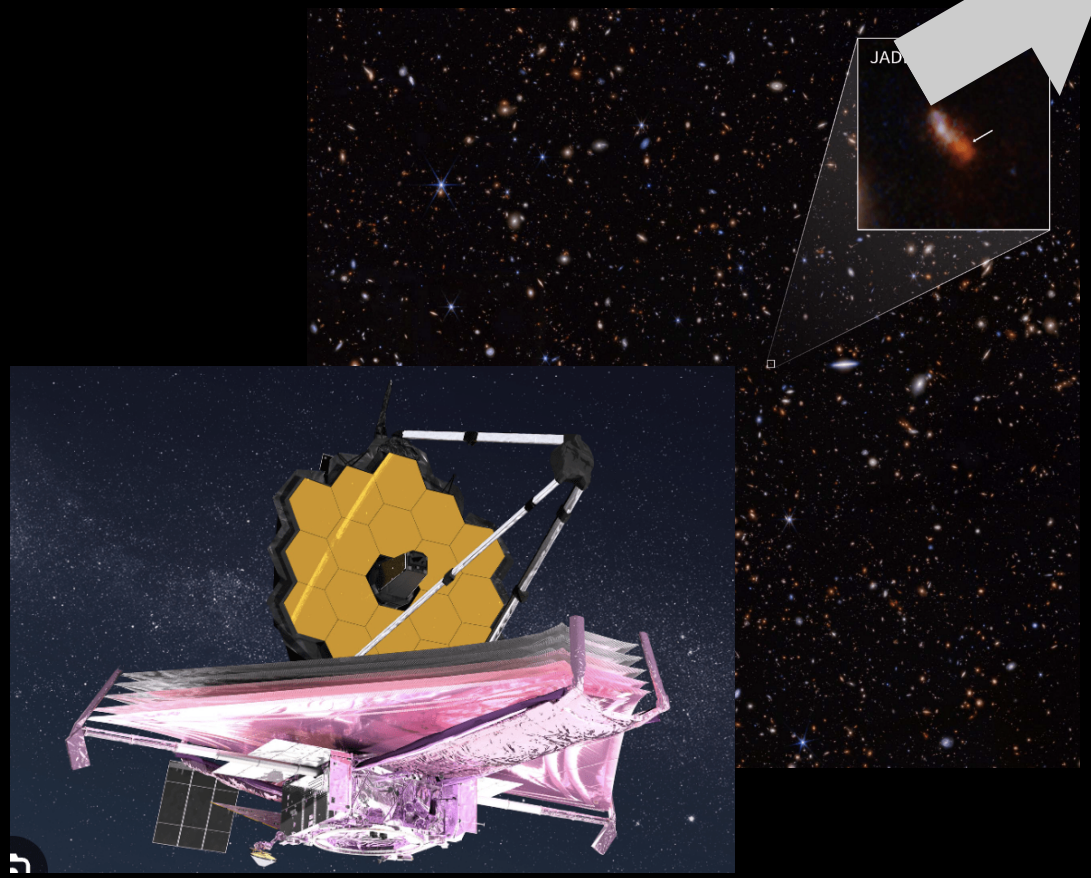
Can we identify all the astronomical objects that our current physics can't explain



Can we identify all the astronomical objects that our current physics can't explain



Can we identify all the astronomical objects that our current physics can't explain ??



The Astrophysical Journal, 964:71 (35pp), 2024 March 20 https://doi.org/10.3847/1538-4357/ad1ee4

The Cosmos in Its Infancy: JADES Galaxy Candidates at $z > 8$ in GOODS-S and GOODS-N

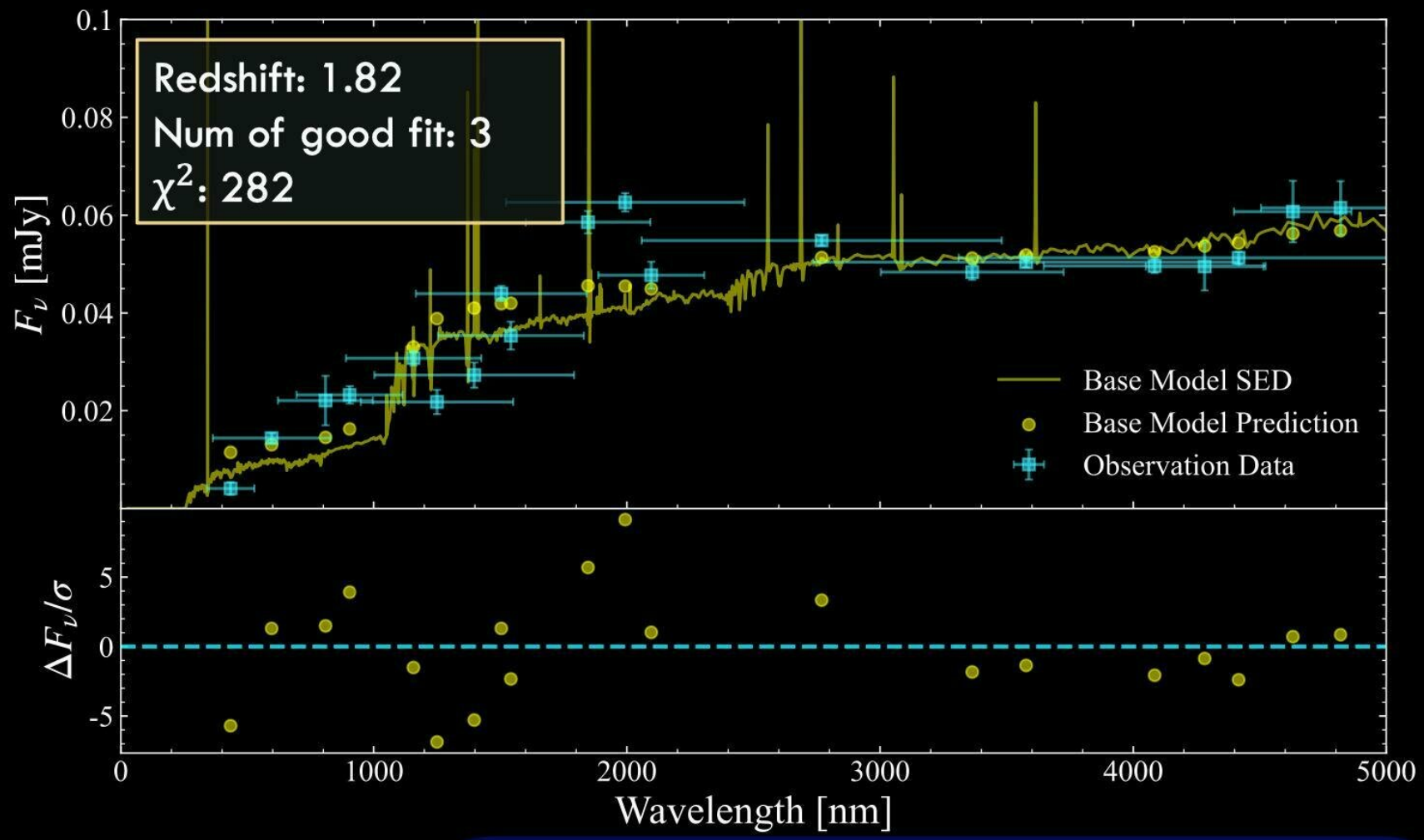
Kevin N. Hainline[1], Benjamin D. Johnson[2], Brant Robertson[3], Sandro Tacchella[4,5], Jakob M. Helton[1], Fengwu Sun[1], Daniel J. Eisenstein[2], Charlotte Simmonds[4,5], Michael W. Topping[1], Lily Whitler[1], Christopher N. A. Willmer[1], Marcia Rieke[1], Katherine A. Suess[6,7], Raphael E. Hviding[1], Alex J. Cameron[8], Stacey Alberts[1], William M. Baker[4,5], Stefi Baum[9], Rachana Bhatawdekar[10,11], Nina Bonaventura[1,12,13], Kristan Boyett[14,15], Andrew J. Bunker[8], Stefano Carniani[16], Stephane Charlot[17], Jacopo Chevallard[8], Zuyi Chen[1], Mirko Curti[4,5,18], Emma Curtis-Lake[19], Francesco D'Eugenio[4,5], Eiichi Egami[1], Ryan Endsley[20], Ryan Hausen[21], Zhiyuan Ji[1], Tobias J. Looser[4,5], Jianwei Lyu[1], Roberto Maiolino[4,5,22], Erica Nelson[23], Dávid Puskás[4,5], Tim Rawle[10], Lester Sandles[4,5], Aayush Saxena[8,22], Renske Smit[24], Daniel P. Stark[1], Christina C. Williams[25], Chris Willott[26], and Joris Witstok[4,5]

[1] Steward Observatory, University of Arizona, 933 N. Cherry Avenue, Tucson, AZ 85721, USA
[2] Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA
[3] Department of Astronomy and Astrophysics, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 96054, USA
[4] Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge, CB3 0HA, UK
[5] Cavendish Laboratory, University of Cambridge, 19 JJ Thomson Avenue, Cambridge, CB3 0HE, UK
[6] Department of Astronomy and Astrophysics, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA
[7] Kavli Institute for Particle Astrophysics and Cosmology and Department of Physics, Stanford University, Stanford, CA 94305, USA
[8] Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford, OX1 3RH, UK
[9] Department of Physics and Astronomy, University of Manitoba, Winnipeg, MB, R3T 2N2, Canada
[10] European Space Agency (ESA), European Space Astronomy Centre (ESAC), Camino Bajo del Castillo s/n, 28692 Villanueva de la Cañada, Madrid, Spain
[11] European Space Agency, ESA/ESTEC, Keplerlaan 1, 2201 AZ Noordwijk, The Netherlands
[12] Cosmic Dawn Center (DAWN), Copenhagen, Denmark
[13] Niels Bohr Institute, University of Copenhagen, Jagtvej 128, DK-2200, Copenhagen, Denmark
[14] School of Physics, University of Melbourne, Parkville, VIC 3010, Australia

Astronomical research is *more than* just fitting data

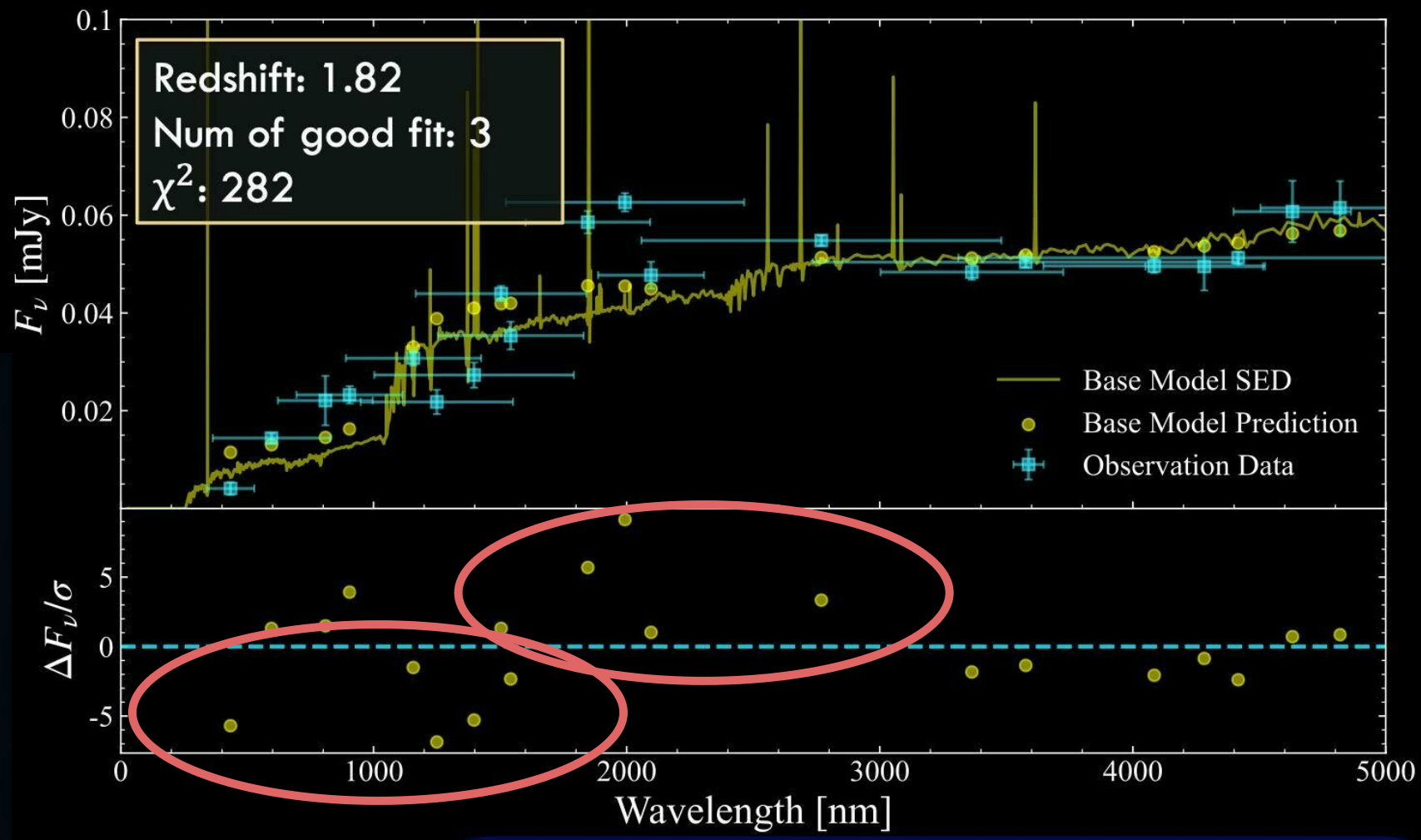
Action space is vast and transcends mathematical formalism

A default fit with an SED model

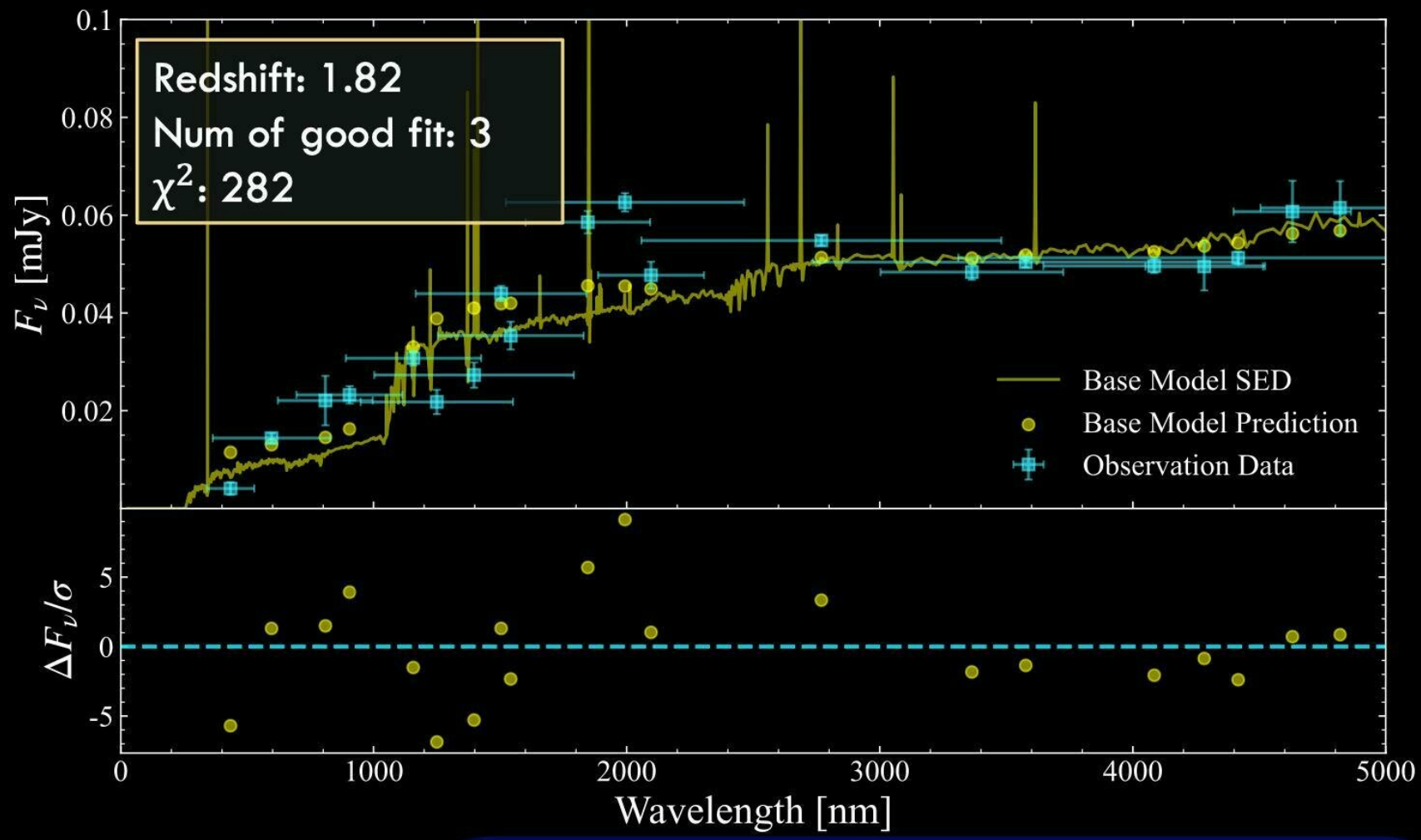


Action space is vast and transcends mathematical formalism

A default fit with an SED model

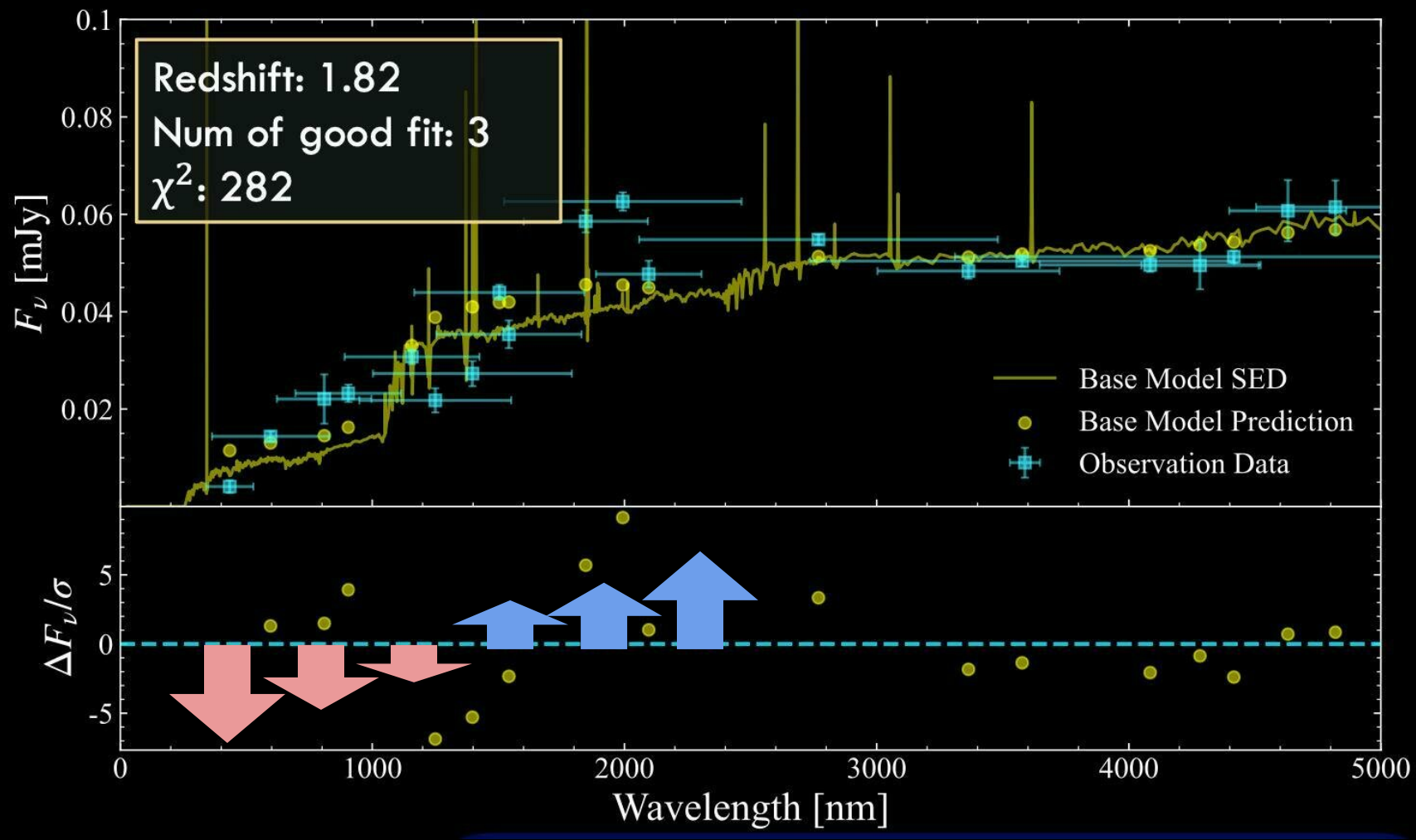


Action space is vast and transcends mathematical formalism



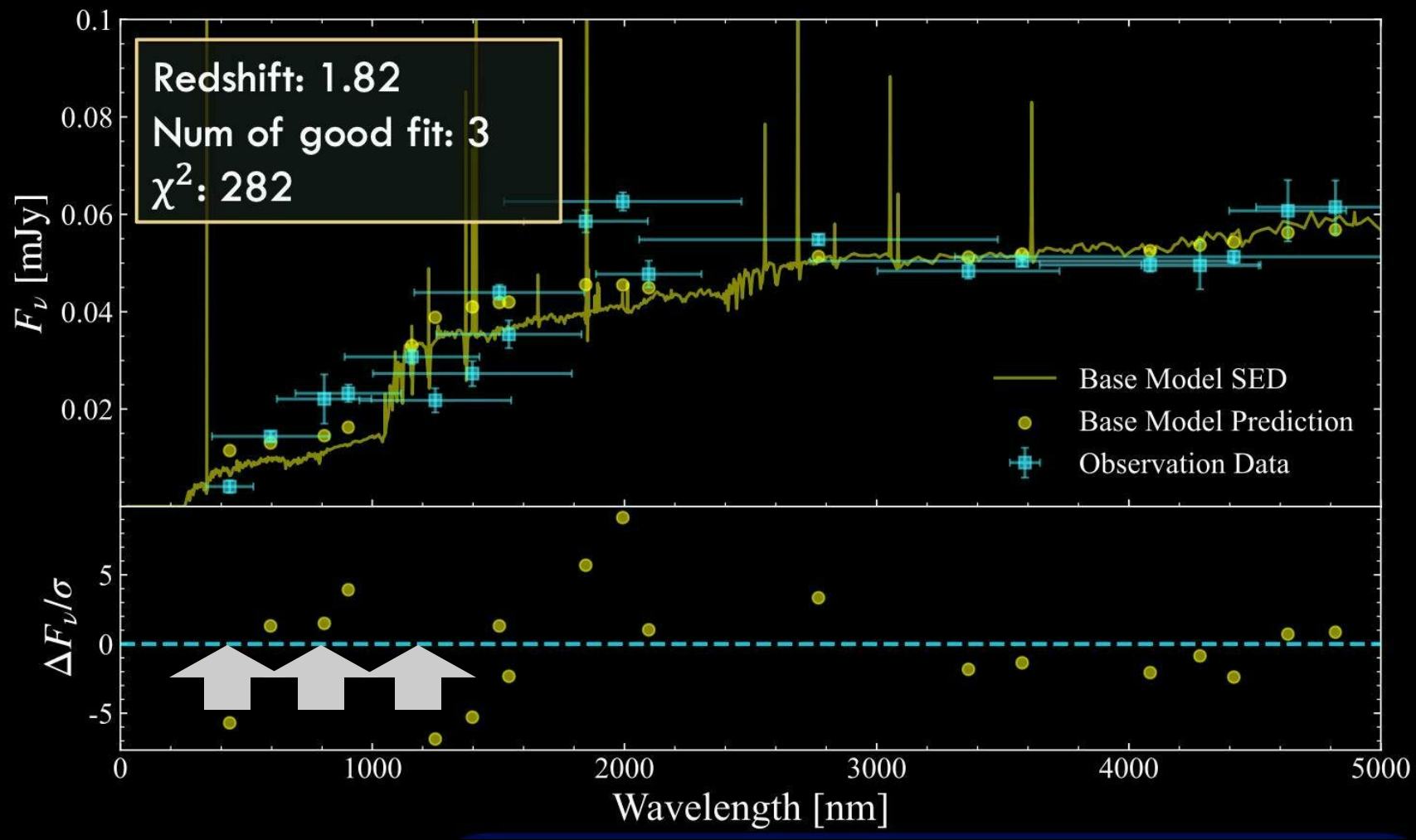
Action space is vast and transcends mathematical formalism

Extinction model ?



Action space is vast and transcends mathematical formalism

Young stellar population?



Many real-world projects lack a *mathematical* reward function

Many real-world projects lack a *mathematical* reward function

- The objective goes beyond minimizing a single error metric.

Many real-world projects lack a *mathematical* reward function

- The objective goes beyond minimizing a single error metric.
- Real-world action spaces are vast and hard to parameterize.

Many real-world projects lack a *mathematical* reward function

- The objective goes beyond minimizing a single error metric.
- Real-world action spaces are vast and hard to parameterize.
- Many tasks may require modifying assumptions / physical models, not just optimizing over all parameters

Can a large-language model learn
from *its own experience*?

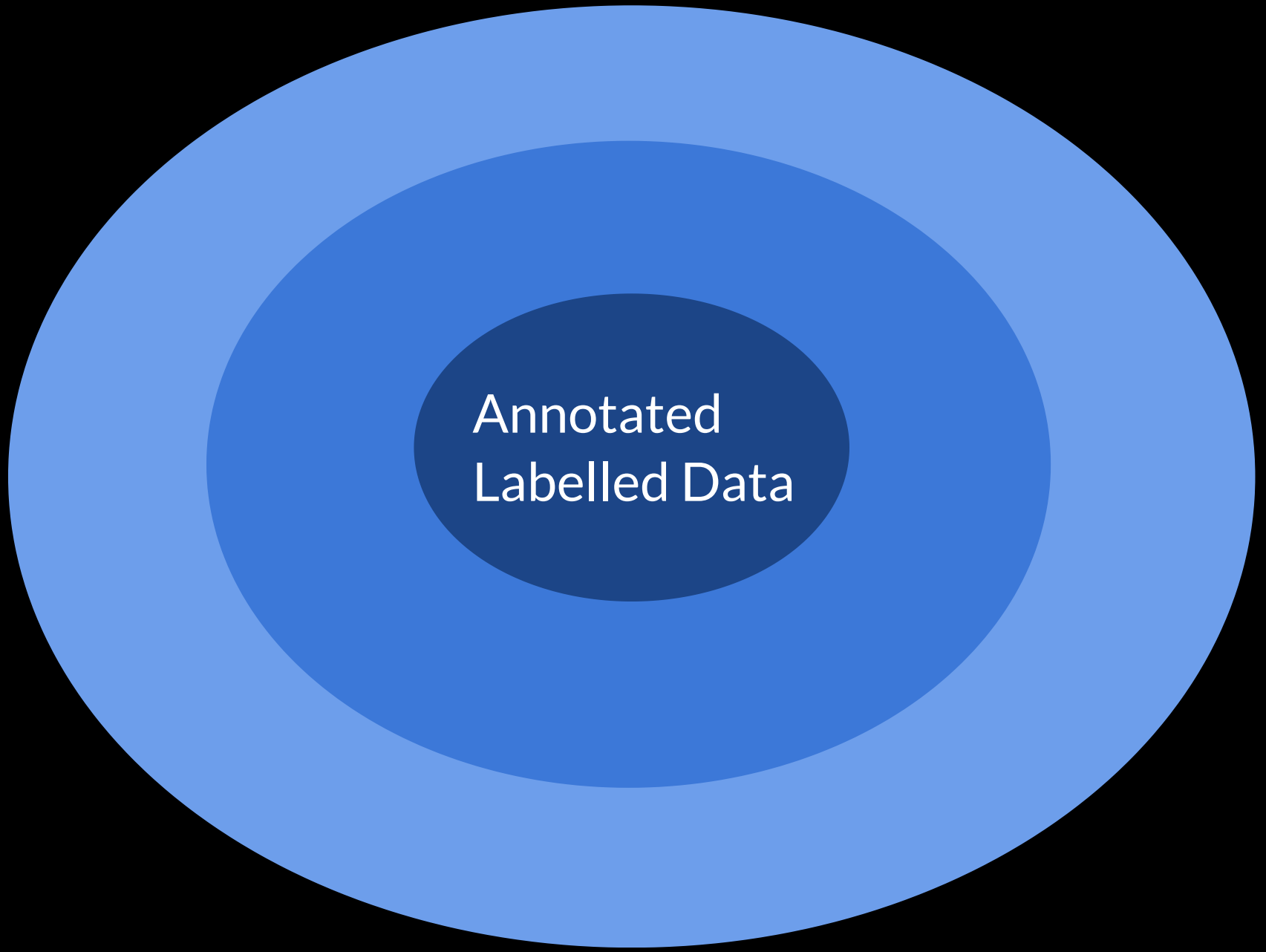
Introducing OpenAI o1

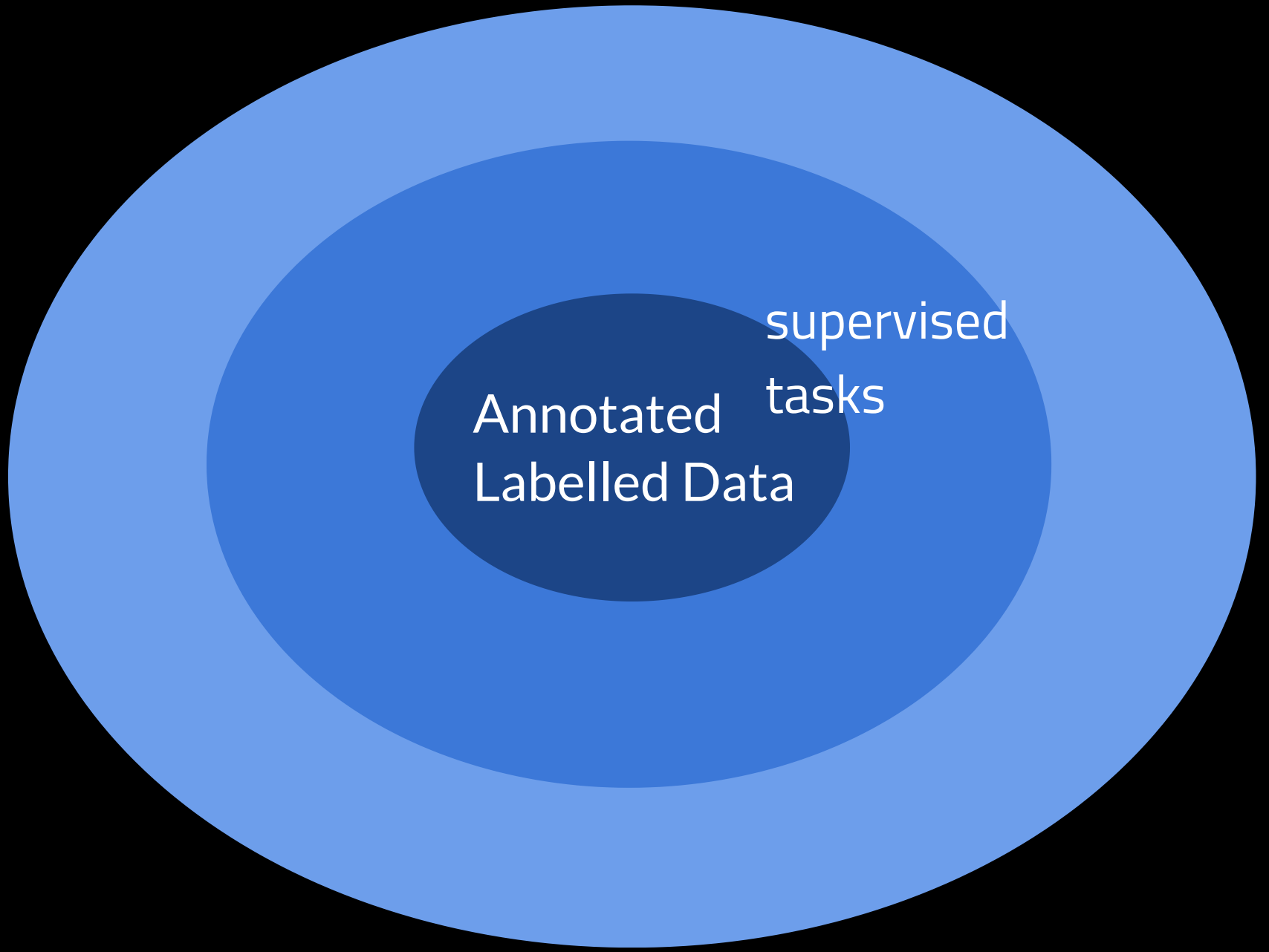
We've developed a new series of AI models designed to spend more time thinking before they respond. Here is the latest news on o1 research, product and other updates.

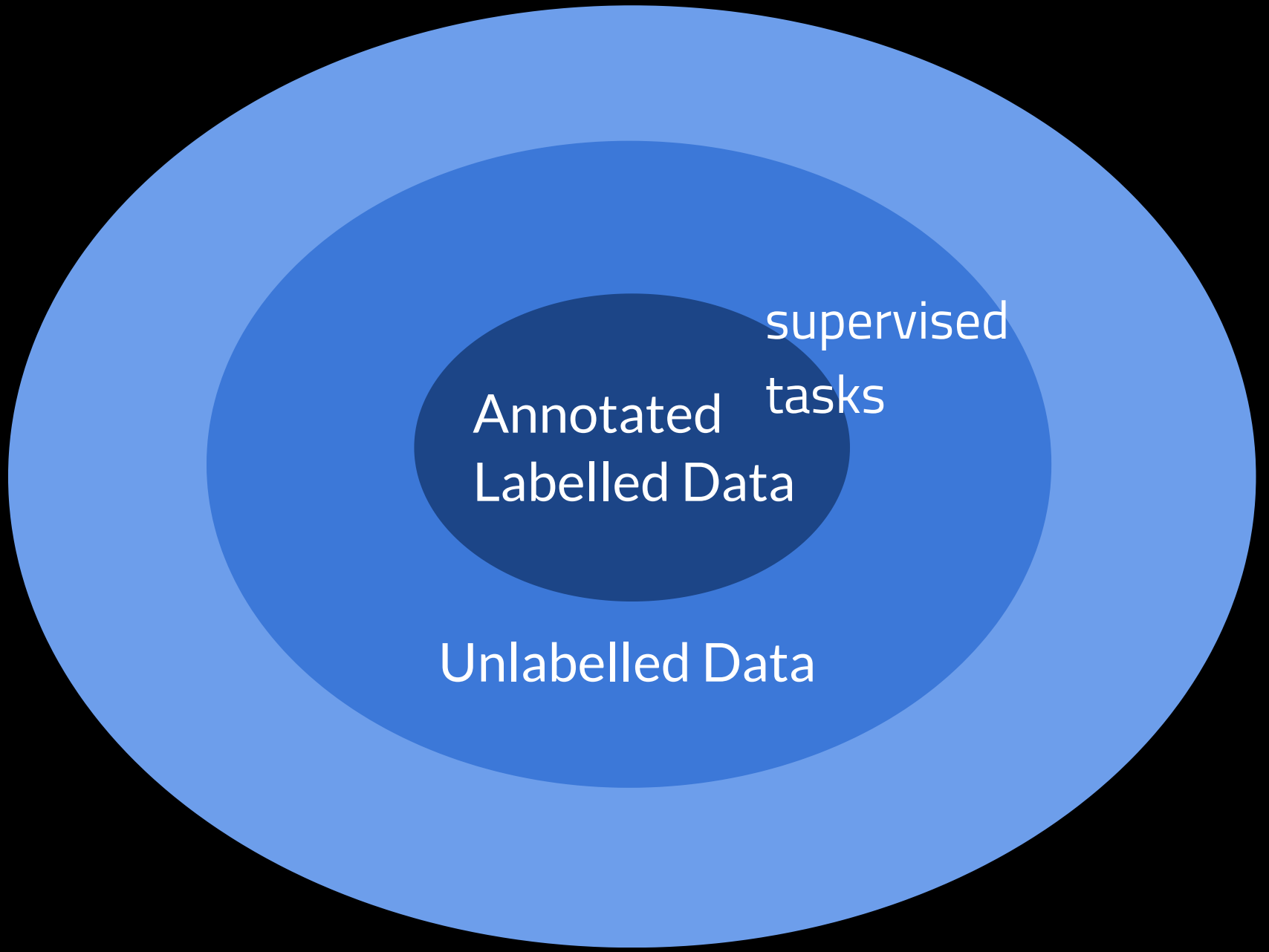
Try it in ChatGPT Plus ↗ Try it in the API ↗

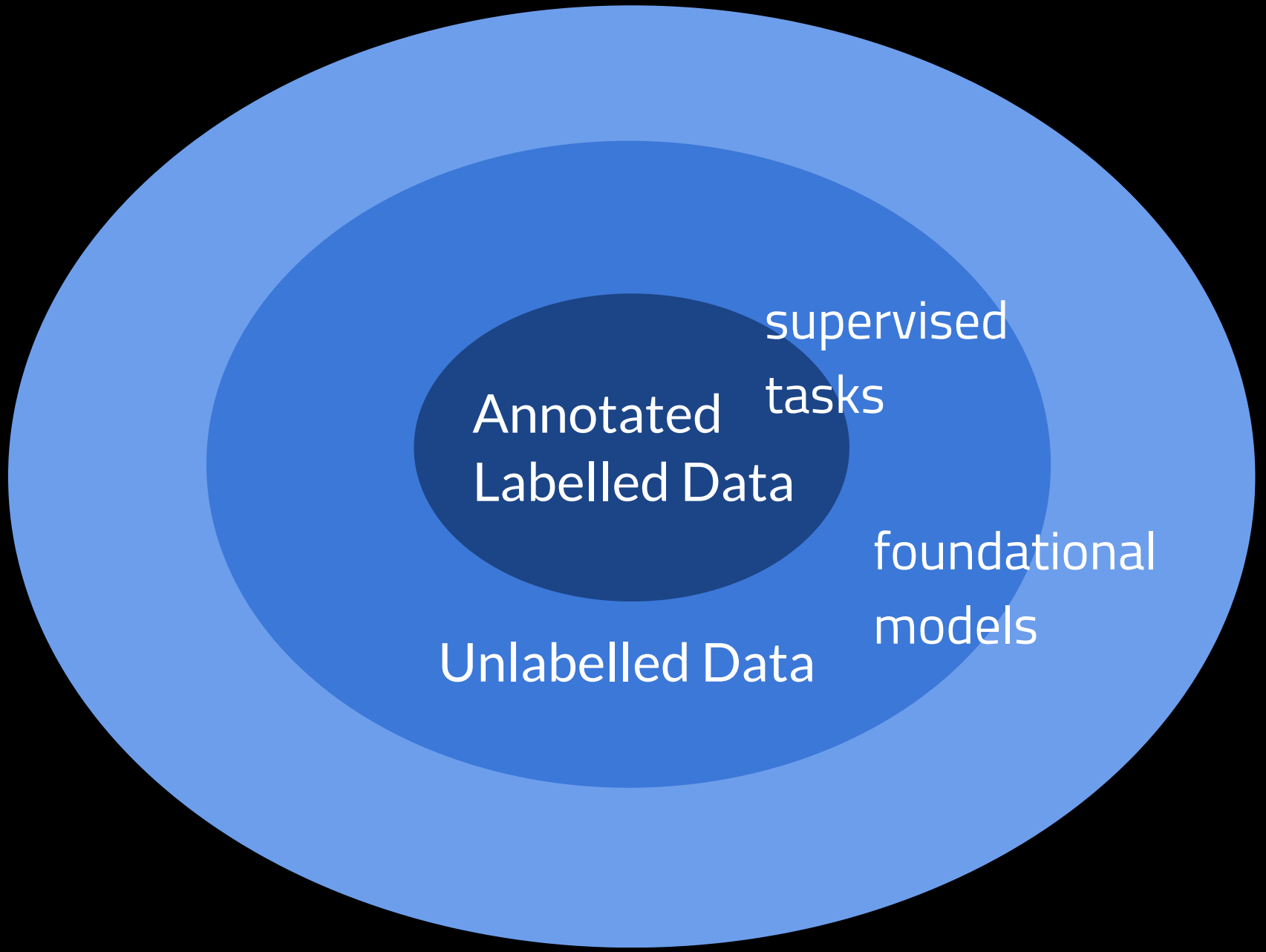
Sort

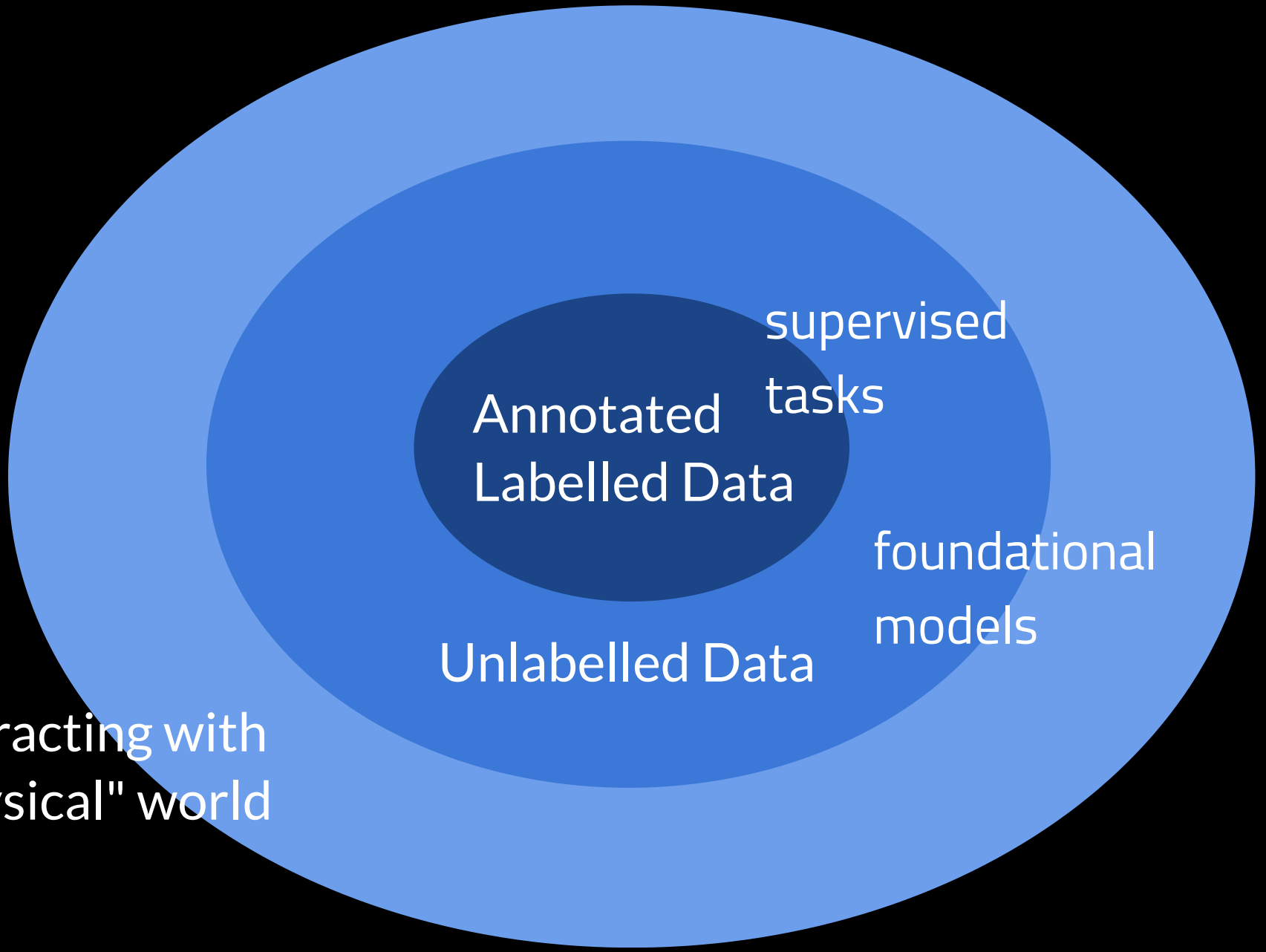
Self-Play Reinforcement Learning

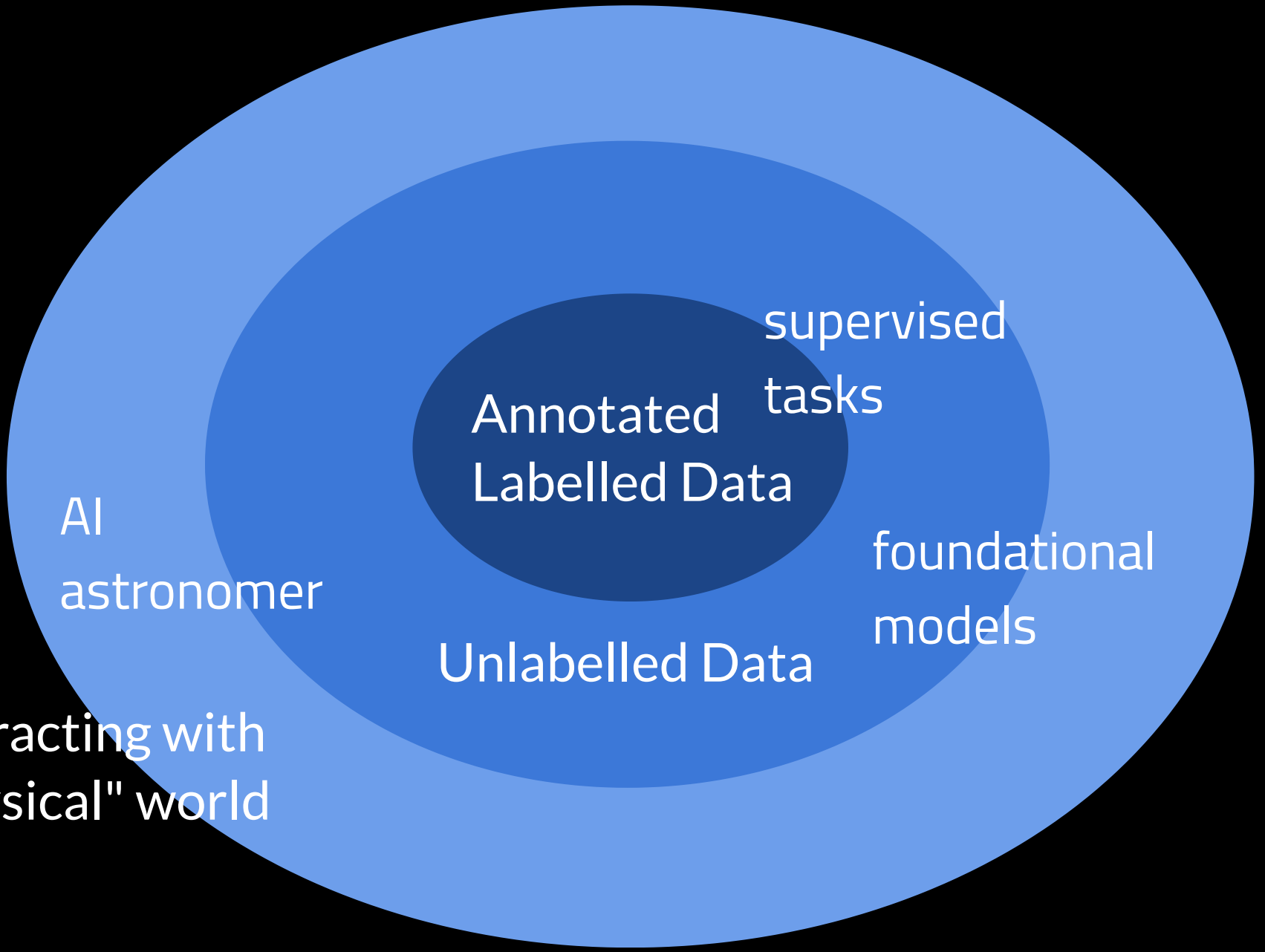

Annotated
Labelled Data


supervised
tasks
Annotated
Labelled Data


supervised
tasks
Annotated
Labelled Data
Unlabelled Data


supervised
tasks
Annotated
Labelled Data
foundational
models
Unlabelled Data

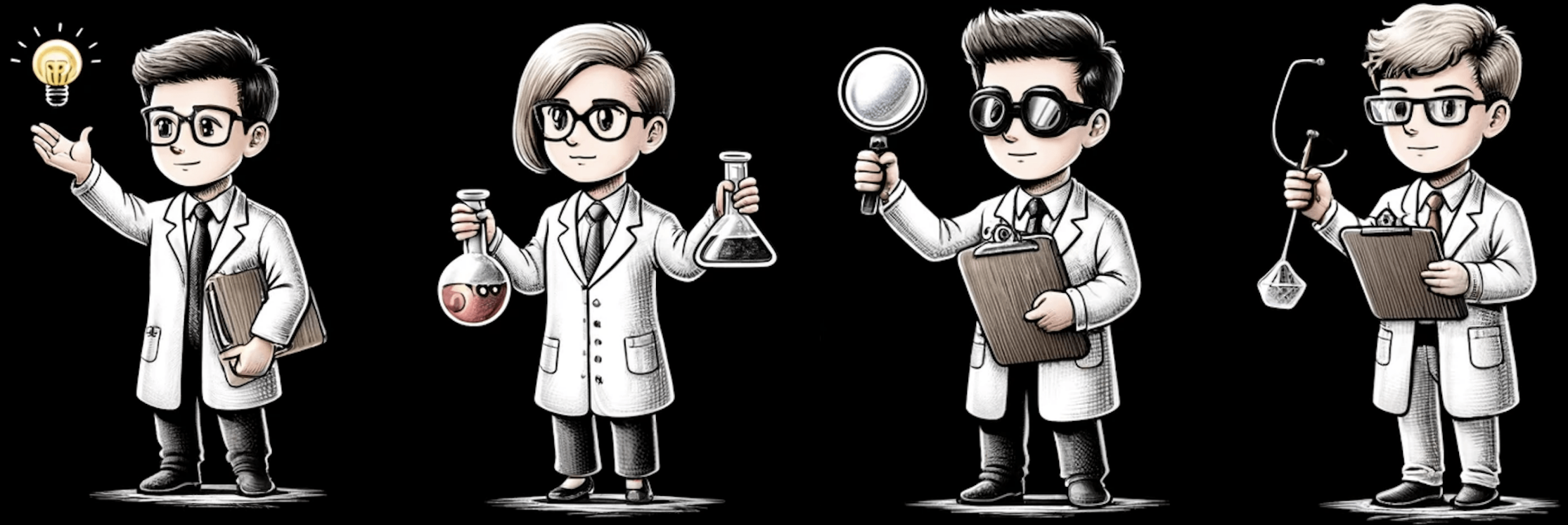

supervised
tasks
Annotated
Labelled Data
foundational
models
Unlabelled Data
Interacting with
"physical" world


supervised
tasks
Annotated
Labelled Data
AI
astronomer
foundational
models
Unlabelled Data
Interacting with
"physical" world

Introducing *Mephisto**

* In the classic tale of Faust, Mephisto is a demon who tempts the scholar Faust with *knowledge* and power in exchange for his soul.

A *collaboration* of multiple AI agents (LLM models)



Proposing actions

Execute actions

State evolution

Knowledge distillation

A *collaboration* of multiple AI agents (LLM models)

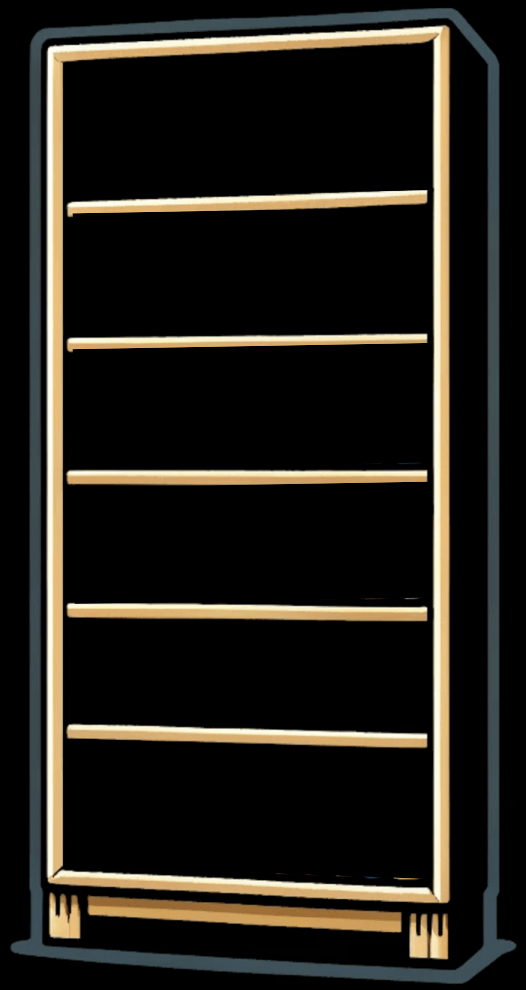
Proposing actions

Execute actions

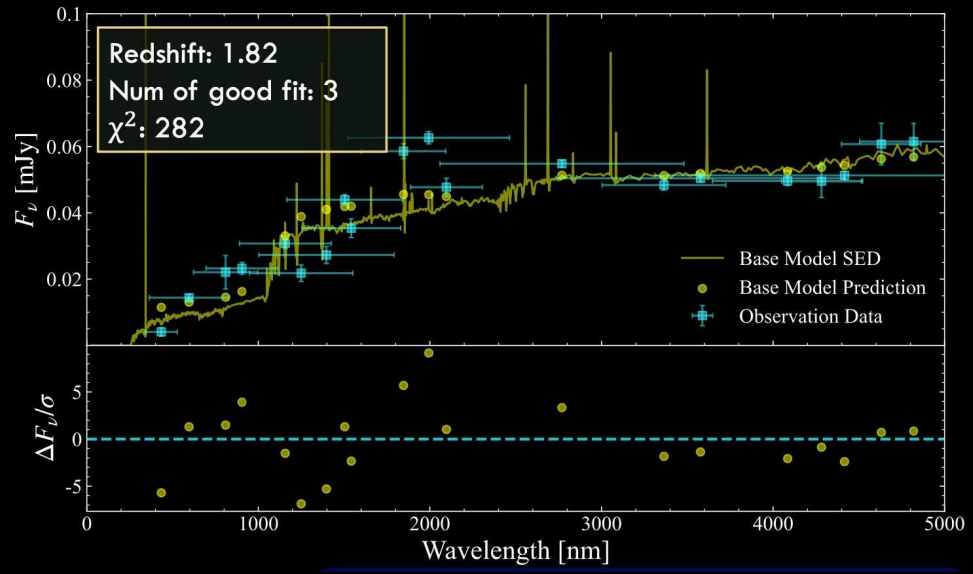
State evolution

Knowledge distillation

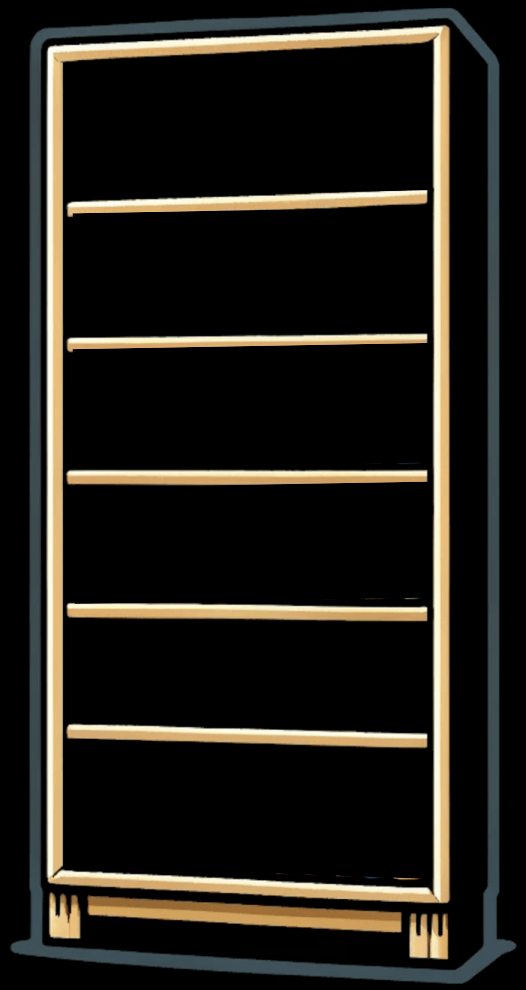
Enabling AI to collect *"knowledge"* through exploration



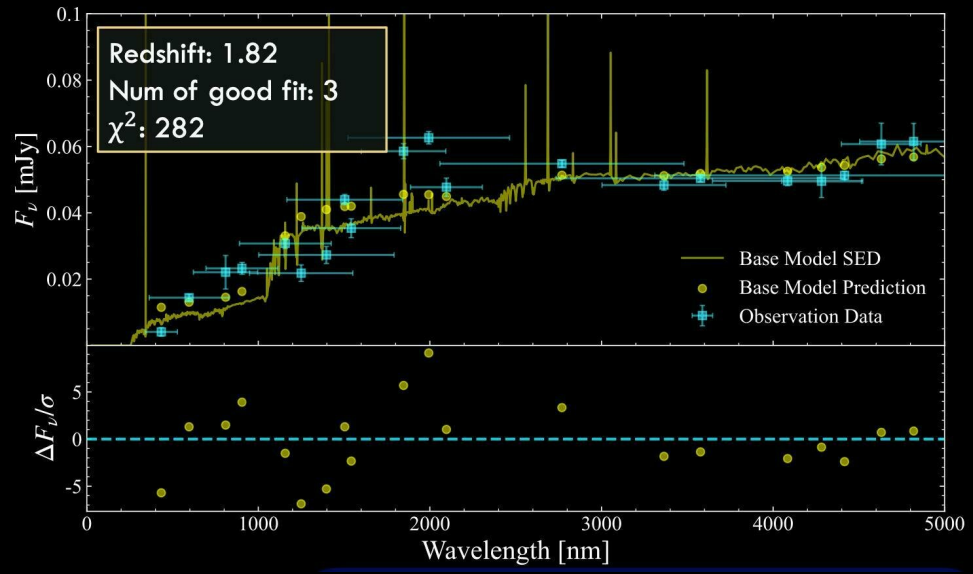
Knowledge base



Enabling AI to collect *"knowledge"* through exploration

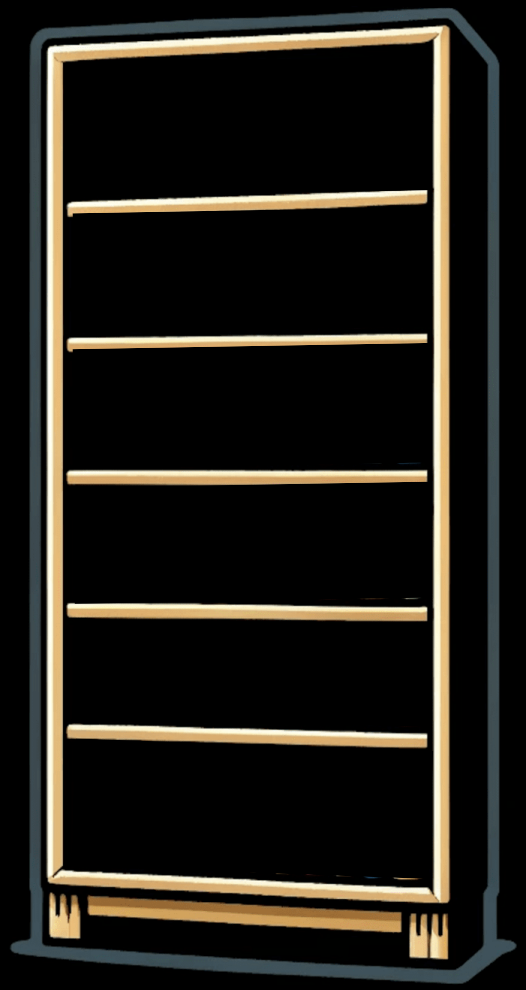


Knowledge base

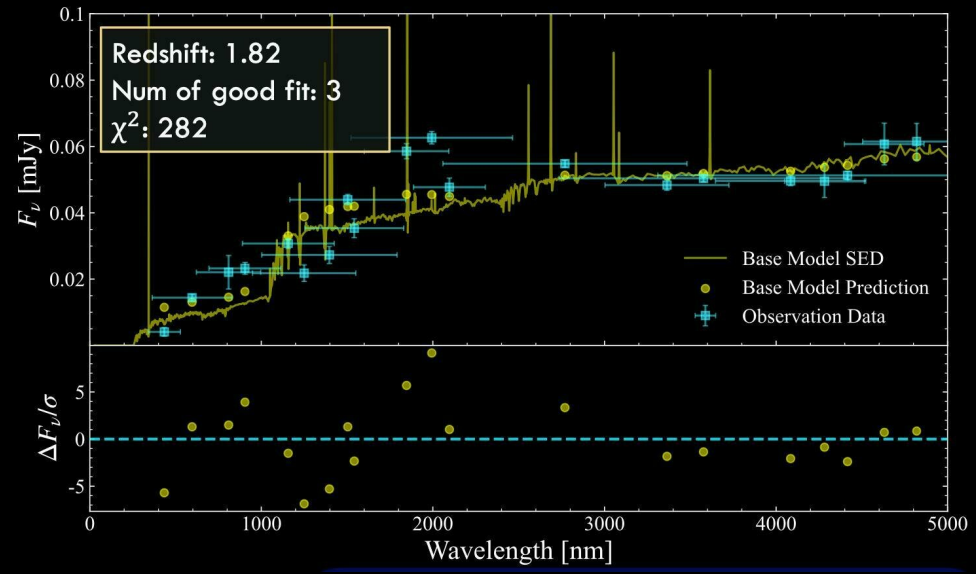


Proposing Actions - e.g., different physical models / parameter range

Enabling AI to collect *"knowledge"* through exploration



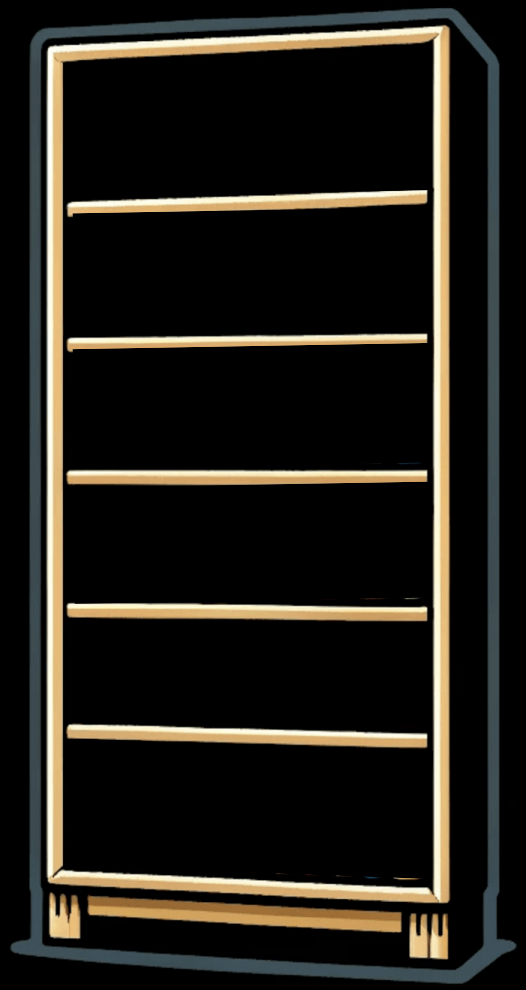
Knowledge base



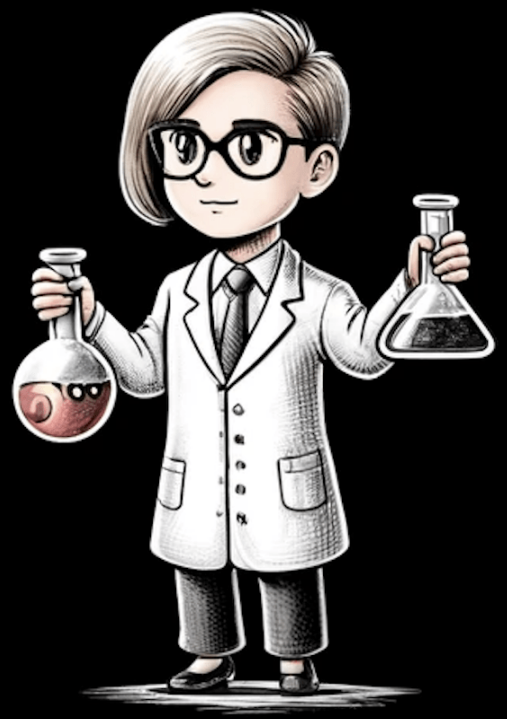
1 2 3 4

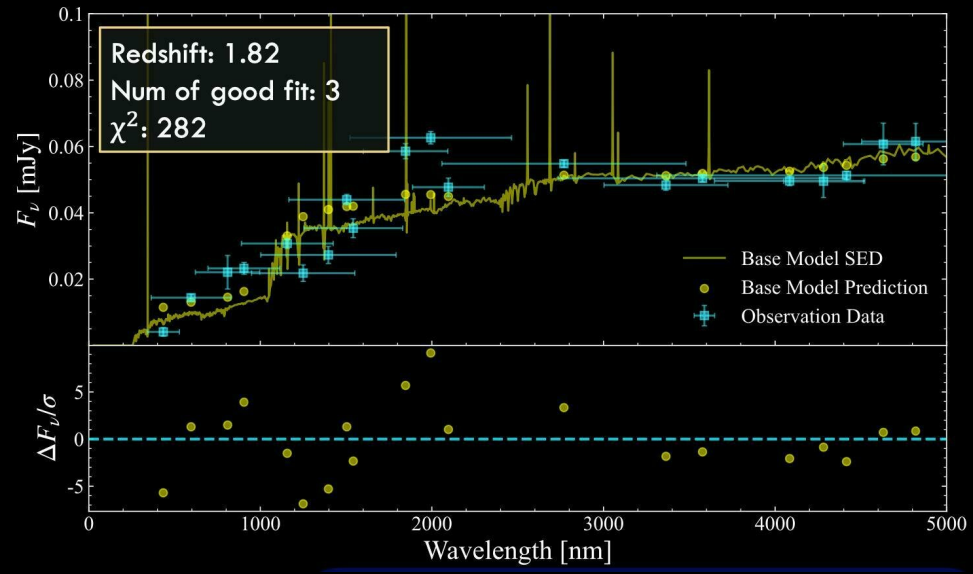
Proposing Actions - e.g., different physical models / parameter range

Enabling AI to collect *"knowledge"* through exploration



Knowledge base

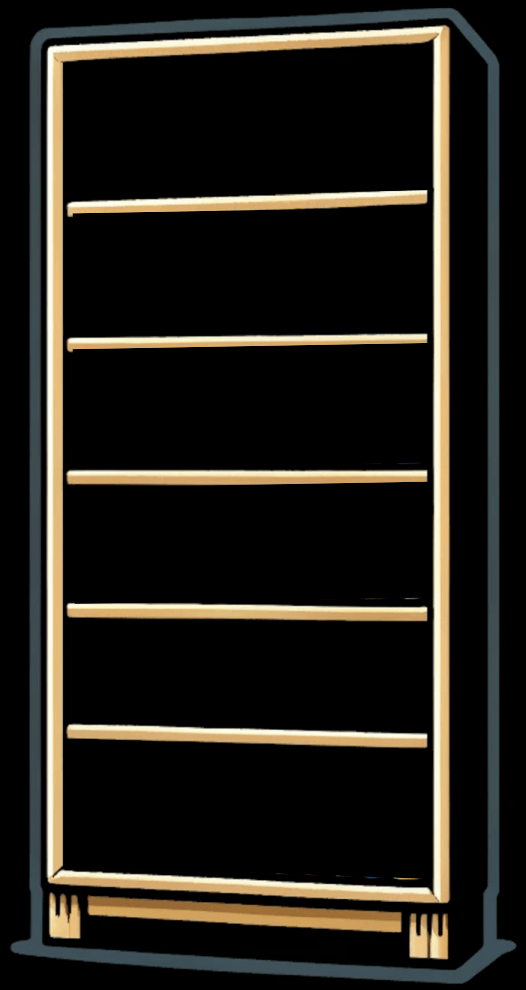




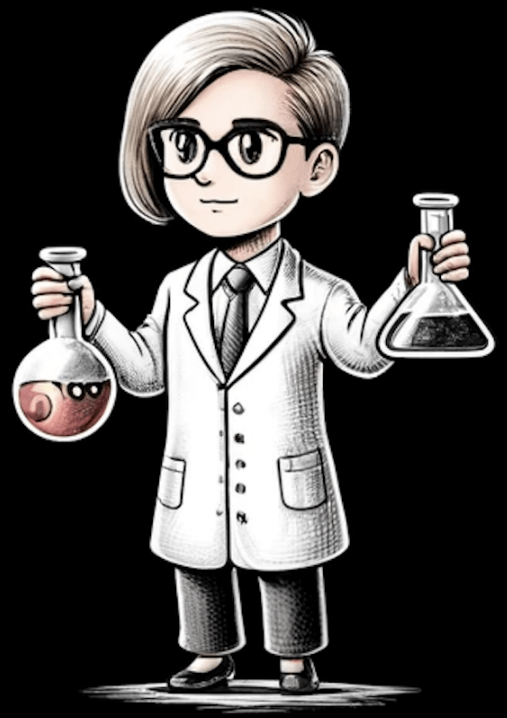
1 2 3 4

Execute Actions - write configuration files, run the codes, automously

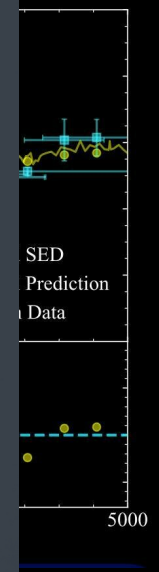
Enabling AI to collect *"knowledge"* through exploration



Knowledge base



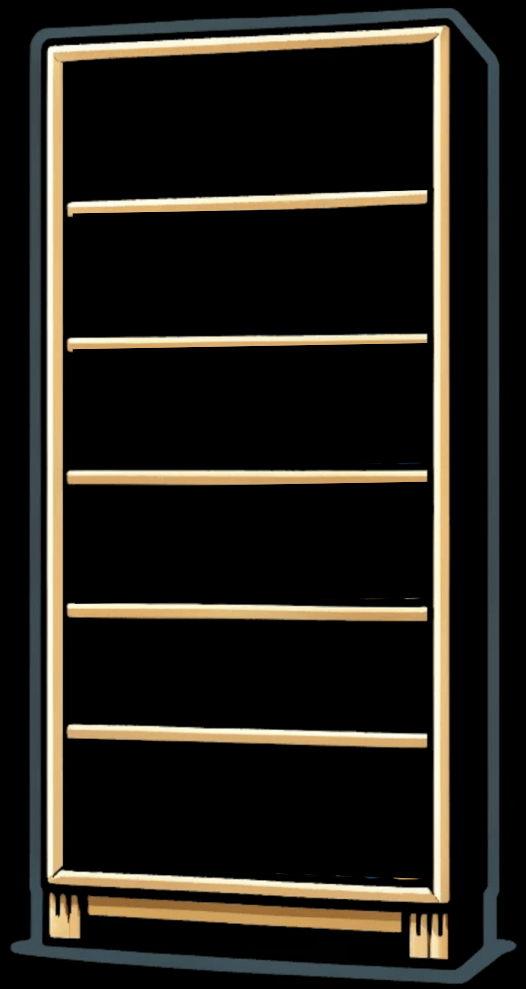
```python
def check_db():
    if not os.path.isfile(FILE_URI):
        db.create_all()

@app.route("/")
def home():
    check_db()
    all_books = db.session.query(Book).all()
    return render_template("index.html", books=all_books)

@app.route("/edit", methods=["GET","POST"])
def edit():

    if request.method == 'POST':
        book_id = request.form["id"]
        book_to_update = Book.query.get(book_id)
        book_to_update.rating = request.form["rating"]
        db.session.commit()
        return redirect(url_for('home'))
```



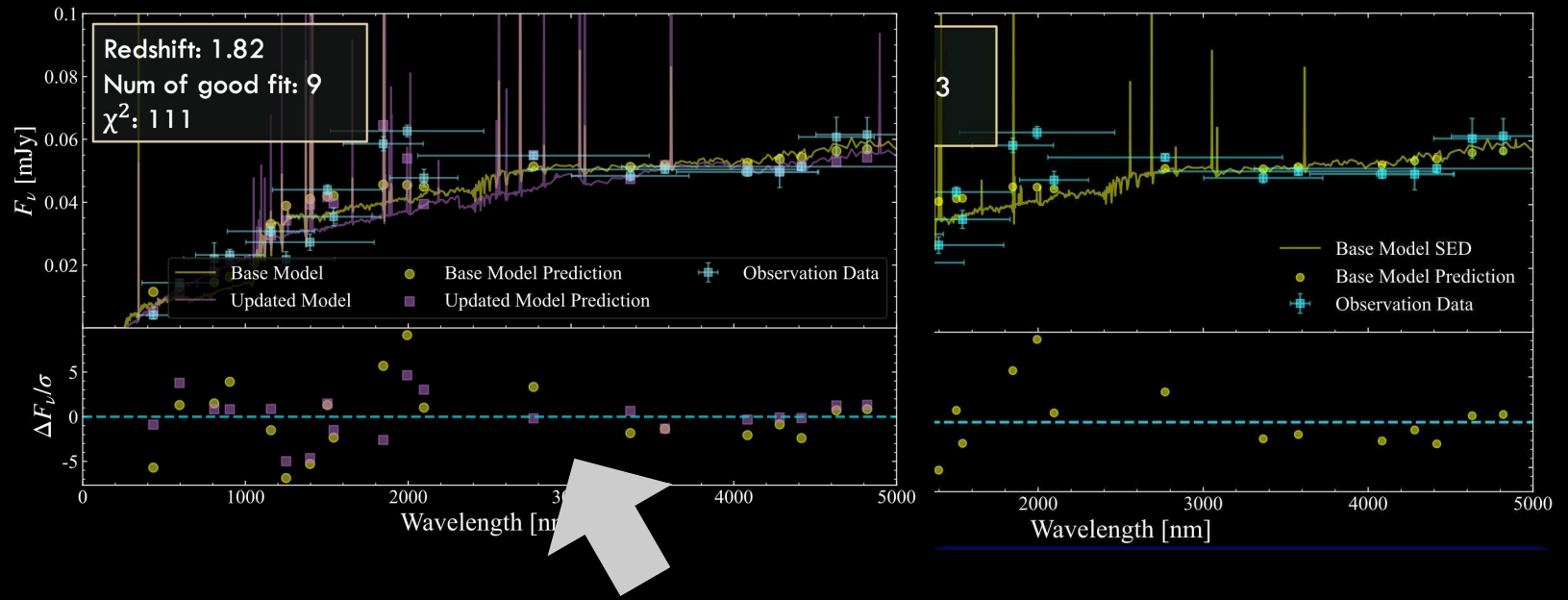
1 2 3 4

Execute Actions - write configuration files, run the codes, automously

Enabling AI to collect *"knowledge"* through exploration

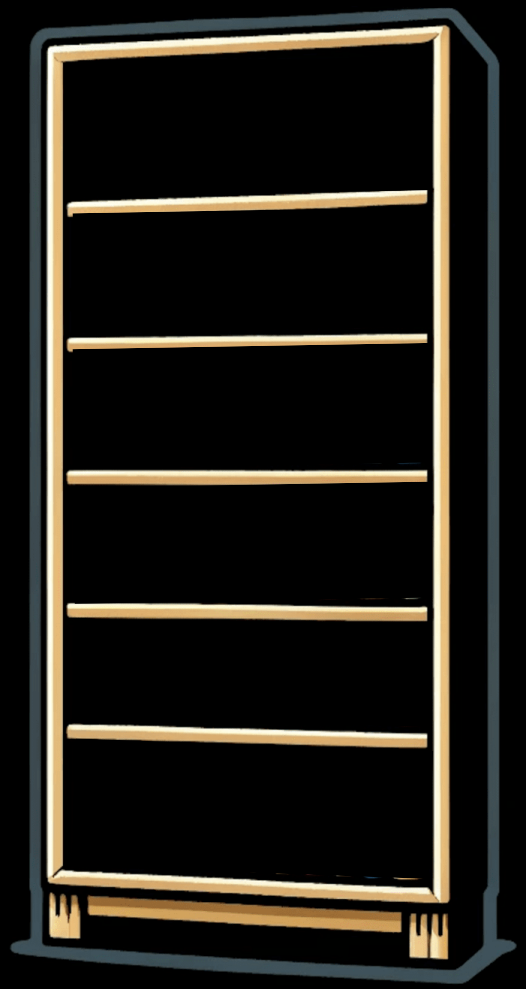


Knowledge base



1 2 3 4

Enabling AI to collect *"knowledge"* through exploration

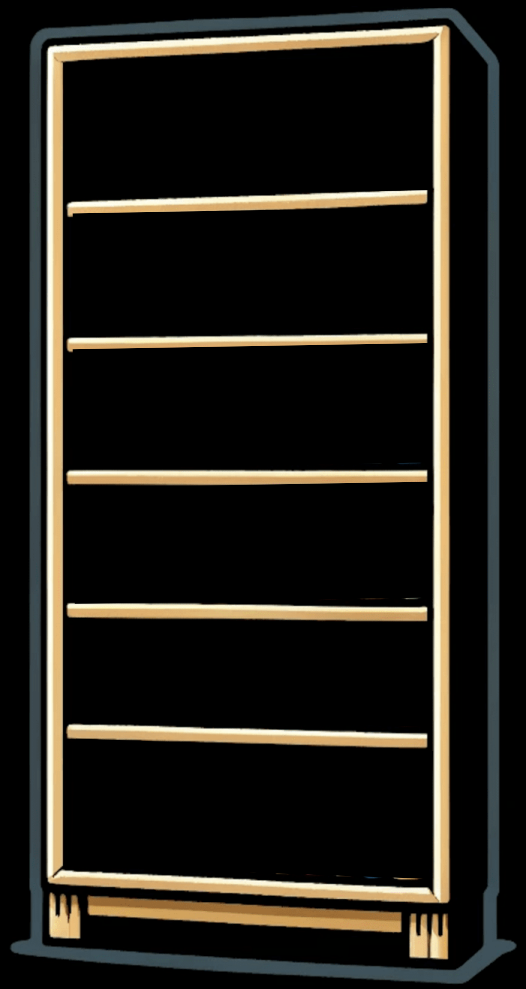


Knowledge base

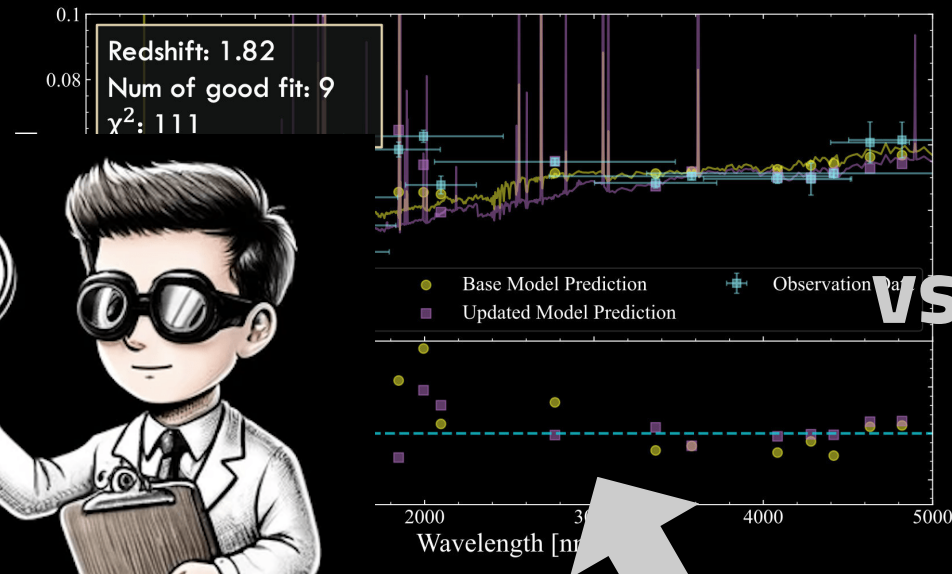
1 2 3 4

State Evaluation - evaluate the results (beyond a single error metric)

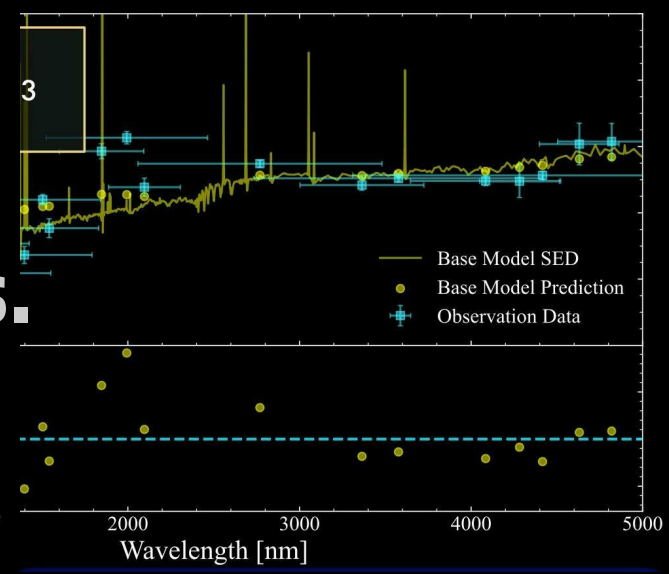
Enabling AI to collect *"knowledge"* through exploration



Knowledge base

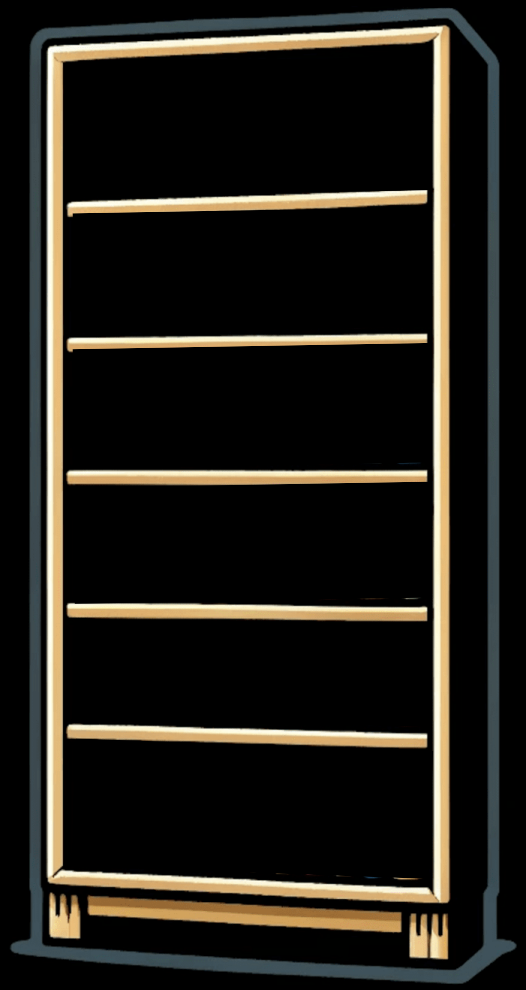


vs.

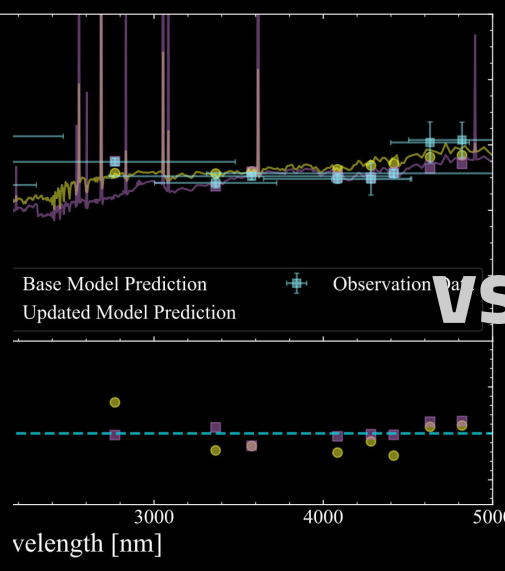


State Evaluation - evaluate the results (beyond a single error metric)

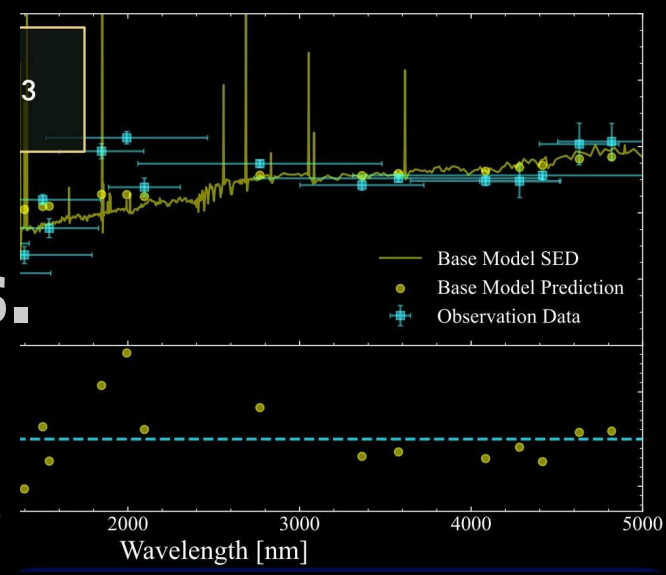
Enabling AI to collect *"knowledge"* through exploration



Knowledge base



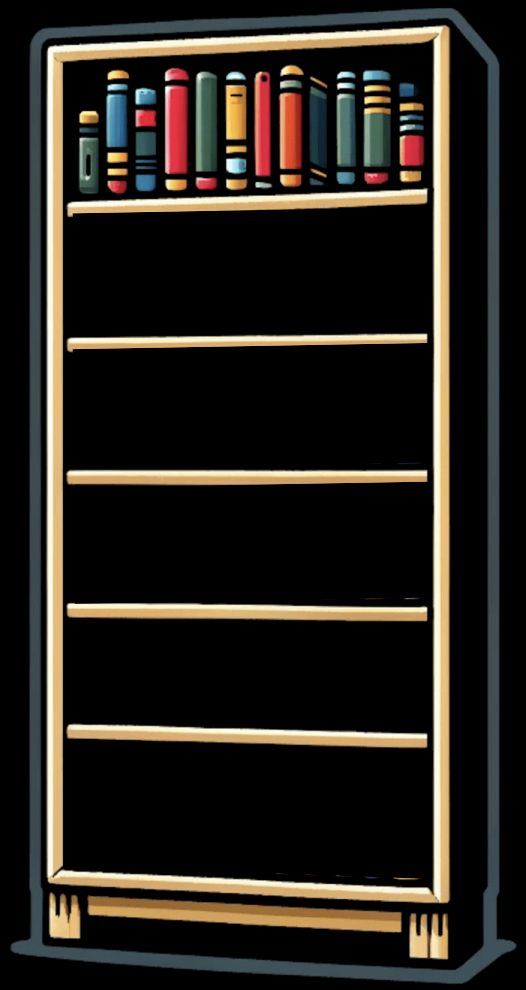
vs.



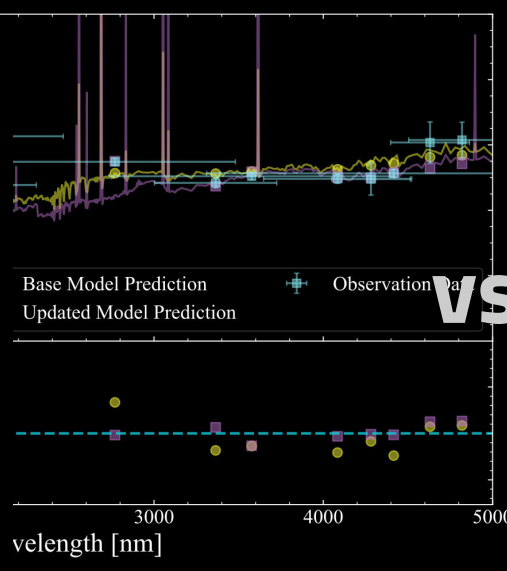
1 vs. 2 vs. 3 vs. 4

Knowledge Distillation - summarise useful actions given the previous state

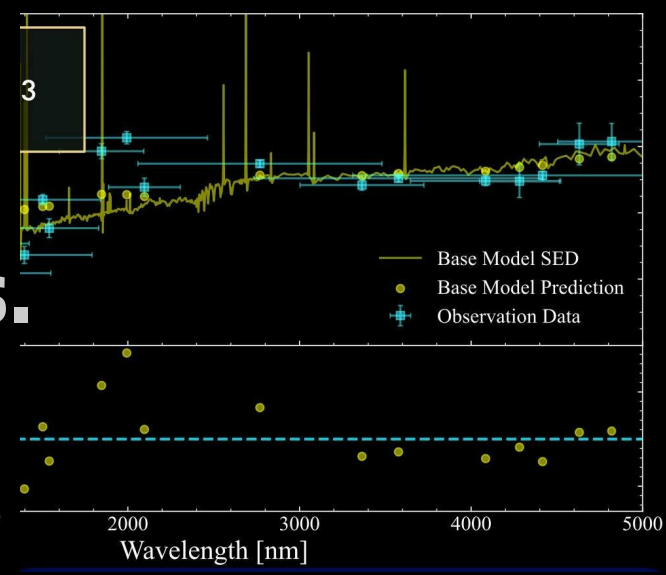
Enabling AI to collect *"knowledge"* through exploration



Knowledge base



vs.



1 vs. 2 vs. 3 vs. 4

Knowledge Distillation - summarise useful actions given the previous state

Example of learned "knowledge"

Example of learned "knowledge"

" If the fit is *overestimated in the UV and optical* bands,

Example of learned "knowledge"

" If the fit is *overestimated in the UV and optical* bands,

increasing the E_BV_lines parameter may lead to a better fit by accounting for more *dust attenuation* in these bands. "

Example of learned "knowledge"

Example of learned "knowledge"

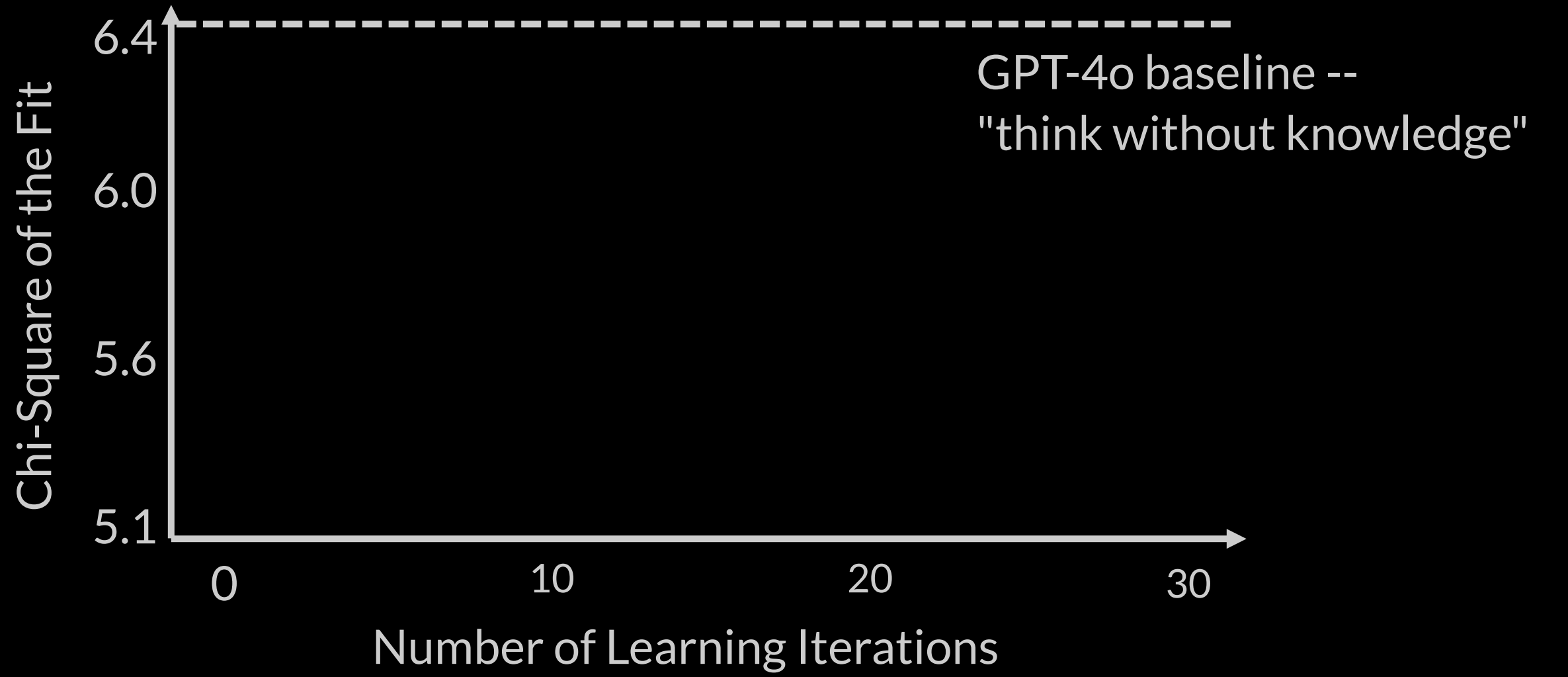
" If there is a gross *underestimation in the MWIR bands*,

Example of learned "knowledge"

" If there is a gross *underestimation in the MWIR bands*,

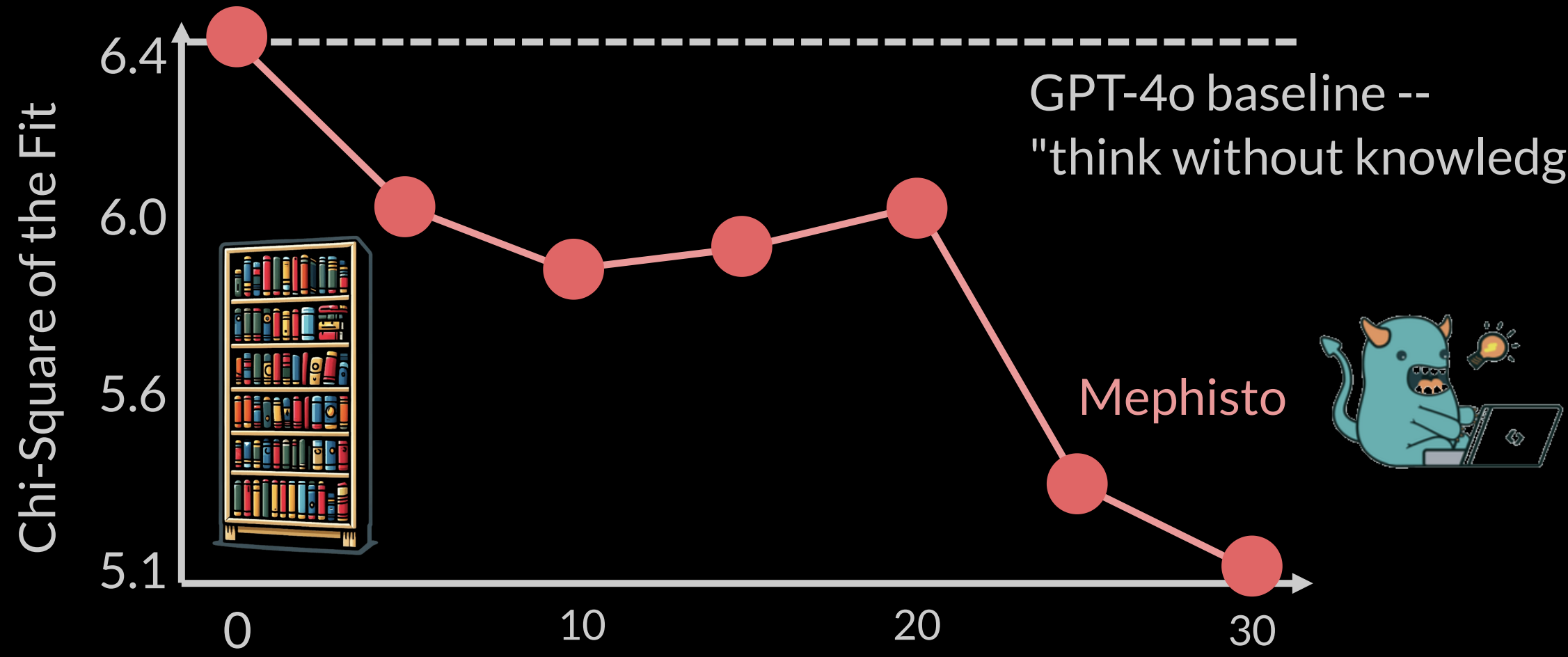
consider exploring a wider range of *fracAGN values* in the agn module to improve the fit in these bands "

LLMs with self-play RL *outperforms* native LLMs



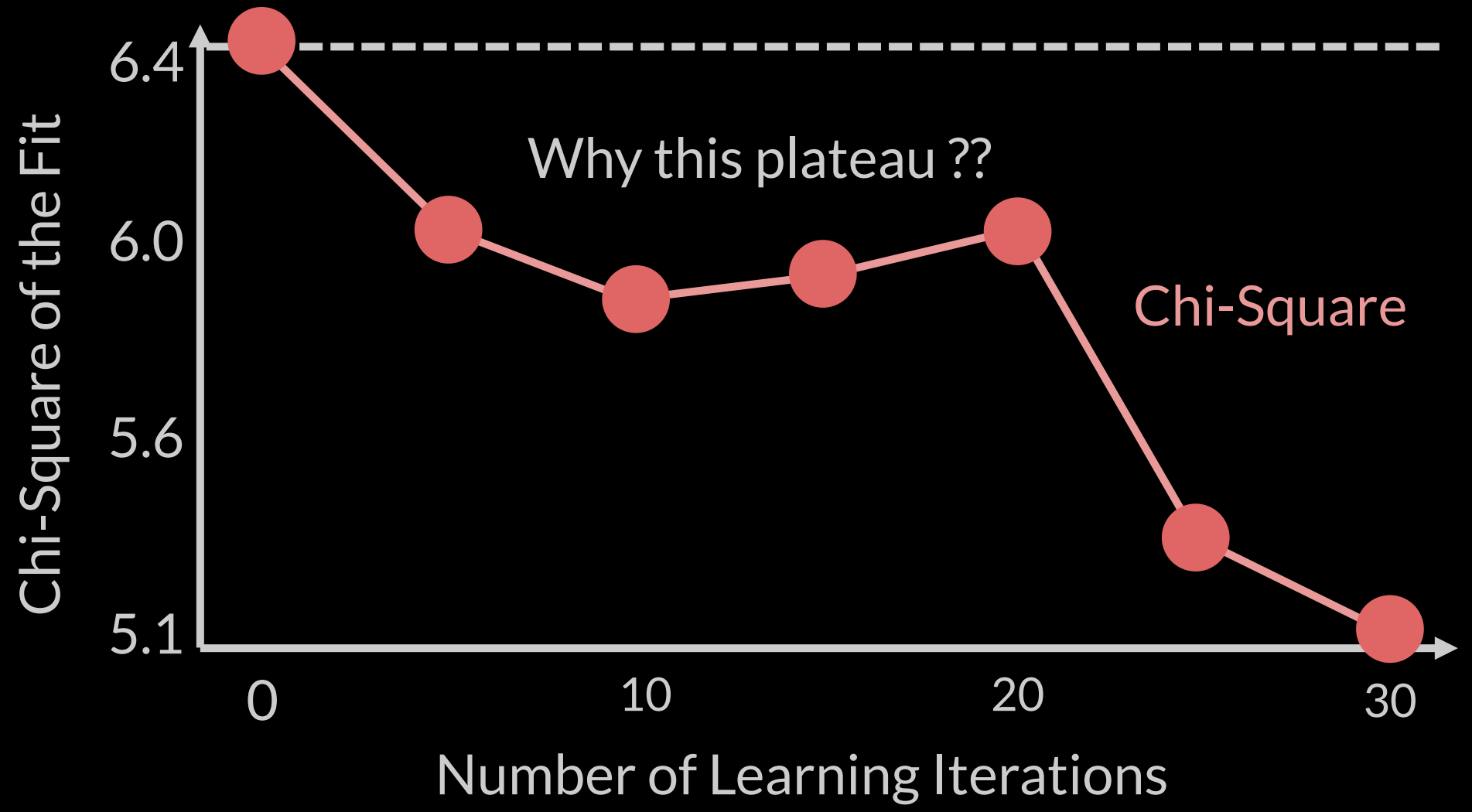
Sun, YST+, 2024b

LLMs with self-play RL *outperforms* native LLMs



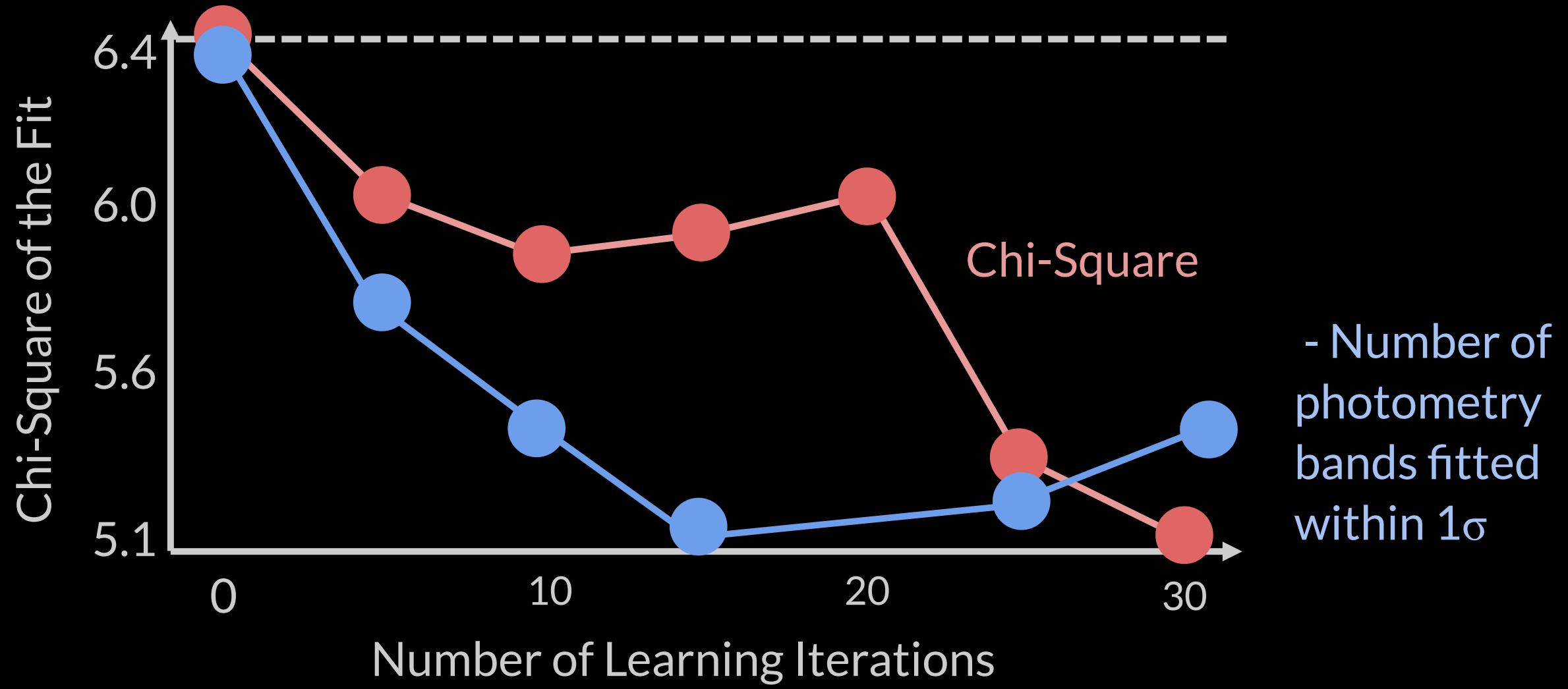
Sun, YST+, 2024b

LLMs with self-play RL *outperforms* native LLMs



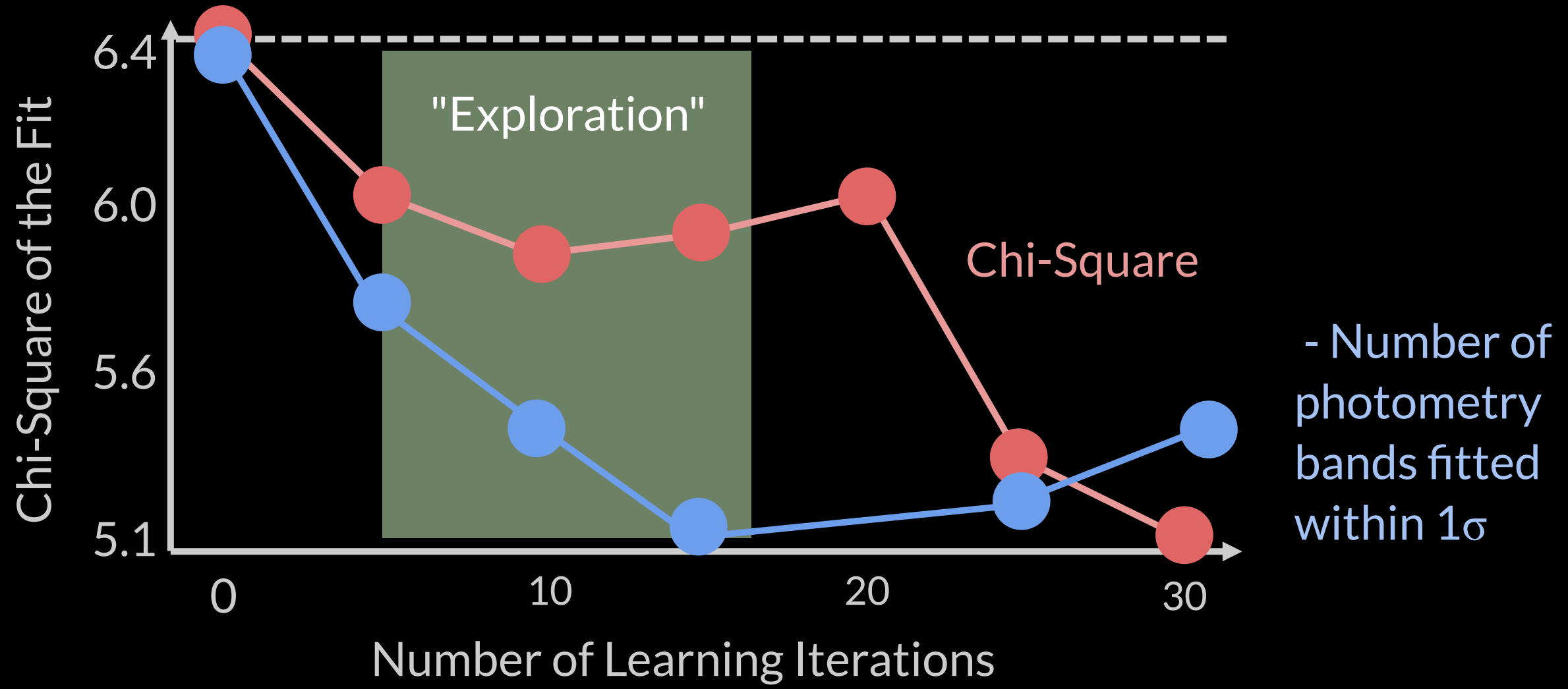
Sun, YST+, 2024b

LLMs with self-play RL *outperforms* native LLMs



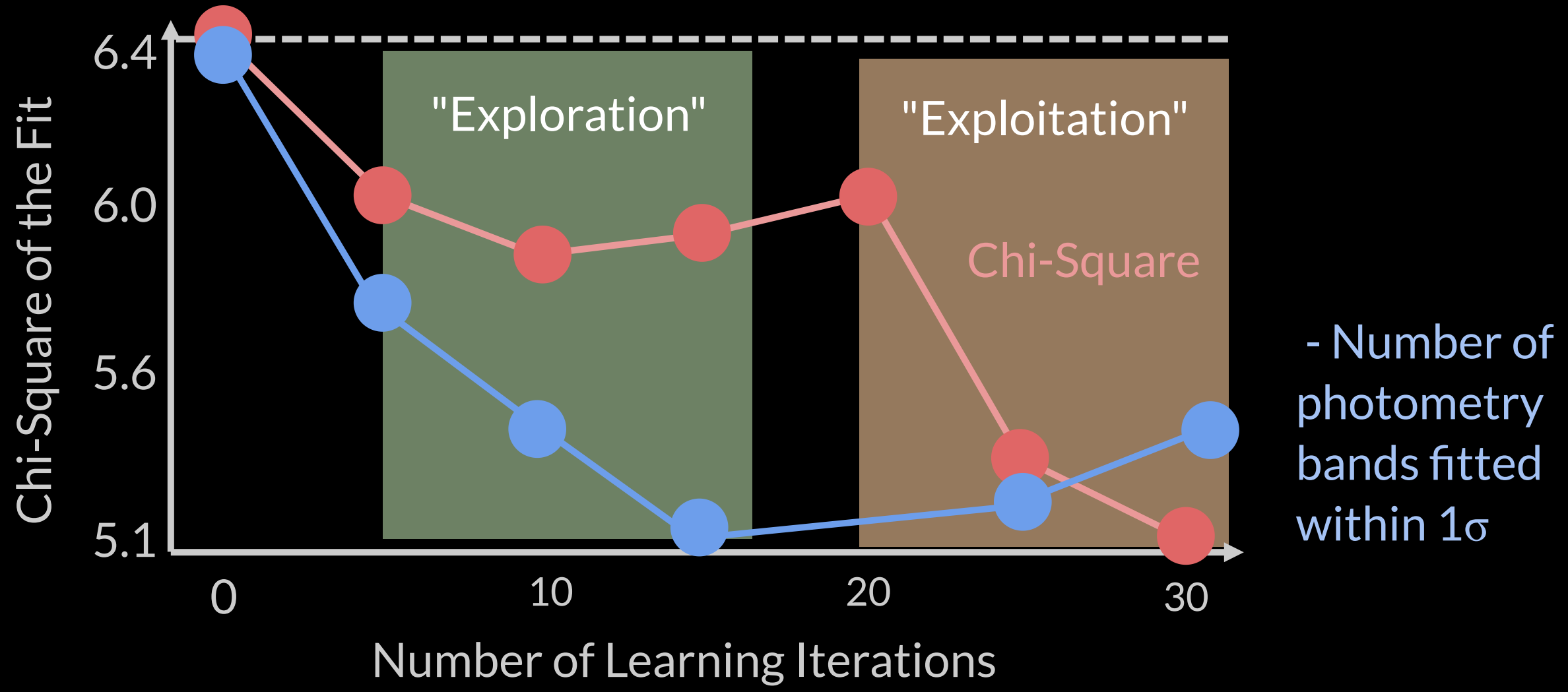
Sun, YST+, 2024b

LLMs with self-play RL *outperforms* native LLMs



Sun, YST+, 2024b

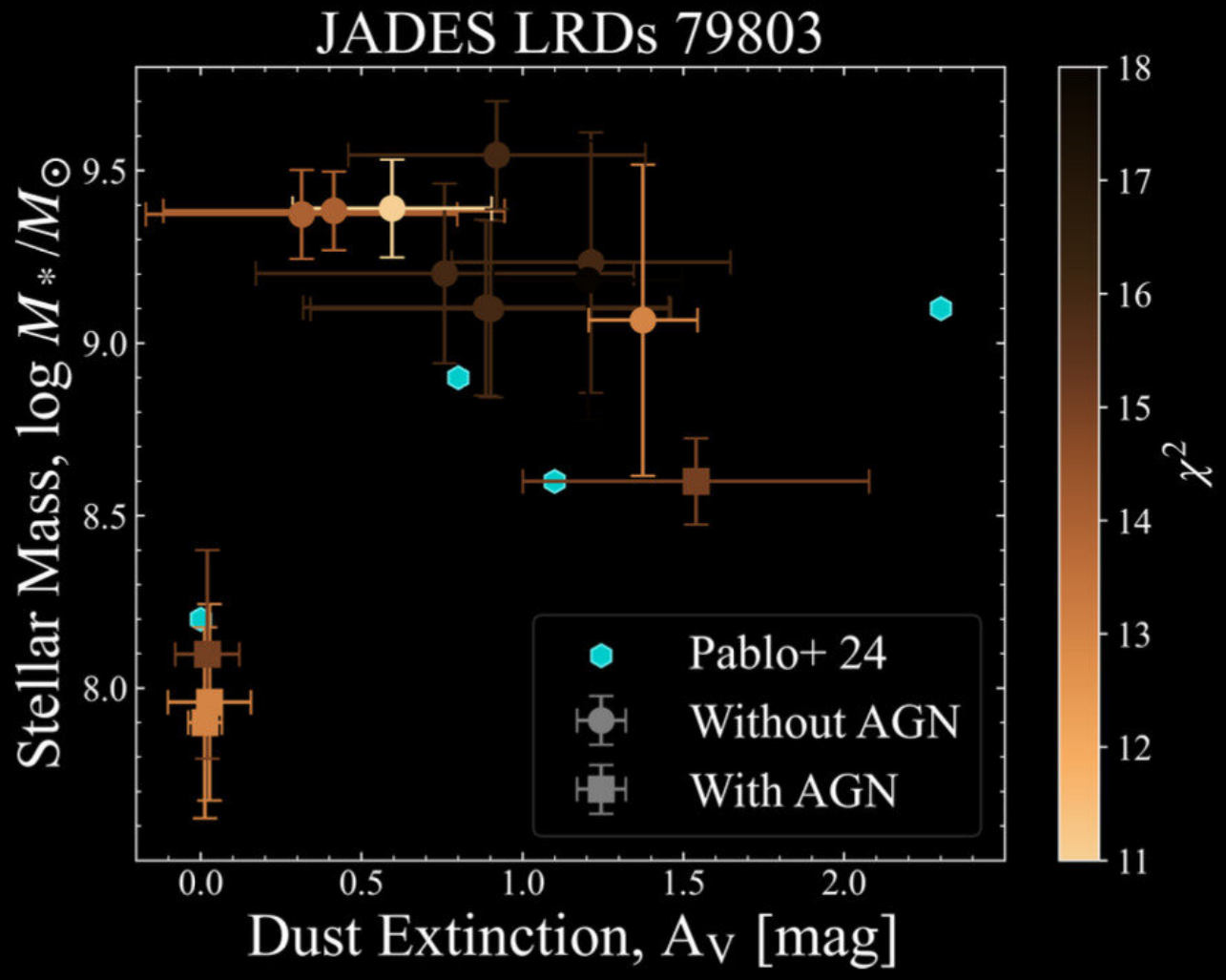
LLMs with self-play RL *outperforms* native LLMs

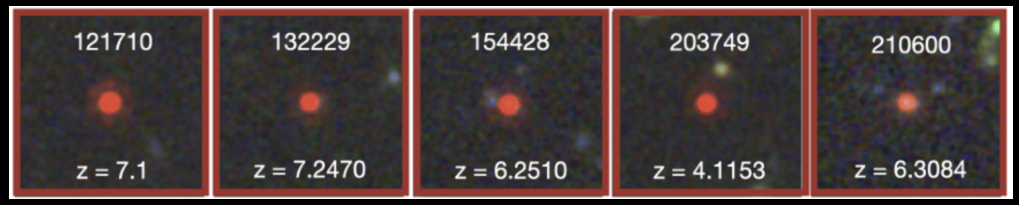


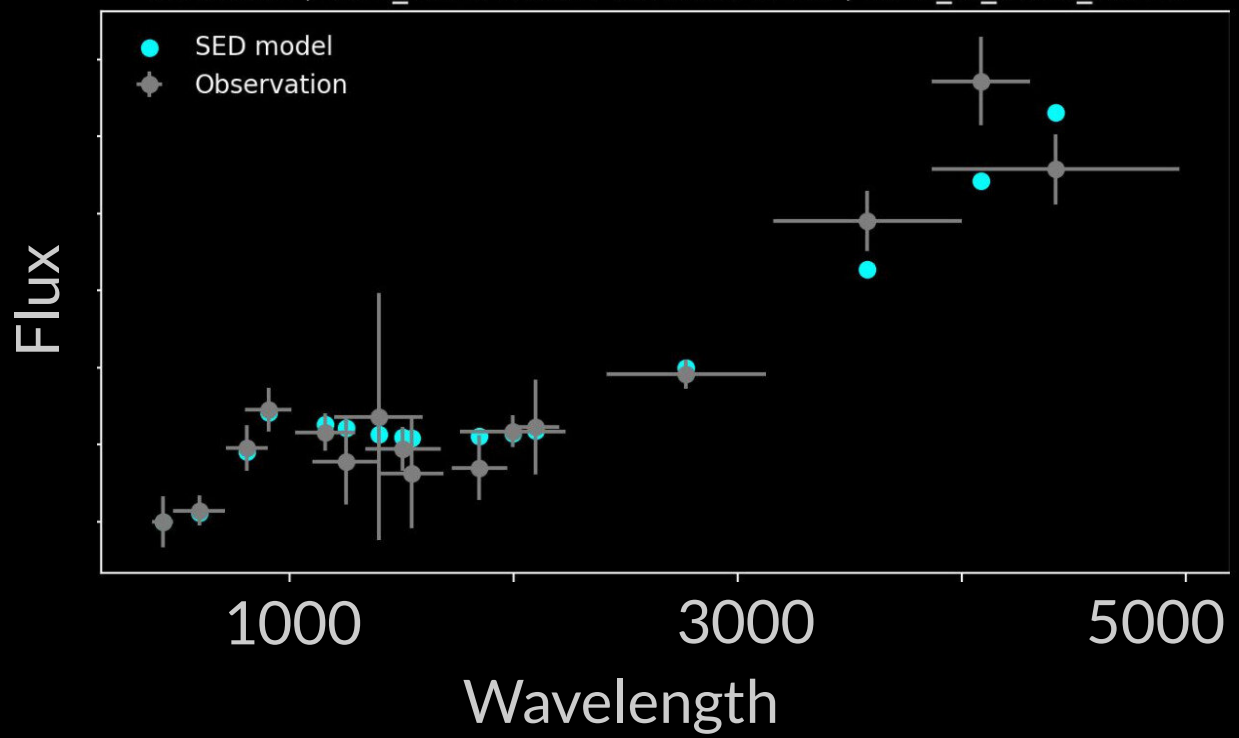
Sun, YST+, 2024b

LLMs can reach *human-level reasoning* for specific astronomy analysis tasks through *self-play* reinforcement learning

Explaining James Webb's *"little red dot" galaxies* with mephisto





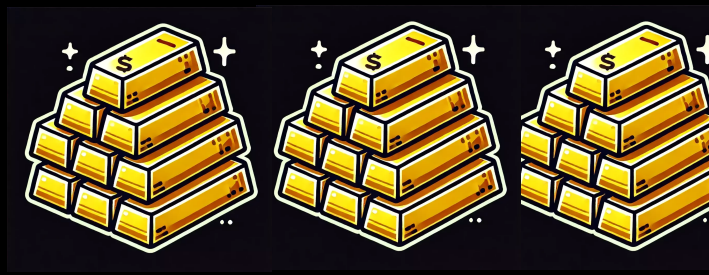


Potentially identifying *all* the astronomical objects' SED that our current physics can't explain

Provided that we have *a capable model* that can generate inference quickly and *cost efficiently*....

Provided that we have *a capable model* that can generate inference quickly and *cost efficiently*....

capable model **vs.** cost efficiency



In the SED case study, we need *~0.1M tokens* per source

capable model **vs.** cost efficiency

e.g., GPT-4o (this study)

In the SED case study, we need *~0.1M tokens* per source

capable model **vs.** cost efficiency

e.g., GPT-4o (this study)

Model	Input	Output
gpt-4o	$5.00 / 1M tokens	$15.00 / 1M tokens
gpt-4o-2024-05-13	$5.00 / 1M tokens	$15.00 / 1M tokens

In the SED case study, we need *~0.1M tokens* per source

capable model **vs.** cost efficiency

e.g., GPT-4o (this study) = USD 1 *per source*

Model	Input	Output
gpt-4o	$5.00 / 1M tokens	$15.00 / 1M tokens
gpt-4o-2024-05-13	$5.00 / 1M tokens	$15.00 / 1M tokens

In the SED case study, we need *~0.1M tokens* per source

1B sources = $1 billion

e.g., *Roman* Space Telescope, *Euclid* Space Telescope

1B sources = $1 billion

e.g., *Roman* Space Telescope, *Euclid* Space Telescope

~ approximately the build cost

How do we get there in a *cost-effective* way?

The first extensive *benchmarking effort* of Large Language Models in terms of astronomy Q&A.

AstroMLab 1: Who Wins Astronomy Jeopardy!?

Yuan-Sen Ting (丁源森),[1,2,3,4] Tuan Dung Nguyen,[5] Tirthankar Ghosal,[6] Rui Pan (潘瑞),[7] Hardik Arora,[8] Zechang Sun (孙泽昌),[9] Tijmen de Haan,[10,11] Nesar Ramachandra,[12] Azton Wells,[12] Sandeep Madireddy,[13] and Alberto Accomazzi[14]

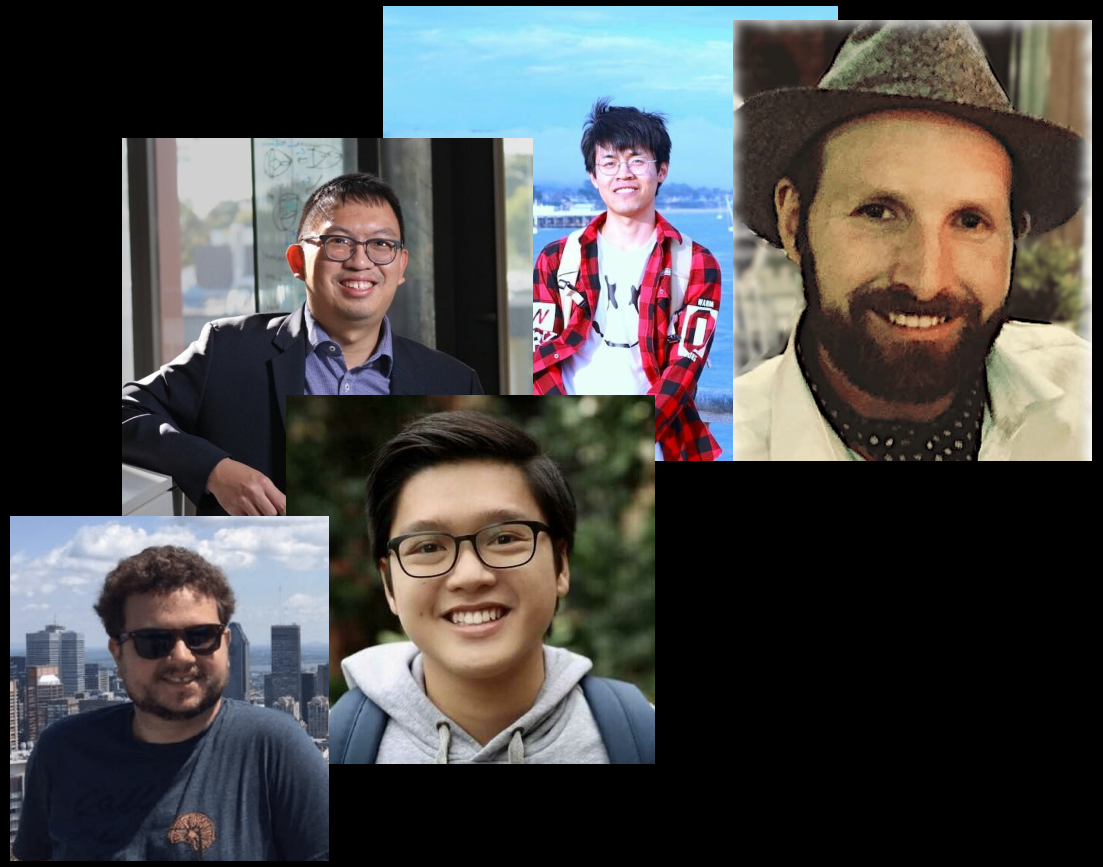
[1] *Research School of Astronomy & Astrophysics, Australian National University, Cotter Rd., Weston, ACT 2611, Australia*
[2] *School of Computing, Australian National University, Acton, ACT 2601, Australia*
[3] *Department of Astronomy, The Ohio State University, Columbus, OH 43210, USA*
[4] *Center for Cosmology and AstroParticle Physics (CCAPP), The Ohio State University, Columbus, OH 43210, USA*
[5] *Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA*
[6] *National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA*
[7] *Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong*
[8] *Indian Institute of Technology Patna, Bihta, Bihar 801106, India*
[9] *Department of Astronomy, MongManWai Building, Tsinghua University, Beijing 100084, China*
[10] *Institute of Particle and Nuclear Studies, High Energy Accelerator Research Organization, Tsukuba, Ibaraki 305-0801, Japan*
[11] *International Center for Quantum-field Measurement Systems for Studies of the Universe and Particles (QUP-WPI), High Energy Accelerator Research Organization (KEK), Tsukuba, Ibaraki 305-0801, Japan*
[12] *Computational Science Division, Argonne National Laboratory, Lemont, IL 60439, USA*
[13] *Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL 60439, USA*
[14] *Center for Astrophysics, Harvard & Smithsonian, Cambridge, MA 02138, USA*

ABSTRACT

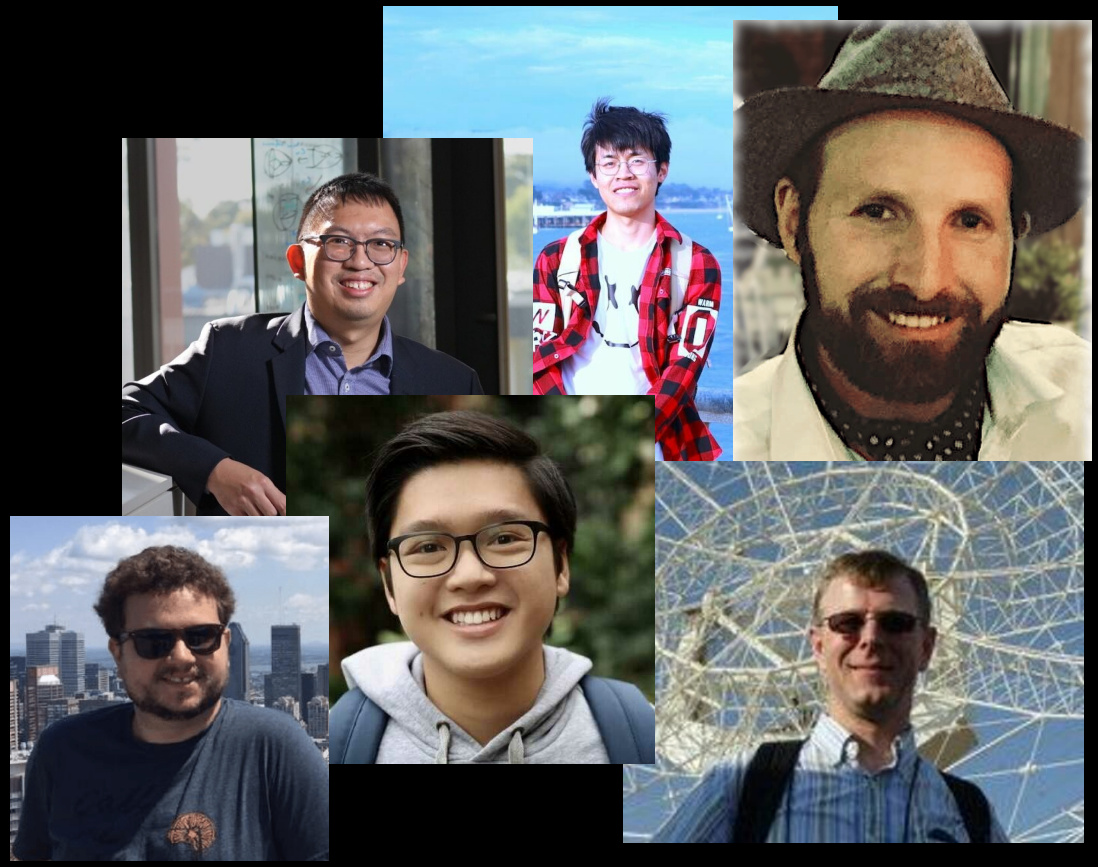
We present a comprehensive evaluation of proprietary and open-weights large language models (LLMs) using the first astronomy-specific benchmarking dataset. This dataset comprises 4,425 multiple-choice questions curated from the Annual Review of Astronomy and Astrophysics, covering a broad range of astrophysical topics. Our analysis examines model performance across various


$100,000
FINAL JEOPARDY!
$50,000
FINAL JEOPARDY!
$25,000
FINAL JEOPARDY!

AstroMLab (astromlab.org)



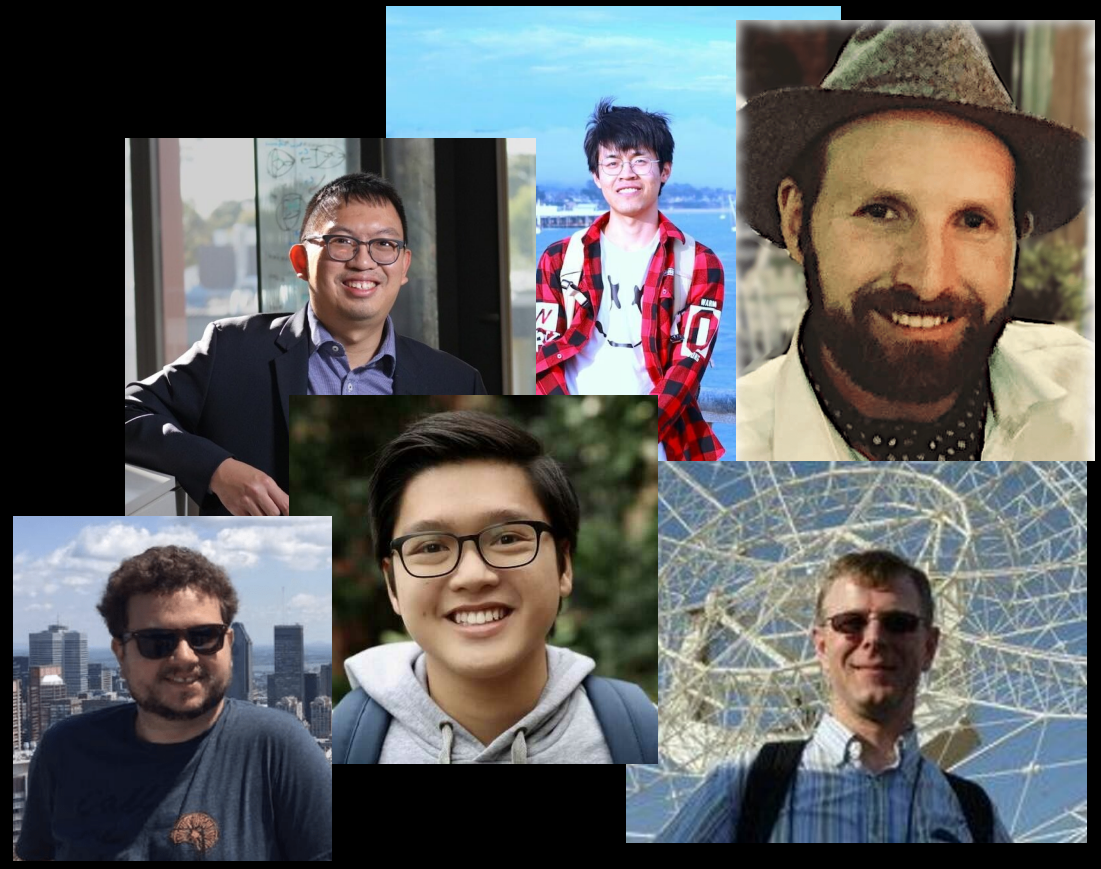
AstroMLab (astromlab.org)



Harvard-Smithsonian ADS

AstroMLab (astromlab.org)

Natural Language Processing experts



Harvard-Smithsonian ADS

AstroMLab (astromlab.org)

Natural Language Processing experts

Oak Ridge National Lab

Harvard-Smithsonian ADS

AstroMLab (astromlab.org)

Natural Language Processing experts

Oak Ridge National Lab

Argonne National Lab

Harvard-Smithsonian ADS

AstroMLab (astromlab.org)

Natural Language Processing experts

Oak Ridge National Lab

Argonne National Lab

U. Illinois Urbana-Champaign

Harvard-Smithsonian ADS

Curation of 5000 high quality astronomy QA *benchmark* dataset

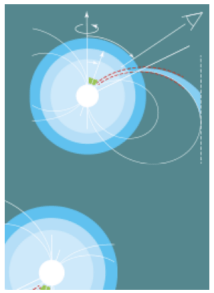
ANNUAL REVIEWS

Publications A-Z Journal Information

Home / A-Z Publications / Annual Review of Astronomy and Astrophysics

Annual Review of Astronomy and Astrophysics

Home | About | Current | Early Publication | Previous Volumes | Editorial Committee



ISSN: 0066-4146
eISSN: 1545-4282

AIMS AND SCOPE OF JOURNAL: The *Annual Review of Astronomy and Astrophysics* covers the significant developments in the field of astronomy and astrophysics, including: the sun; solar system and extrasolar planets; stars; the interstellar medium; galaxy and galaxies; active galactic nuclei; cosmology; and instrumentation and techniques, and the history of the development of new areas of research.

Published Since	1963
Journal Status	Active
DOI:	https://doi.org/10.1146/astro.683
Impact Factor	33.3

Current Volume is OA

Latest Articles

Argonne NATIONAL LABORATORY

Benchmark multiple choice question - *example*

What is the primary reason for the decline in the number density of luminous quasars at redshifts greater than 5?

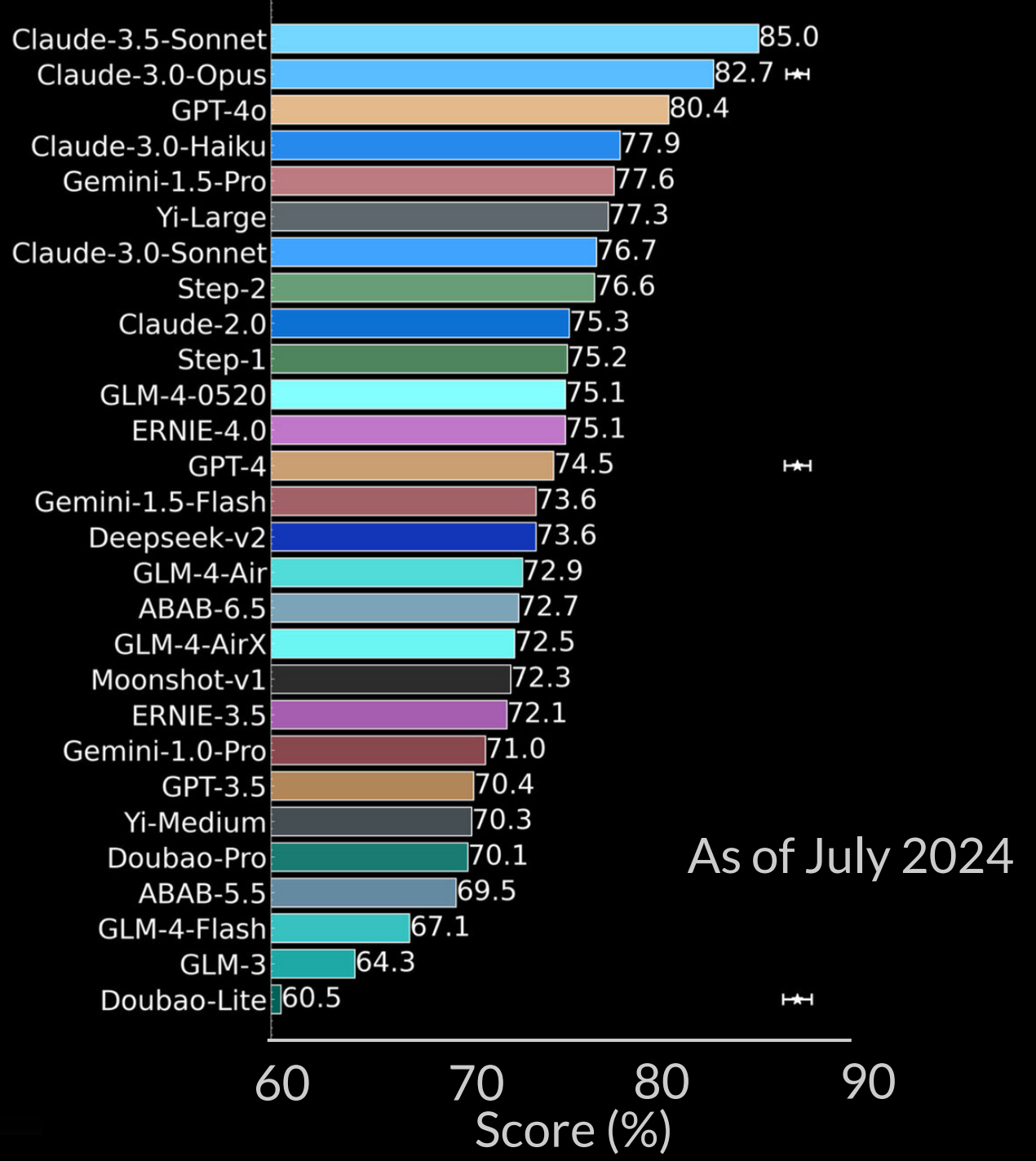
- A decrease in the overall star formation rate, leading to fewer potential host galaxies for quasars.
- A decrease in the number of massive black hole seeds that can form and grow into supermassive black holes.
- An increase in the neutral hydrogen fraction in the intergalactic medium, which obscures the quasars' light.
- An increase in the average metallicity of the Universe, leading to a decrease in the efficiency of black hole accretion.

Benchmark multiple choice question - *example*

What is the primary reason for the decline in the number density of luminous quasars at redshifts greater than 5?

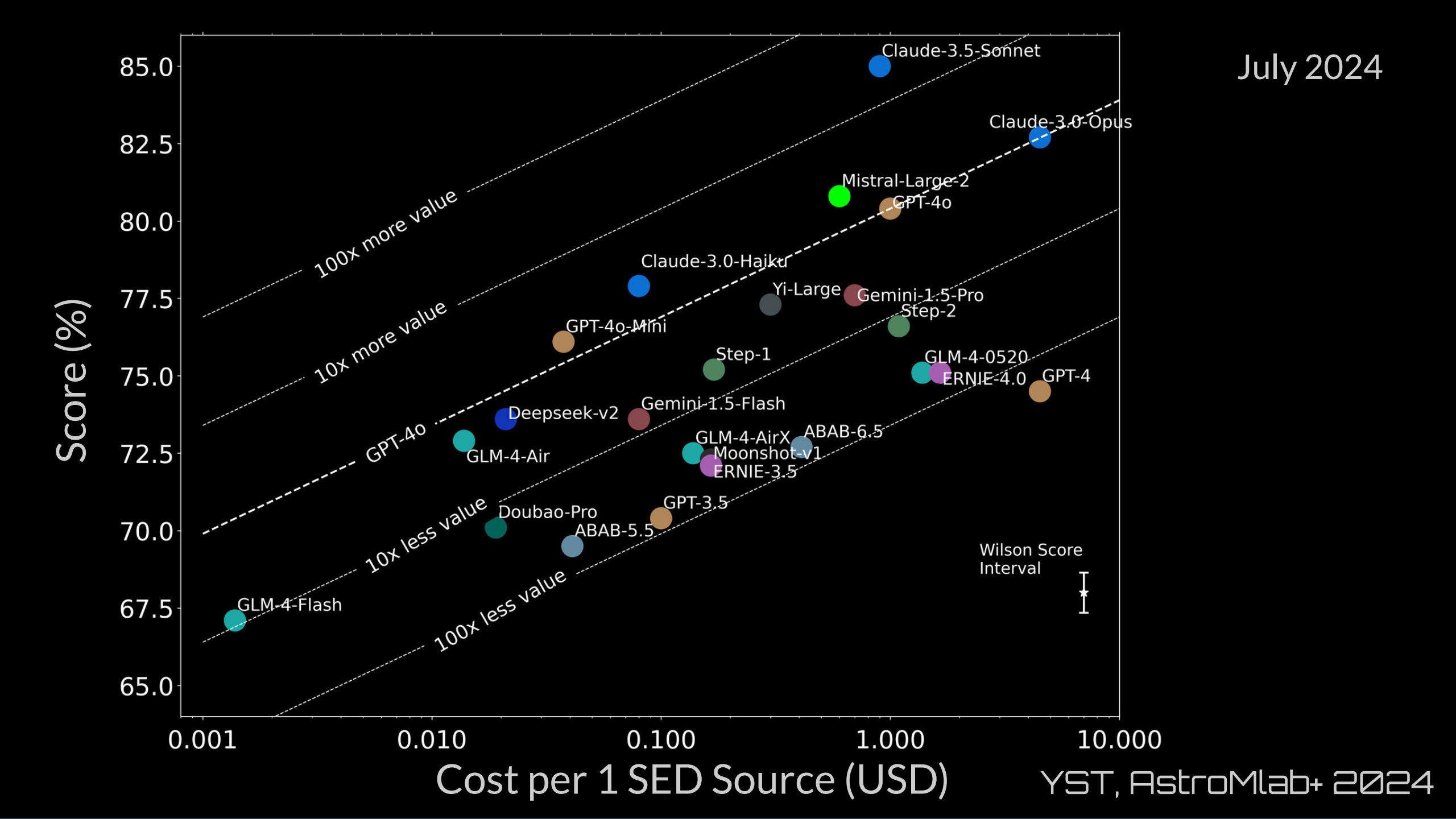
- A decrease in the overall star formation rate, leading to fewer potential host galaxies for quasars.
- A decrease in the number of massive black hole seeds that can form and grow into supermassive black holes.
- An increase in the neutral hydrogen fraction in the intergalactic medium, which obscures the quasars' light.
- An increase in the average metallicity of the Universe, leading to a decrease in the efficiency of black hole accretion.

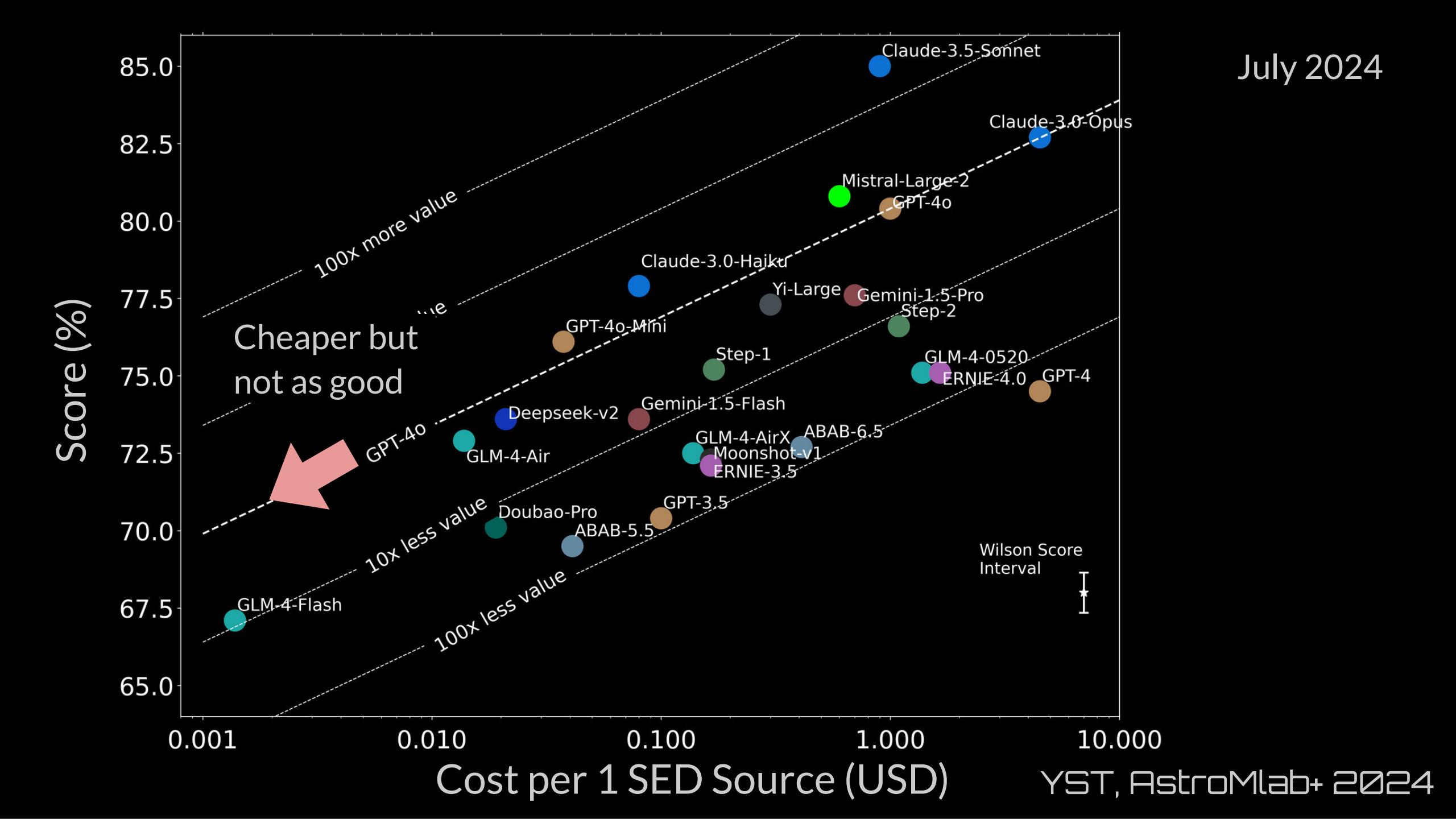
Proprietary models

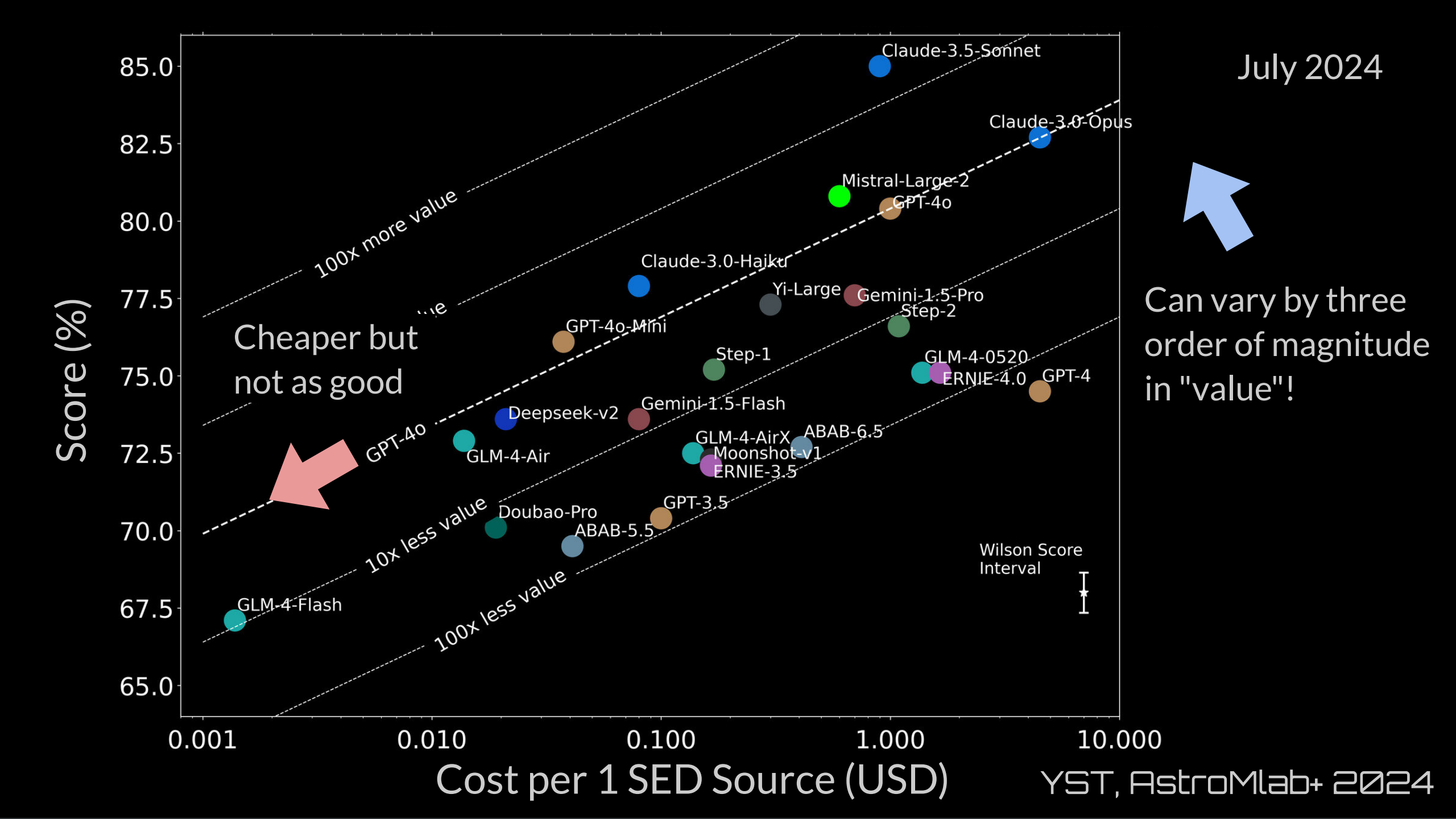


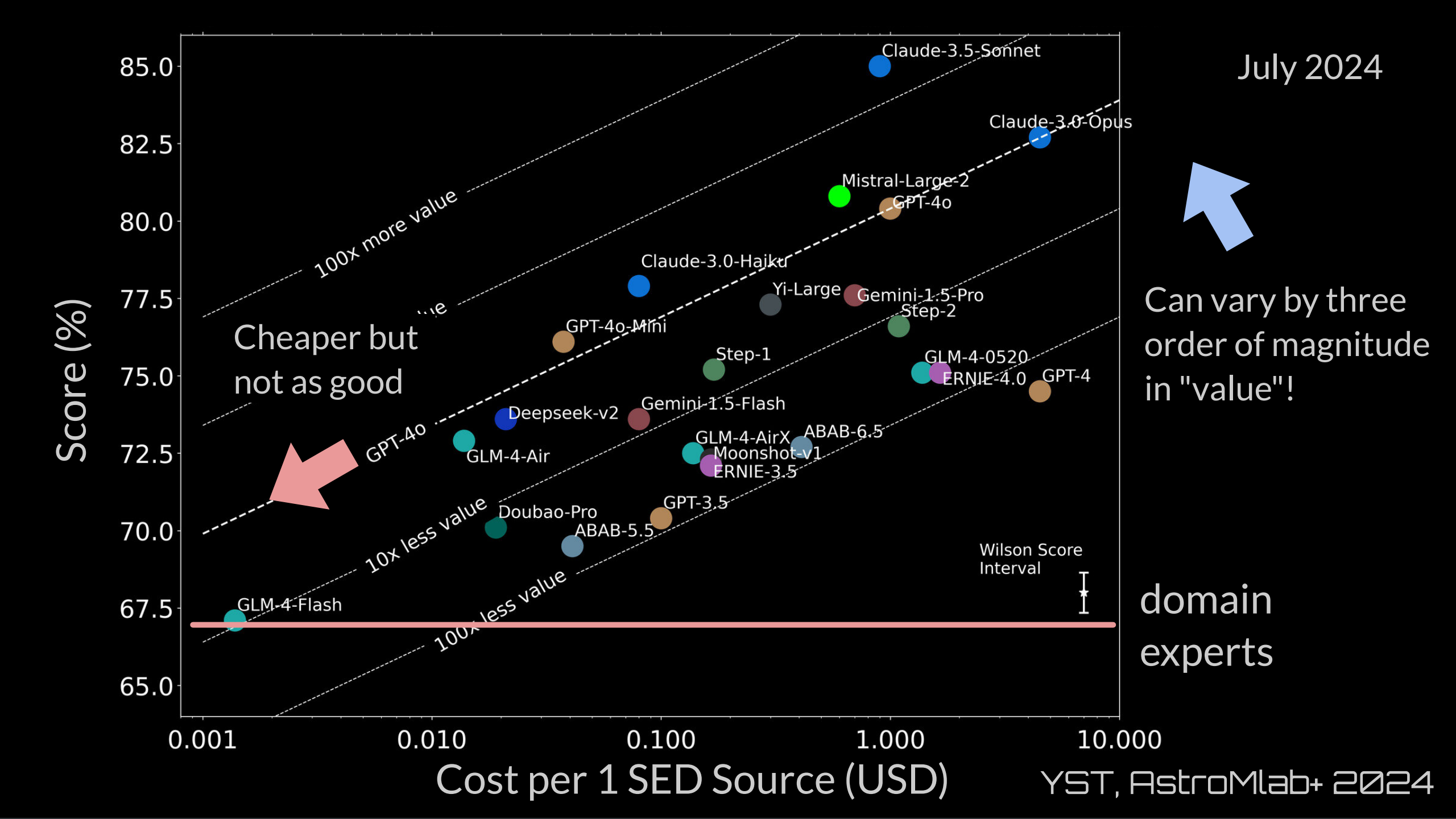
As of July 2024

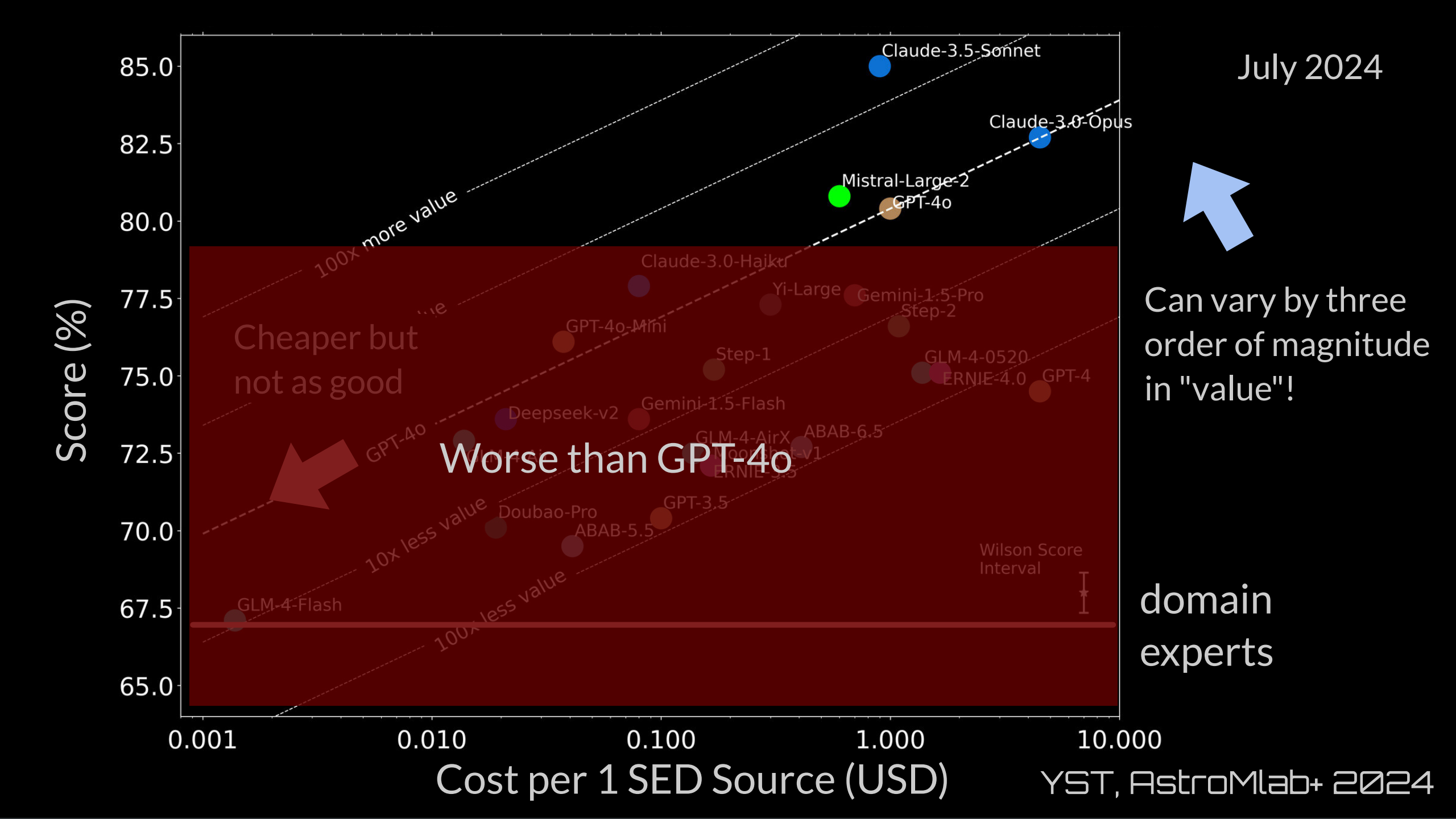
YST, AstroMlab+ 2024

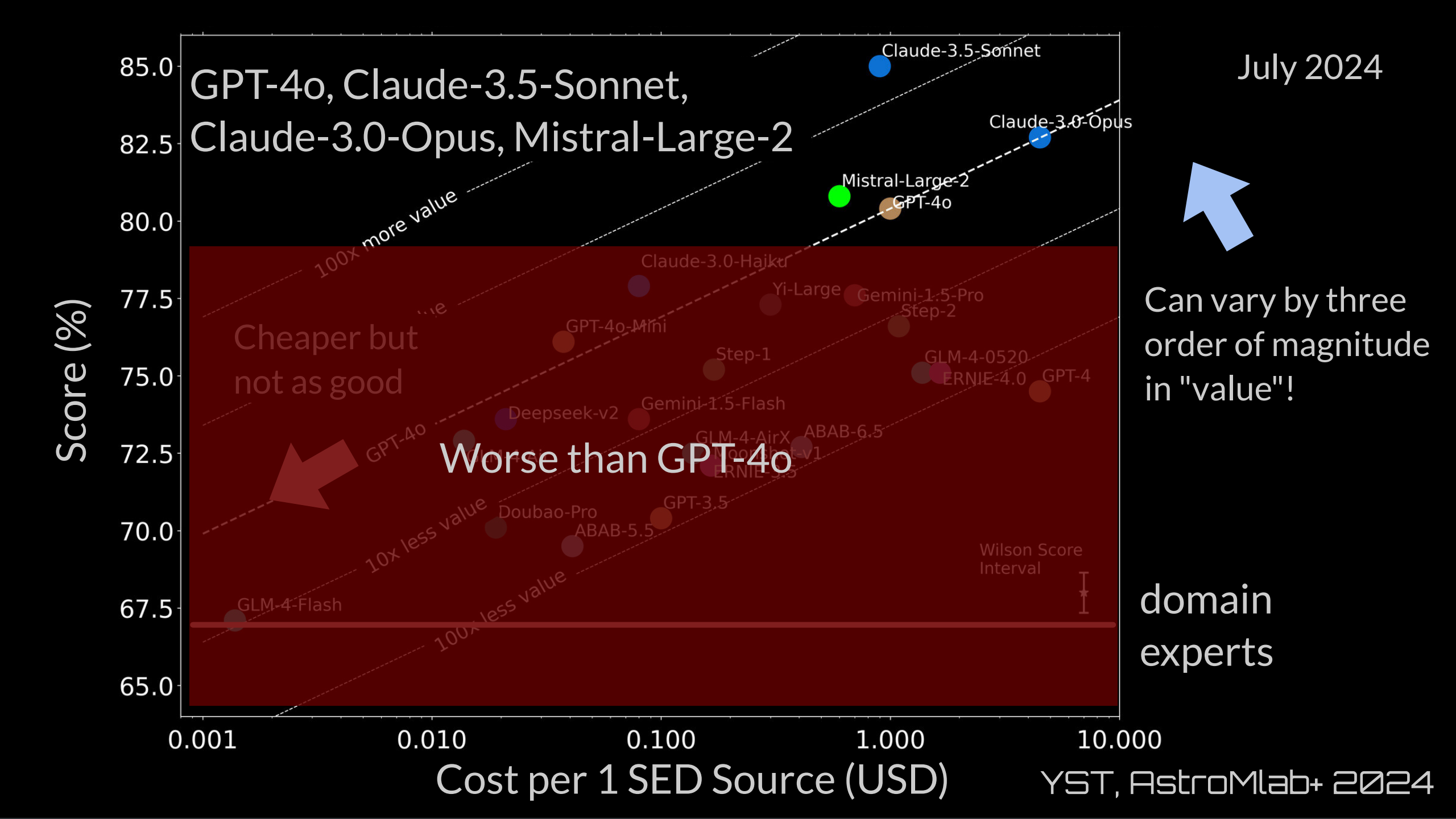

July 2024
Claude-3.5-Sonnet
Claude-3.0-Opus
Mistral-Large-2
GPT-4o
Claude-3.0-Haiku
Yi-Large
Gemini-1.5-Pro
Step-2
GPT-4o-Mini
Step-1
GLM-4-0520
ERNIE-4.0
GPT-4
Deepseek-v2
Gemini-1.5-Flash
GLM-4-AirX
ABAB-6.5
GLM-4-Air
Moonshot-v1
ERNIE-3.5
GPT-3.5
Doubao-Pro
ABAB-5.5
GLM-4-Flash
100x more value
10x more value
GPT-4o
10x less value
100x less value
Wilson Score Interval
Score (%)
85.0
82.5
80.0
77.5
75.0
72.5
70.0
67.5
65.0
0.001
0.010
0.100
1.000
10.000
Cost per 1 SED Source (USD)
YST, AstroMlab+ 2024

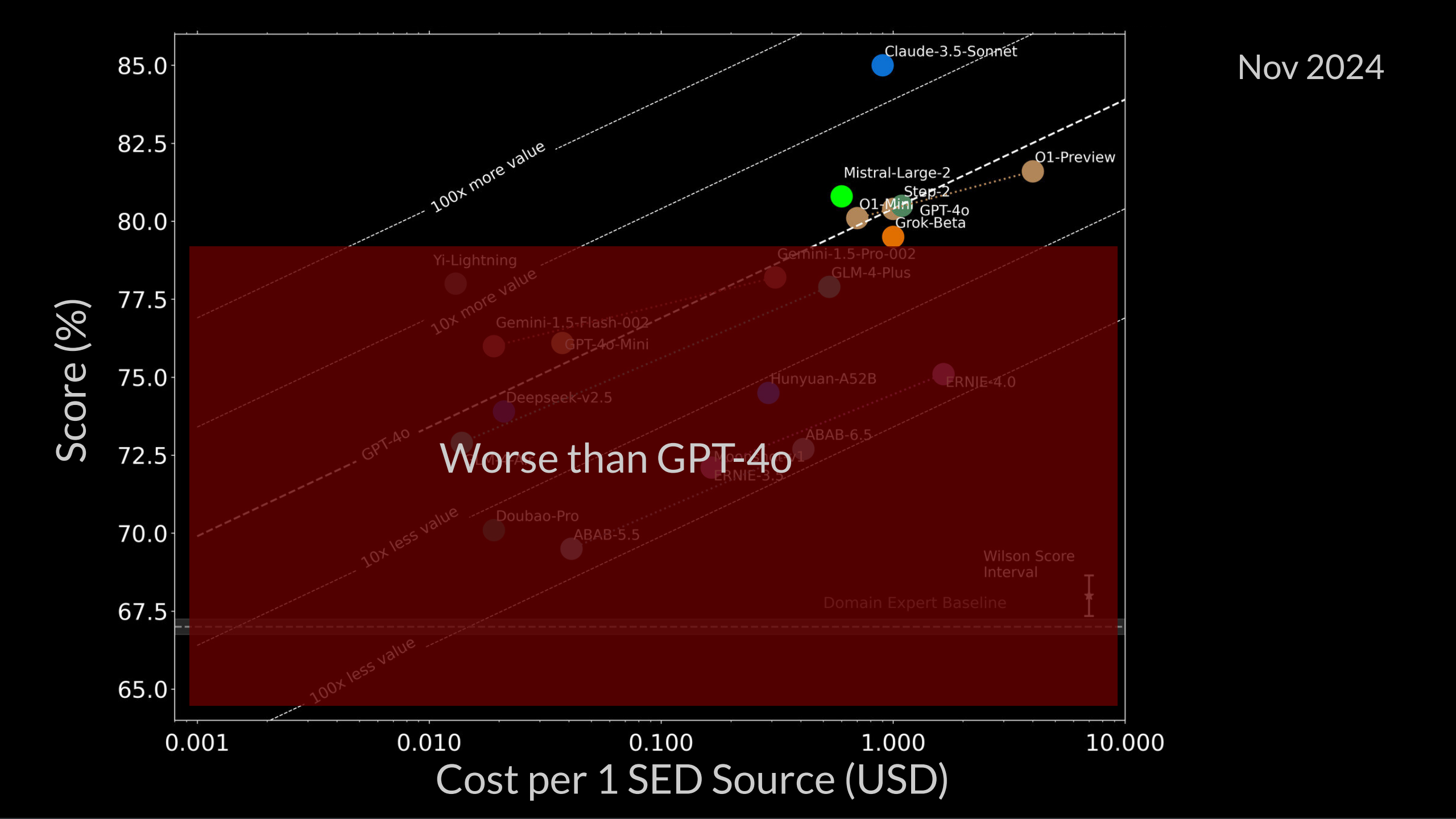

July 2024
Score (%)
85.0
82.5
80.0
77.5
75.0
72.5
70.0
67.5
65.0
0.001
0.010
0.100
1.000
10.000
Cost per 1 SED Source (USD)
100x more value
GPT-4o
10x less value
100x less value
Cheaper but
not as good
Claude-3.5-Sonnet
Claude-3.0-Opus
Mistral-Large-2
GPT-4o
Claude-3.0-Haiku
Yi-Large
Gemini-1.5-Pro
Step-2
GPT-4o-Mini
Step-1
GLM-4-0520
ERNIE-4.0
GPT-4
Deepseek-v2
Gemini-1.5-Flash
GLM-4-Air
GLM-4-AirX
ABAB-6.5
Moonshot-v1
ERNIE-3.5
GPT-3.5
Doubao-Pro
ABAB-5.5
GLM-4-Flash
Wilson Score
Interval
YST, AstroMlab+ 2024

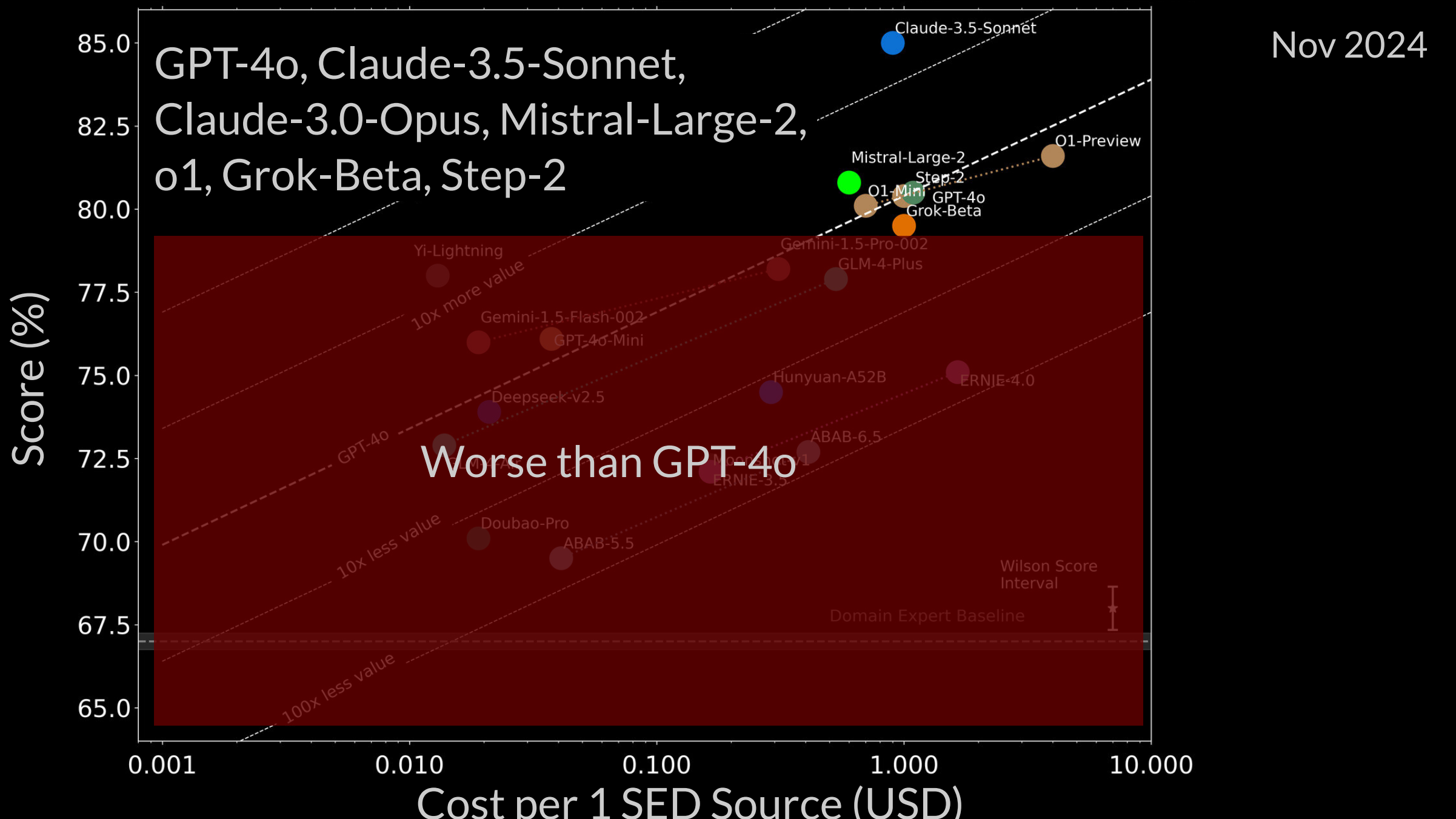

July 2024
Claude-3.5-Sonnet
Claude-3.0-Opus
Mistral-Large-2
GPT-4o
Claude-3.0-Haiku
Yi-Large
Gemini-1.5-Pro
Step-2
GPT-4o-Mini
Step-1
GLM-4-0520
ERNIE-4.0
GPT-4
Deepseek-v2
Gemini-1.5-Flash
GLM-4-AirX
ABAB-6.5
GLM-4-Air
Moonshot-v1
ERNIE-3.5
GPT-3.5
Doubao-Pro
ABAB-5.5
GLM-4-Flash
100x more value
GPT-4o
10x less value
100x less value
Cheaper but not as good
Can vary by three order of magnitude in "value"!
Wilson Score Interval
Score (%)
85.0
82.5
80.0
77.5
75.0
72.5
70.0
67.5
65.0
0.001
0.010
0.100
1.000
10.000
Cost per 1 SED Source (USD)
YST, AstroMlab+ 2024


July 2024
Claude-3.5-Sonnet
Claude-3.0-Opus
Mistral-Large-2
GPT-4o
Claude-3.0-Haiku
Yi-Large
Gemini-1.5-Pro
Step-2
GPT-4o-Mini
Step-1
GLM-4-0520
ERNIE-4.0
GPT-4
Deepseek-v2
Gemini-1.5-Flash
GLM-4-Air
GLM-4-AirX
ABAB-6.5
Moonshot-v1
ERNIE-3.5
GPT-3.5
Doubao-Pro
ABAB-5.5
GLM-4-Flash
100x more value
GPT-4o
10x less value
100x less value
Cheaper but not as good
Can vary by three order of magnitude in "value"!
Wilson Score Interval
domain experts
Score (%)
85.0
82.5
80.0
77.5
75.0
72.5
70.0
67.5
65.0
0.001
0.010
0.100
1.000
10.000
Cost per 1 SED Source (USD)
YST, AstroMlab+ 2024


July 2024
Claude-3.5-Sonnet
Claude-3.0-Opus
Mistral-Large-2
GPT-4o
100x more value
Claude-3.0-Haiku
Yi-Large
Gemini-1.5-Pro
Step-2
GPT-4o-Mini
Step-1
GLM-4-0520
ERNIE-4.0
GPT-4
Cheaper but
not as good
Deepseek-v2
Gemini-1.5-Flash
GPT-4o
GLM-4-AirX
ABAB-6.5
Worse than GPT-4o
GPT-3.5
Doubao-Pro
ABAB-5.5
10x less value
Wilson Score
Interval
GLM-4-Flash
100x less value
Can vary by three
order of magnitude
in "value"!
domain
experts
Score (%)
85.0
82.5
80.0
77.5
75.0
72.5
70.0
67.5
65.0
0.001
0.010
0.100
1.000
10.000
Cost per 1 SED Source (USD)
YST, AstroMlab+ 2024

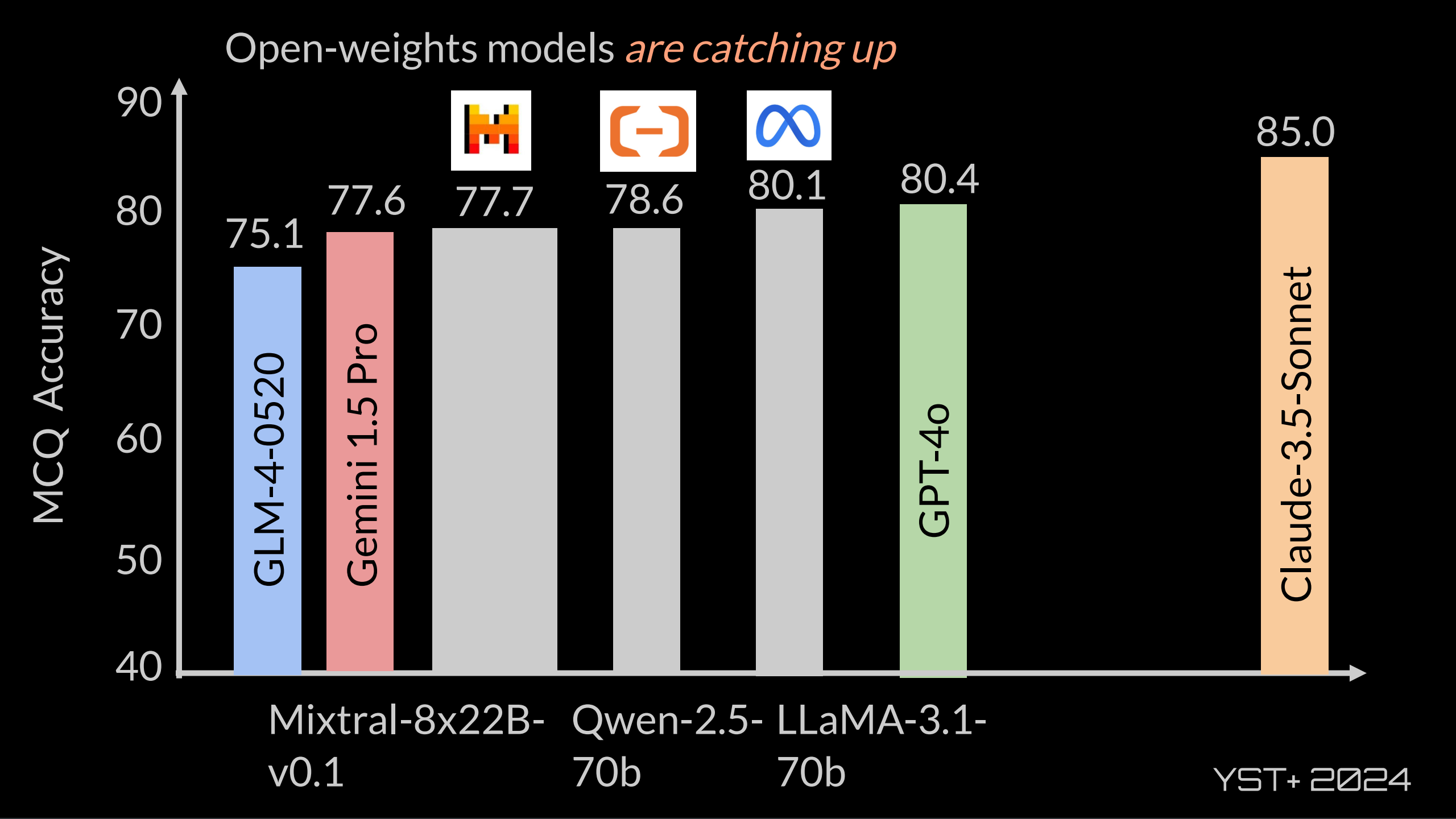

GPT-4o, Claude-3.5-Sonnet,
Claude-3.0-Opus, Mistral-Large-2
July 2024
Claude-3.5-Sonnet
Claude-3.0-Opus
Mistral-Large-2
GPT-4o
100x more value
Claude-3.0-Haiku
Yi-Large
Gemini-1.5-Pro
Step-2
GPT-4o-Mini
Step-1
GLM-4-0520
ERNIE-4.0
GPT-4
Cheaper but
not as good
Deepseek-v2
Gemini-1.5-Flash
GPT-4o
GLM-4-AirX
ABAB-6.5
Worse than GPT-4o
10x less value
Doubao-Pro
GPT-3.5
ABAB-5.5
Wilson Score
Interval
GLM-4-Flash
100x less value
Can vary by three
order of magnitude
in "value"!
domain
experts
Score (%)
85.0
82.5
80.0
77.5
75.0
72.5
70.0
67.5
65.0
0.001
0.010
0.100
1.000
10.000
Cost per 1 SED Source (USD)
YST, AstroMlab+ 2024

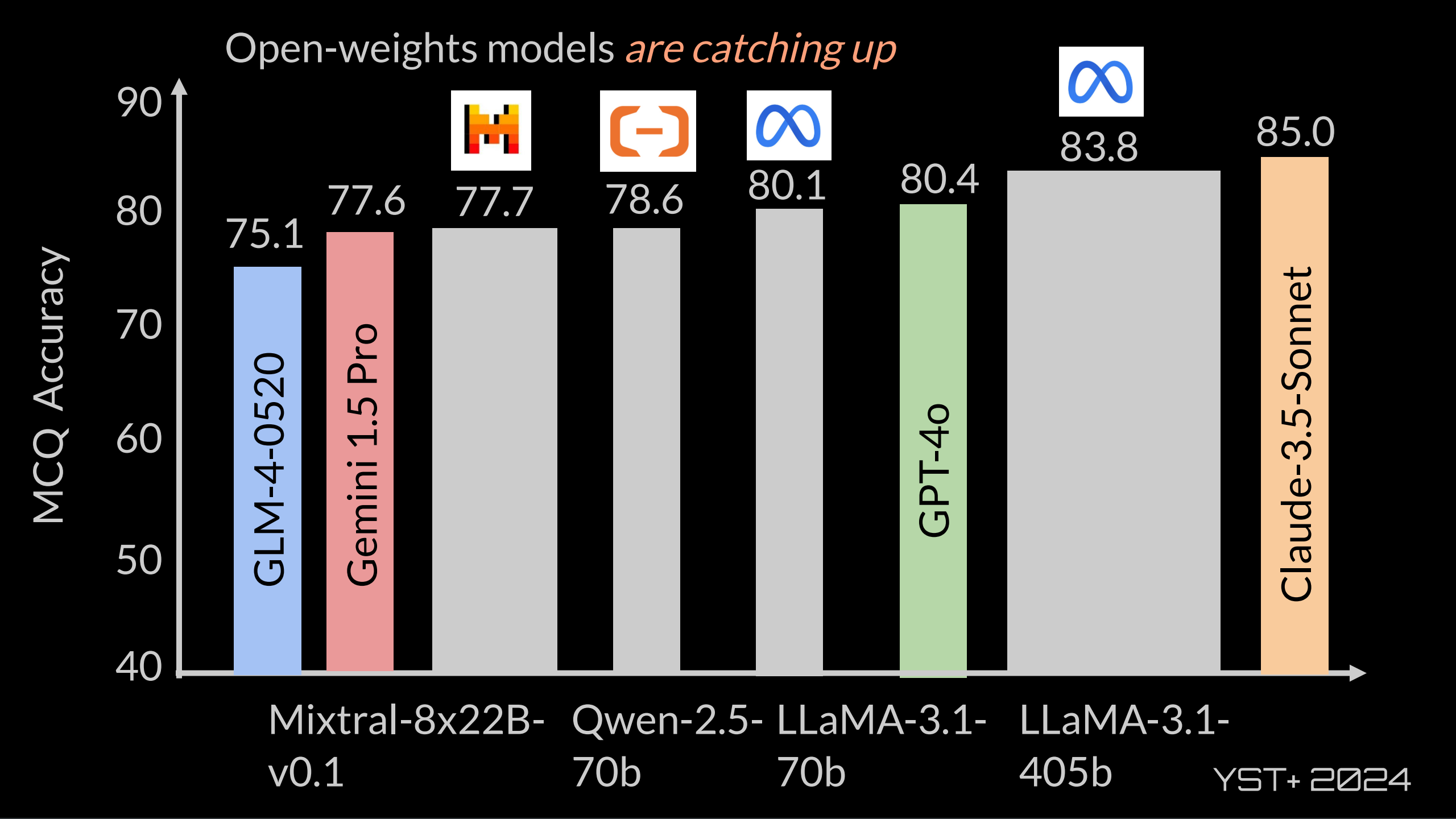

Nov 2024
Score (%)
85.0
82.5
80.0
77.5
75.0
72.5
70.0
67.5
65.0
0.001
0.010
0.100
1.000
10.000
Cost per 1 SED Source (USD)
Claude-3.5-Sonnet
O1-Preview
Mistral-Large-2
Step-2
O1-Mini
GPT-4o
Grok-Beta
Yi-Lightning
Gemini-1.5-Pro-002
GLM-4-Plus
Gemini-1.5-Flash-002
GPT-4o-Mini
Hunyuan-A52B
ERNIE-4.0
Deepseek-v2.5
ABAB-6.5
ERNIE-3.5
Doubao-Pro
ABAB-5.5
100x more value
10x more value
GPT-4o
10x less value
100x less value
Worse than GPT-4o
Wilson Score Interval
Domain Expert Baseline


GPT-4o, Claude-3.5-Sonnet,
Claude-3.0-Opus, Mistral-Large-2,
o1, Grok-Beta, Step-2
Nov 2024
Claude-3.5-Sonnet
O1-Preview
Mistral-Large-2
Step-2
O1-Mini
GPT-4o
Grok-Beta
Yi-Lightning
Gemini-1.5-Pro-002
GLM-4-Plus
10x more value
Gemini-1.5-Flash-002
GPT-4o-Mini
Hunyuan-A52B
ERNIE-4.0
Deepseek-v2.5
GPT-4o
ABAB-6.5
ERNIE-3.5
Doubao-Pro
10x less value
ABAB-5.5
Wilson Score
Interval
Domain Expert Baseline
100x less value
Worse than GPT-4o
Score (%)
85.0
82.5
80.0
77.5
75.0
72.5
70.0
67.5
65.0
0.001
0.010
0.100
1.000
10.000
Cost per 1 SED Source (USD)

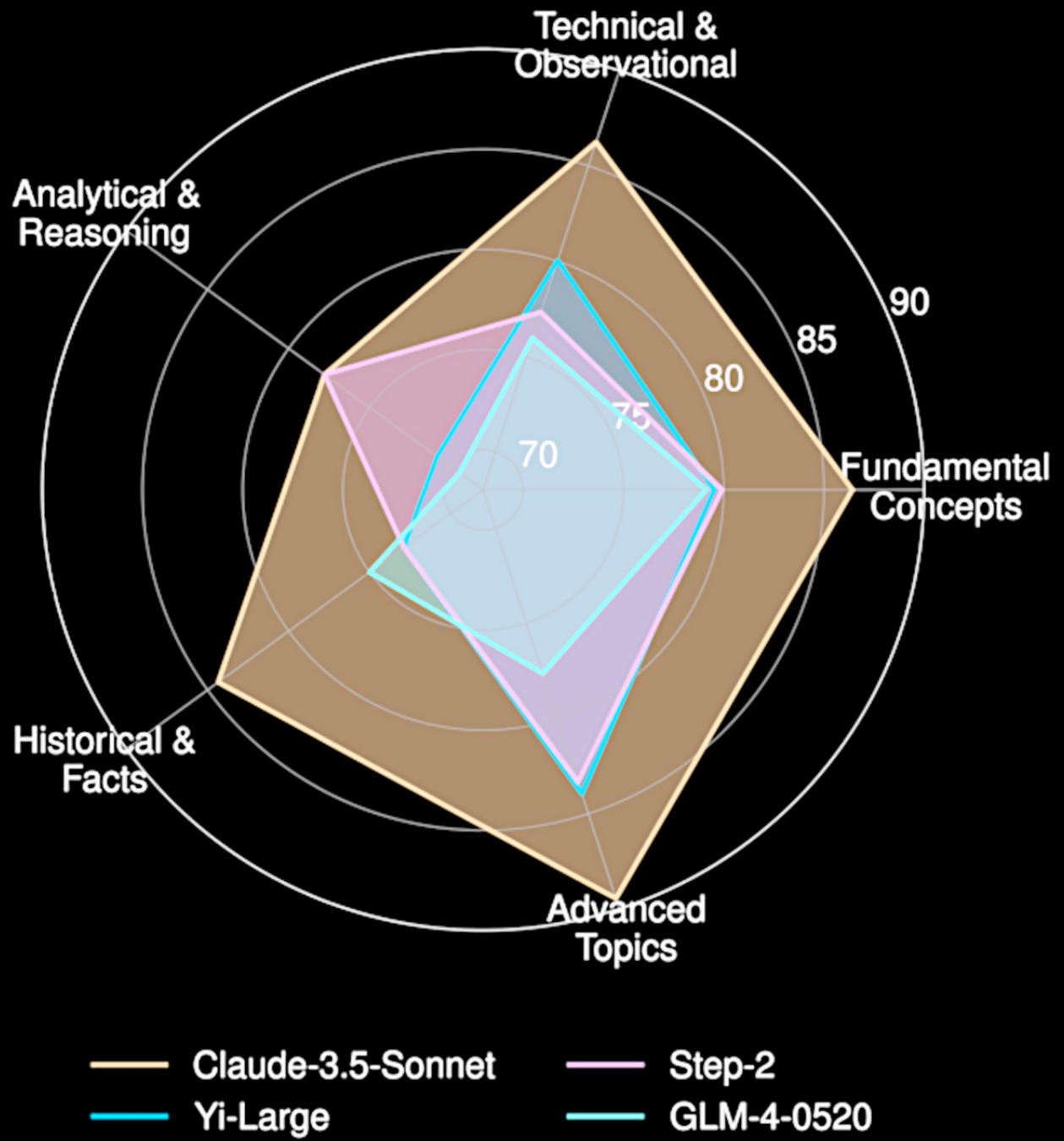
Open-weights large language models?

In academic settings, it's still easier to *secure GPU time* than to get grants for LLMs.

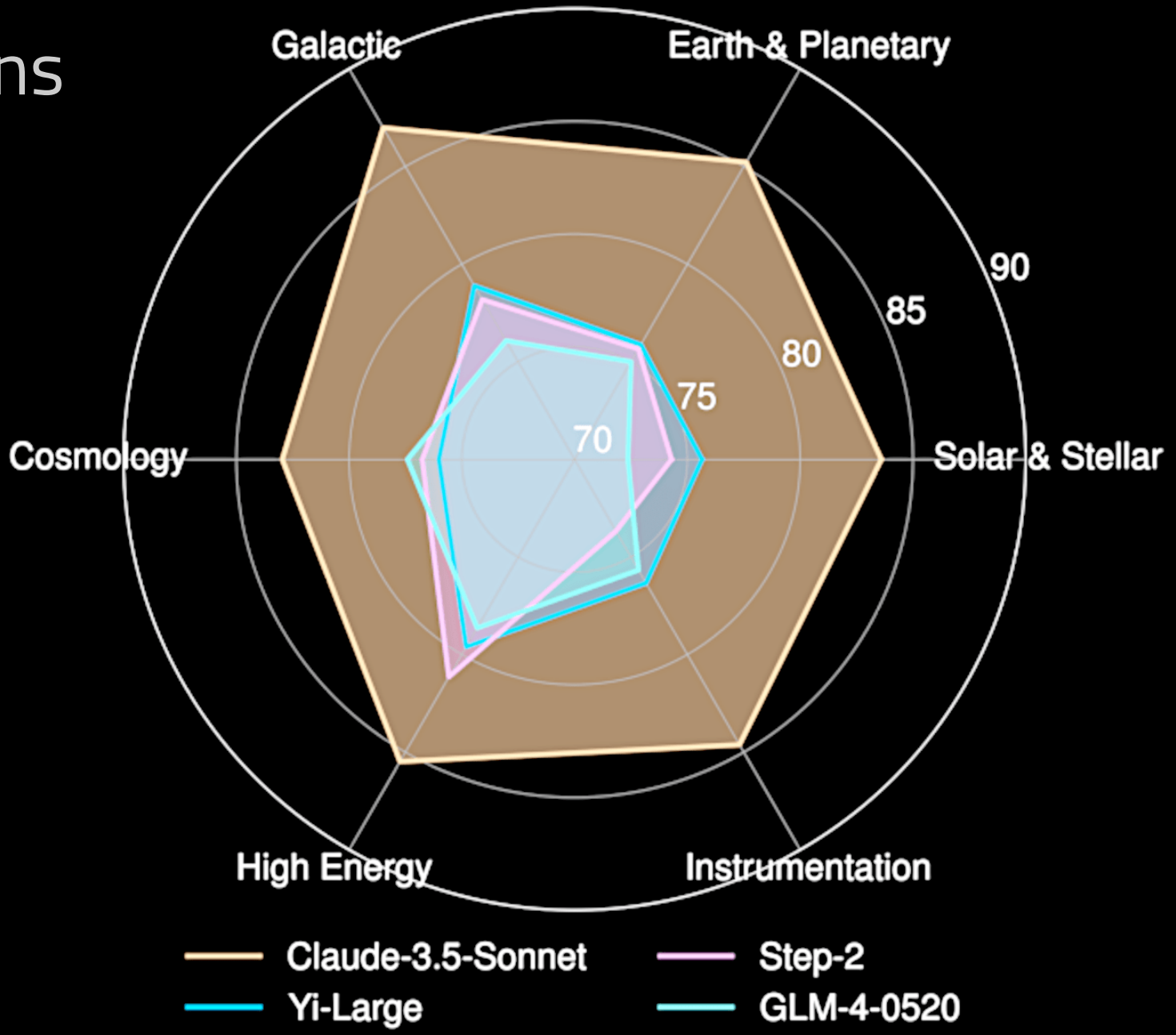

Open-weights models *are catching up*
MCQ Accuracy
90
80
70
60
50
40
75.1
77.6
77.7
78.6
80.1
80.4
85.0
GLM-4-0520
Gemini 1.5 Pro
Mixtral-8x22B-v0.1
Qwen-2.5-70b
LLaMA-3.1-70b
GPT-4o
Claude-3.5-Sonnet
YST+ 2024


Open-weights models *are catching up*
MCQ Accuracy
90
80
70
60
50
40
75.1
77.6
77.7
78.6
80.1
80.4
83.8
85.0
GLM-4-0520
Gemini 1.5 Pro
GPT-4o
Claude-3.5-Sonnet
Mixtral-8x22B-v0.1
Qwen-2.5-70b
LLaMA-3.1-70b
LLaMA-3.1-405b
YST+ 2024

Identifying the limitations

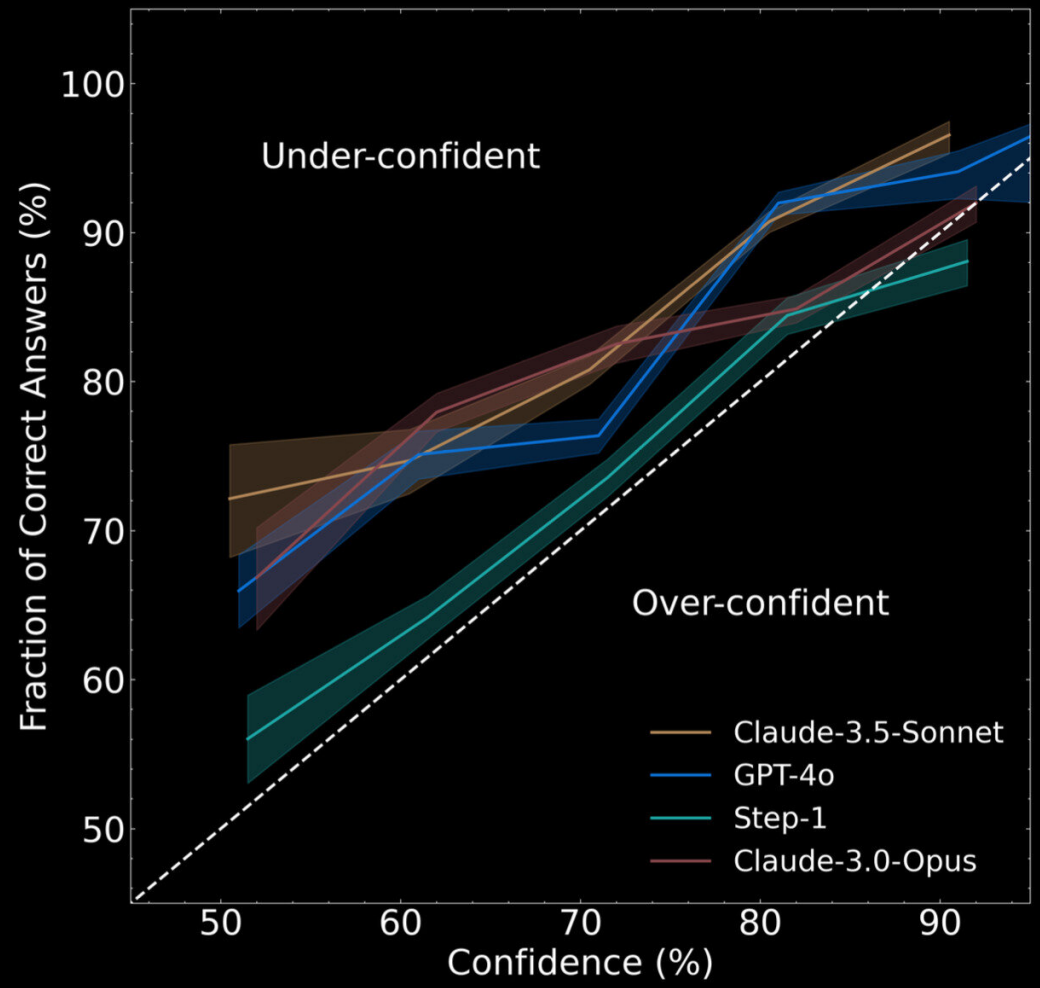


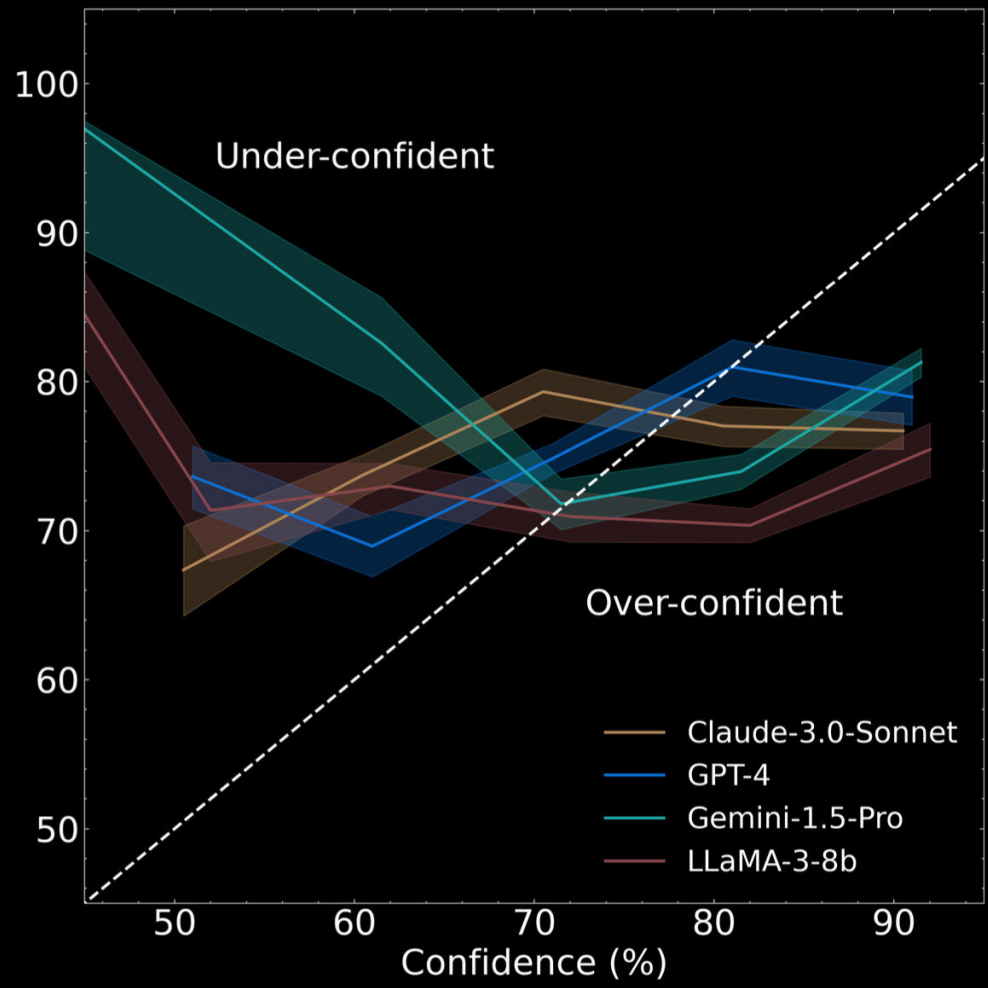
Identifying the limitations



YST, AstroMlab+ 2024

Trustworthiness : Are you sure?





Open-weight large language models
are *as good as* the proprietary models

Open-weight large language models
are *as good as* the proprietary models

at the ~70B level
and beyond

Still it is *not very scalable*

Still it is *not very scalable*

LLaMA-3.1 70b throughput on *four H100 GPUs*

= ~ 100 tokens / second

LLaMA-3.1 70b throughput on *four H100 GPUs*

= ~ 100 tokens / second

1 SED source = 15 GPU minutes

LLaMA-3.1 70b throughput on *four H100 GPUs*

= ~ 100 tokens / second

1 SED source = 15 GPU minutes

1B sources = 10M GPU days

LLaMA-3.1 70b throughput on *four H100 GPUs*

= ~ 100 tokens / second

1 SED source = 15 GPU minutes

1B sources = 10M GPU days

A cluster with *10,000 H100 GPUs*
running for *3 years*

Can we improve *lightweight (e.g., 8B)* open-weights LLMs to perform well on astronomical tasks?

AstroMLab 3: Achieving GPT-4o Level Performance in Astronomy with a Specialized 8B-Parameter Large Language Model

Tijmen de Haan[1], Yuan-Sen Ting (丁源森)[2,3], Tirthankar Ghosal[4], Tuan Dung Nguyen[5], Alberto Accomazzi[6], Azton Wells[7], Nesar Ramachandra[7], Rui Pan (潘瑞)[8], Zechang Sun (孙泽昌)[9]

[1] *International Center for Quantum-field Measurement Systems for Studies of the Universe and Particles (QUP-WPI), High Energy Accelerator Research Organization (KEK), Tsukuba, Ibaraki, Japan*

[2] *Department of Astronomy, The Ohio State University, Columbus, OH, USA*

[3] *Center for Cosmology and AstroParticle Physics (CCAPP), The Ohio State University, Columbus, OH, USA*

[4] *National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA*

[5] *Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA*

[6] *Center for Astrophysics, Harvard & Smithsonian, Cambridge, MA, USA*

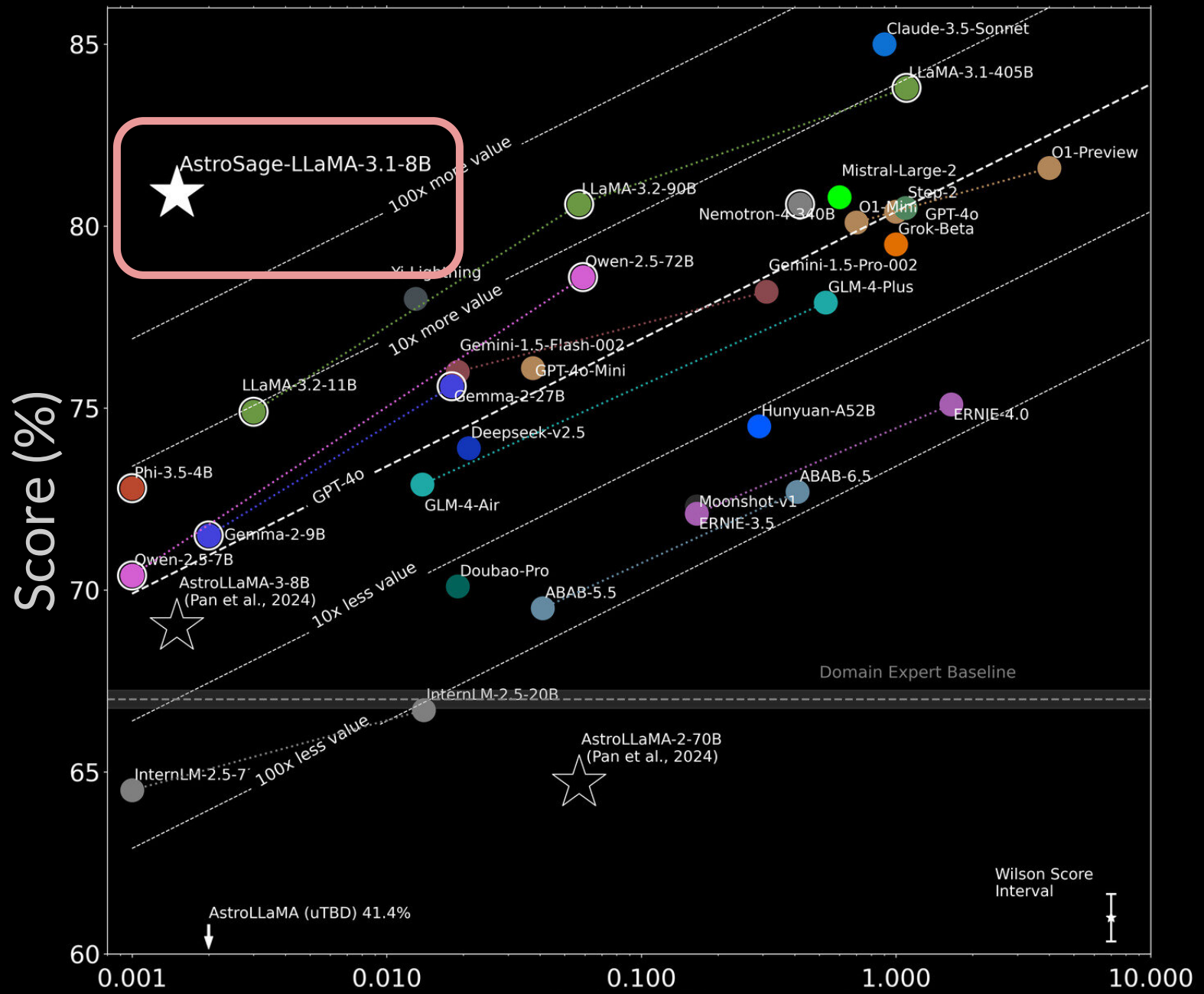
[7] *Computational Science Division, Argonne National Laboratory, Lemont, IL, USA*

[8] *Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong and*

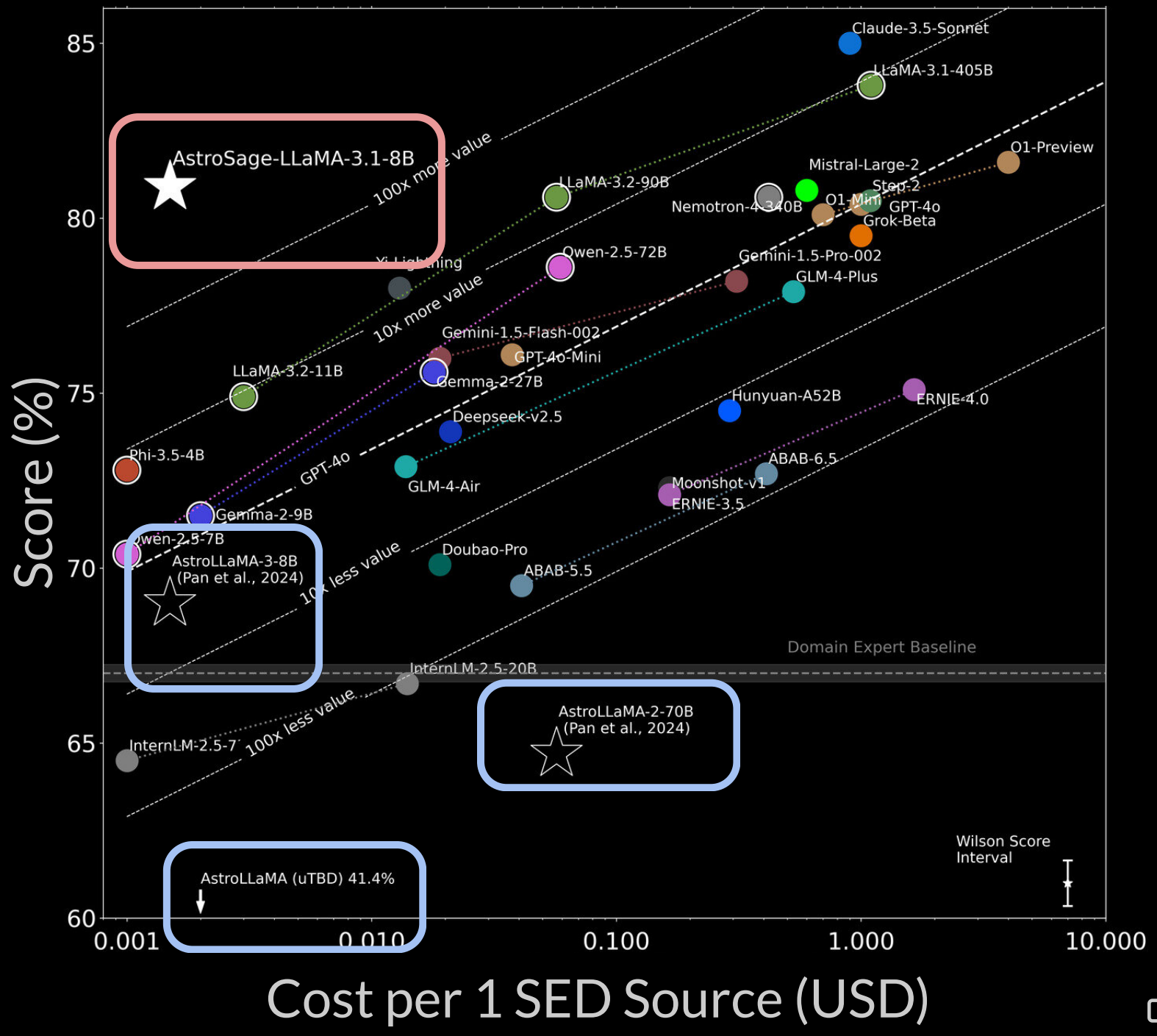
[9] *Department of Astronomy, Tsinghua University, Beijing, People's Republic of China*

AstroSage-Llama-3.1-8B is a domain-specialized natural-language AI assistant tailored for research in astronomy, astrophysics, and cosmology. Trained on the complete collection of astronomy-related arXiv papers from 2007-2024 along with millions of synthetically-generated question-answer

Special thanks to

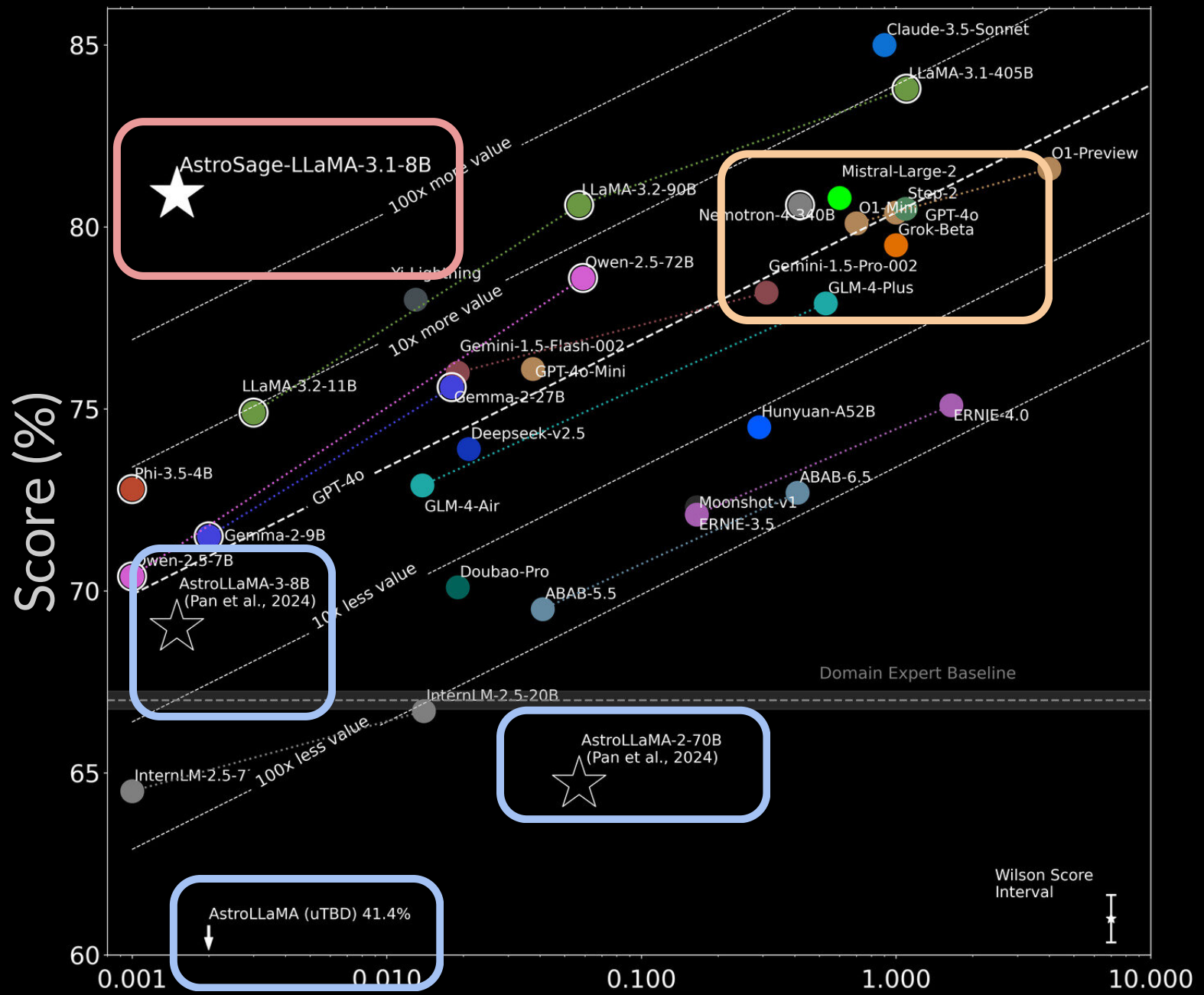


de Haan, AstroMlab+ 2024



The first specialized LLM in astronomy that outperforms its original baseline general model

de Haan, AstroMlab+ 2024



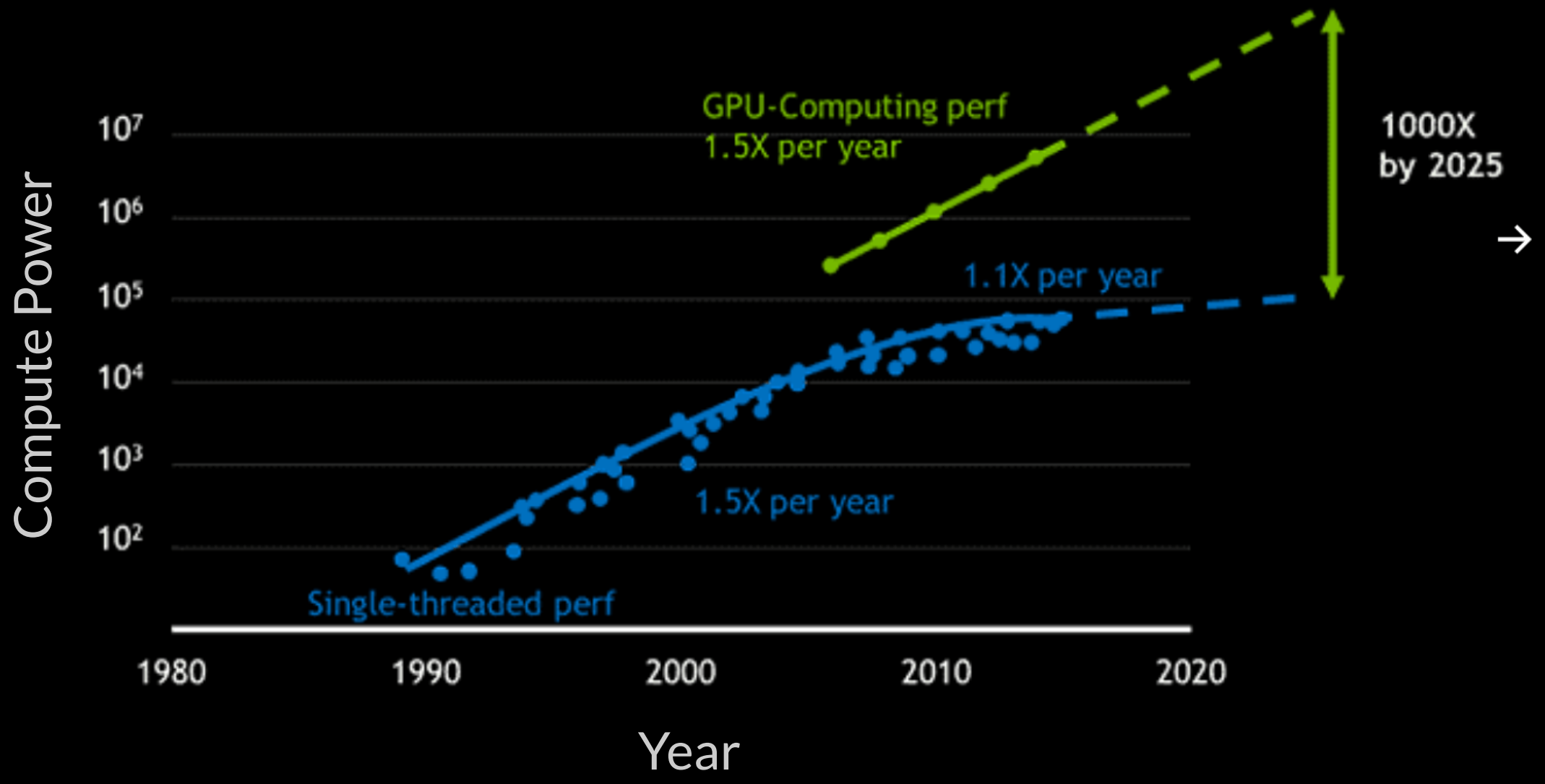
Achieve same score *as GPT-4o* in astronomy Q&A

The first specialized LLM in astronomy that outperforms its original baseline general model

de Haan, AstroMlab+ 2024

For *individual specific* downstream task,
it is possible to train a super-performant
lightweight LLMs.

Huang's Law


GPU-Computing perf
1.5X per year
1000X
by 2025
1.1X per year
1.5X per year
Single-threaded perf
Compute Power
10^7
10^6
10^5
10^4
10^3
10^2
1980
1990
2000
2010
2020
Year

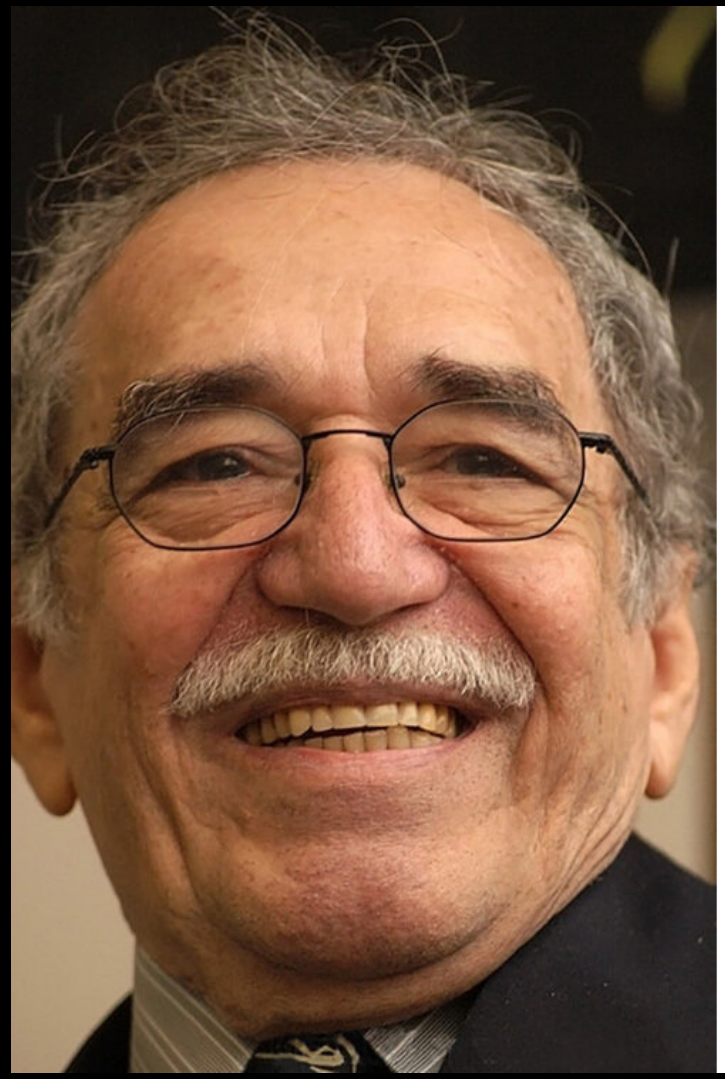
Tech / Tech Trends

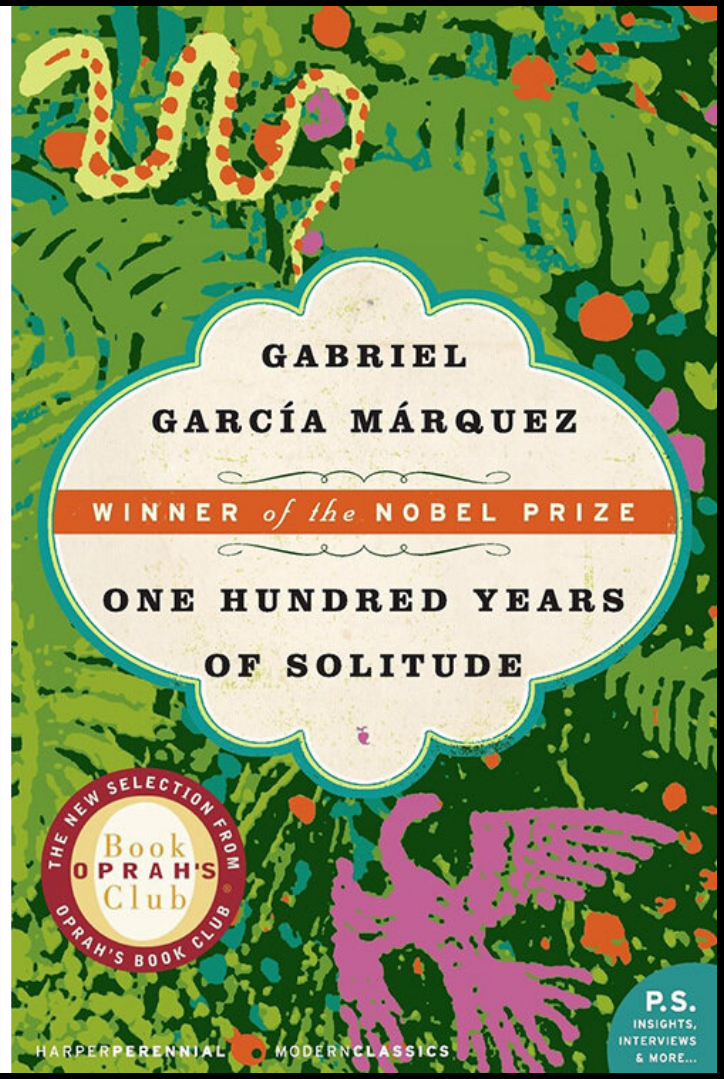
Tech unicorn Zhipu AI joins China's LLM price war amid new funding round

- Zhipu AI's GLM series of large language models now costs 90 per cent less than the current industry average of 1 yuan per 1 million tokens

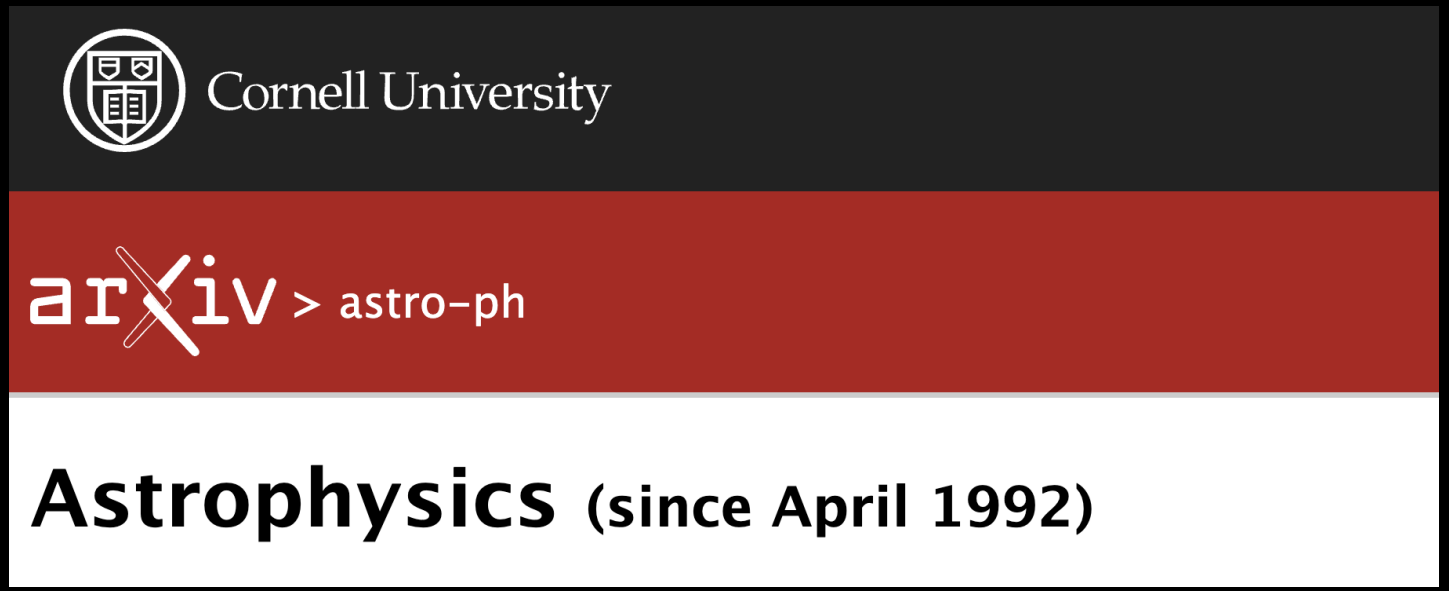
Ben Jiang in Beijing + FOLLOW

Published: 8:00pm, 5 Jun 2024

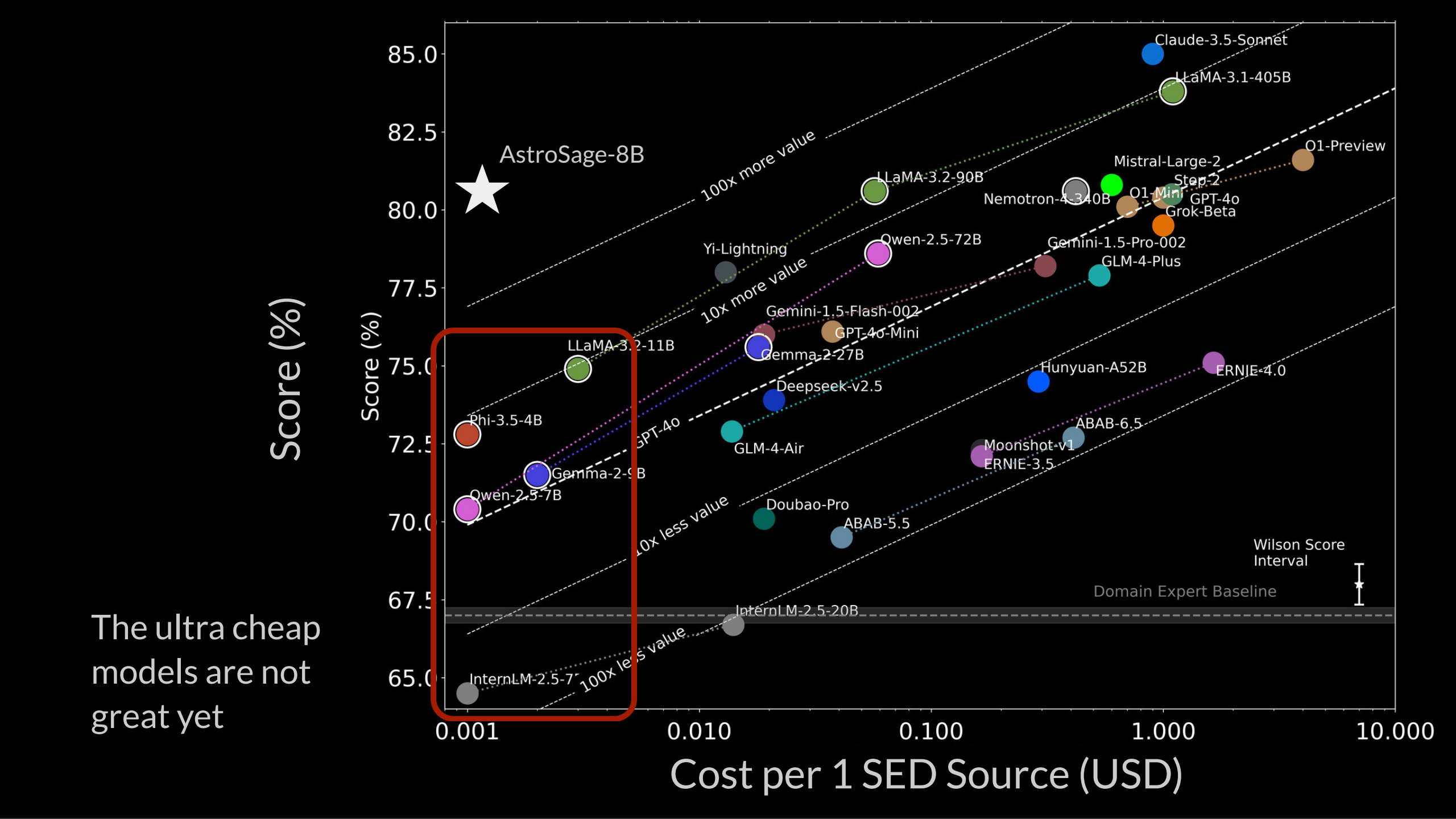



GABRIEL
GARCÍA MÁRQUEZ
WINNER of the NOBEL PRIZE
ONE HUNDRED YEARS
OF SOLITUDE
THE NEW SELECTION FROM
OPRAH'S BOOK CLUB
Book
OPRAH'S
Club
HARPERPERENNIAL MODERNCLASSICS
P.S.
INSIGHTS,
INTERVIEWS
& MORE...

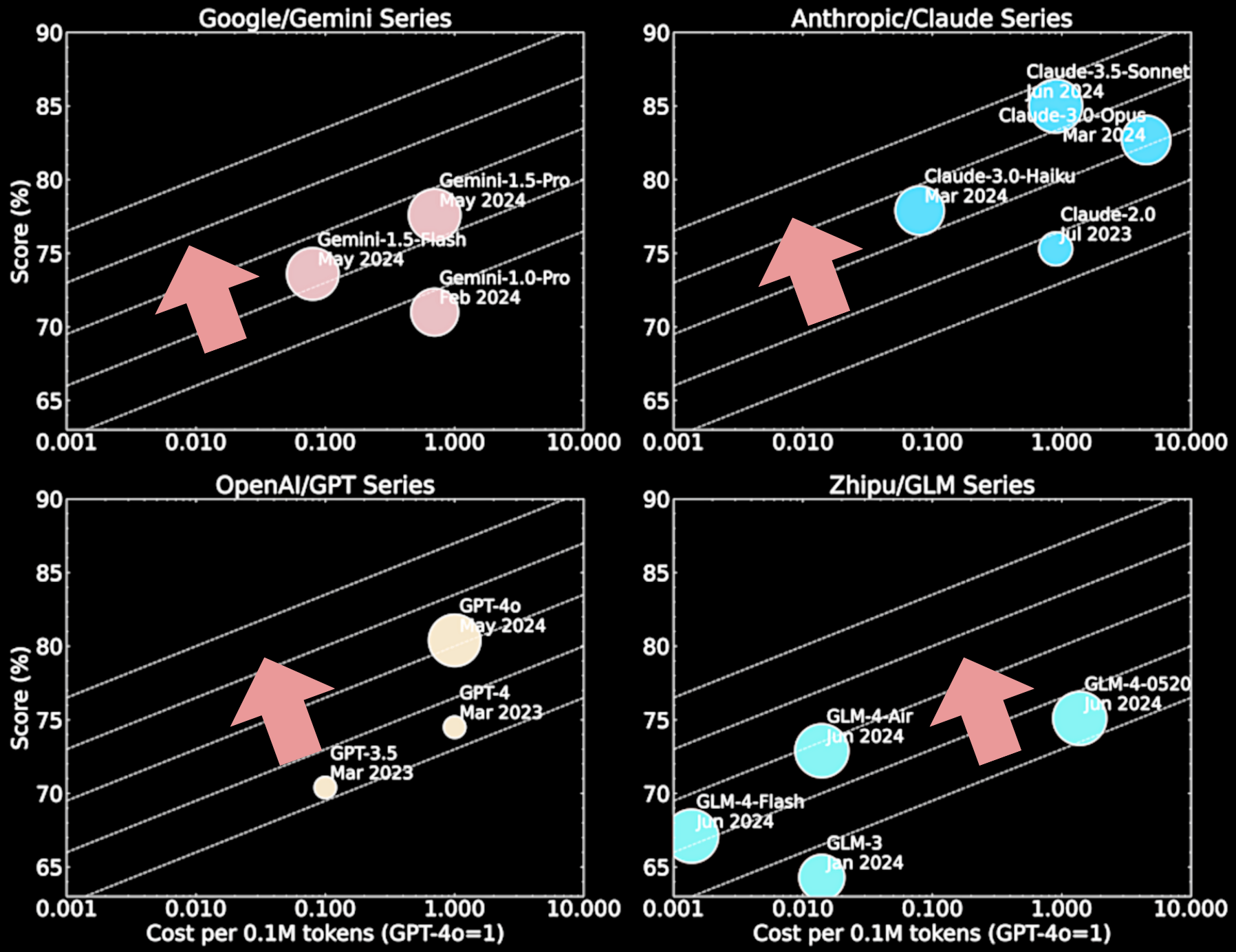
= 0.03 USD


Cornell University
arXiv > astro-ph
Astrophysics (since April 1992)

= 40 USD



The ultra cheap models are not great yet



The price drop has an e-folding time of *approximately 3-6 months*

Data-poor , Theory-rich

???

Collecting more data

Data-poor , Theory-rich

Data-poor , Theory-rich

Data-poor , Theory-rich

Data-rich , Theory-poor


RADNOR
FIRST AID KIT

Data-poor , Theory-rich

三个臭皮匠胜过一个诸葛亮

Millions of LLM agents

"Three cobblers with their wits combined surpass one Zhuge Liang (~Machiavelli) " - Chinese proverb