Expediting Astronomical Discovery with Large Language Models

Yuan-Sen Ting (丁源森)

Australian National University
The Ohio State University

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

Biological Sciences

~ 2 centers

Physical Sciences

~ 1 + (1+1) centers

Environmental Sciences

~ 2 centers

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

Biological Sciences

~ 2 centers

Physical Sciences

~ 1 + (1+1) centers

Environmental Sciences

~ 2 centers

6 centers x

15M ~ 100M

Hype, myth, or *real deal*?

nature

Explore content About the journal Publish with us

nature > articles > article

Article | Open access | Published: 15 July 2021

Highly accurate protein structure prediction with AlphaFold

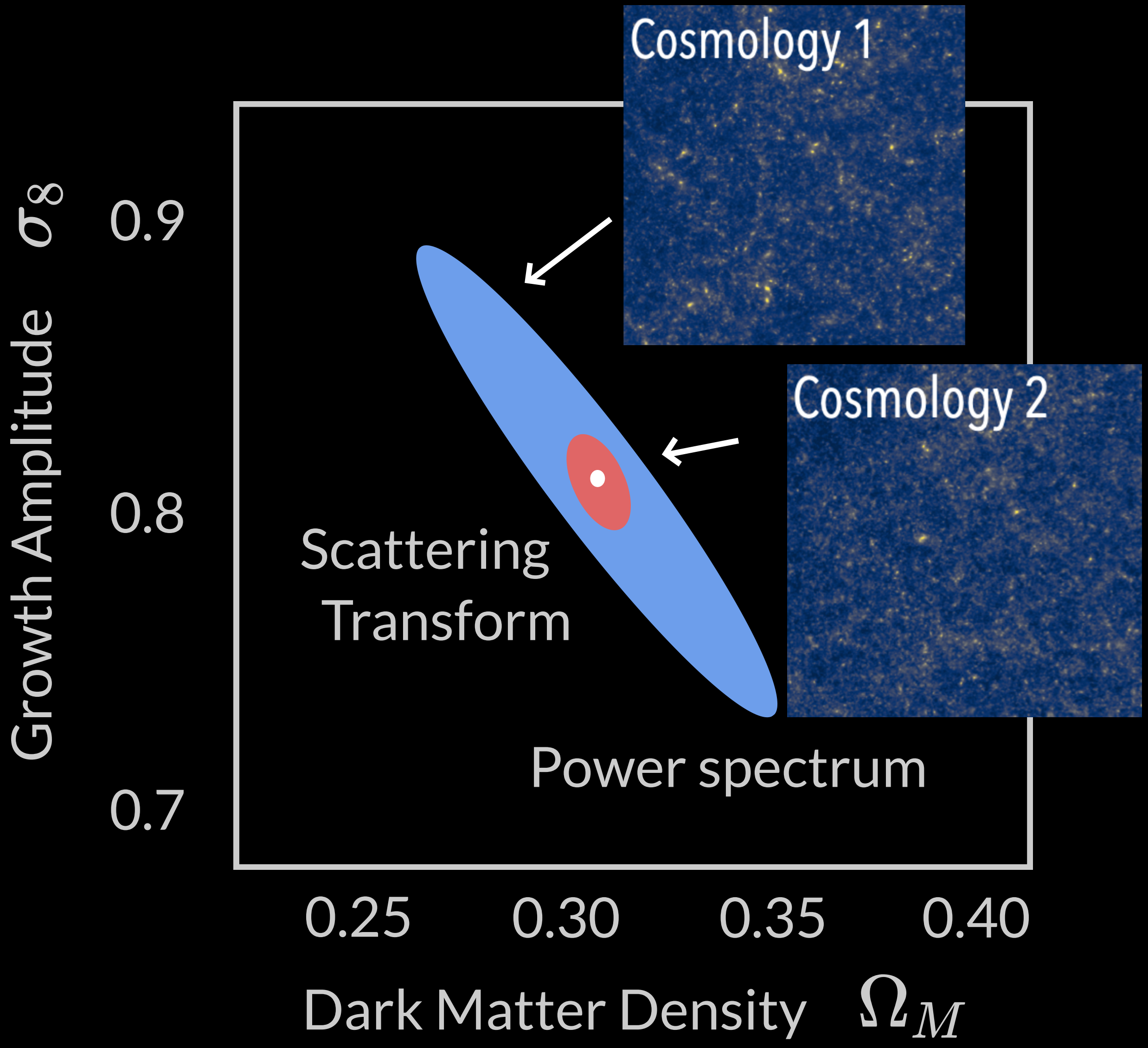
John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain,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

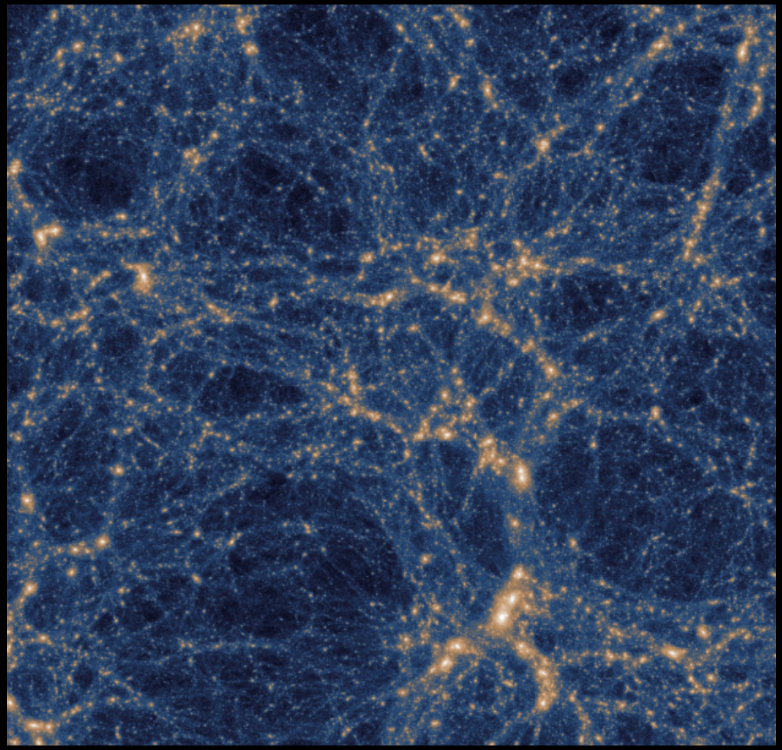
+ Show authors

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

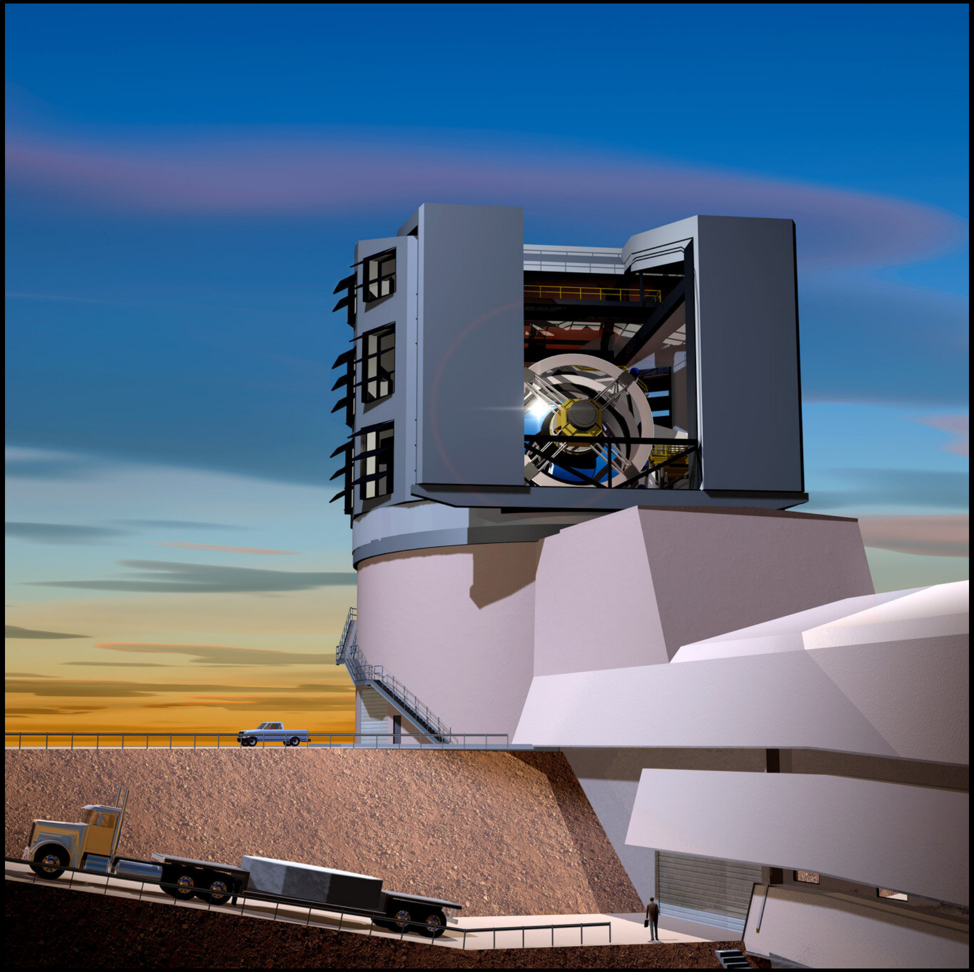




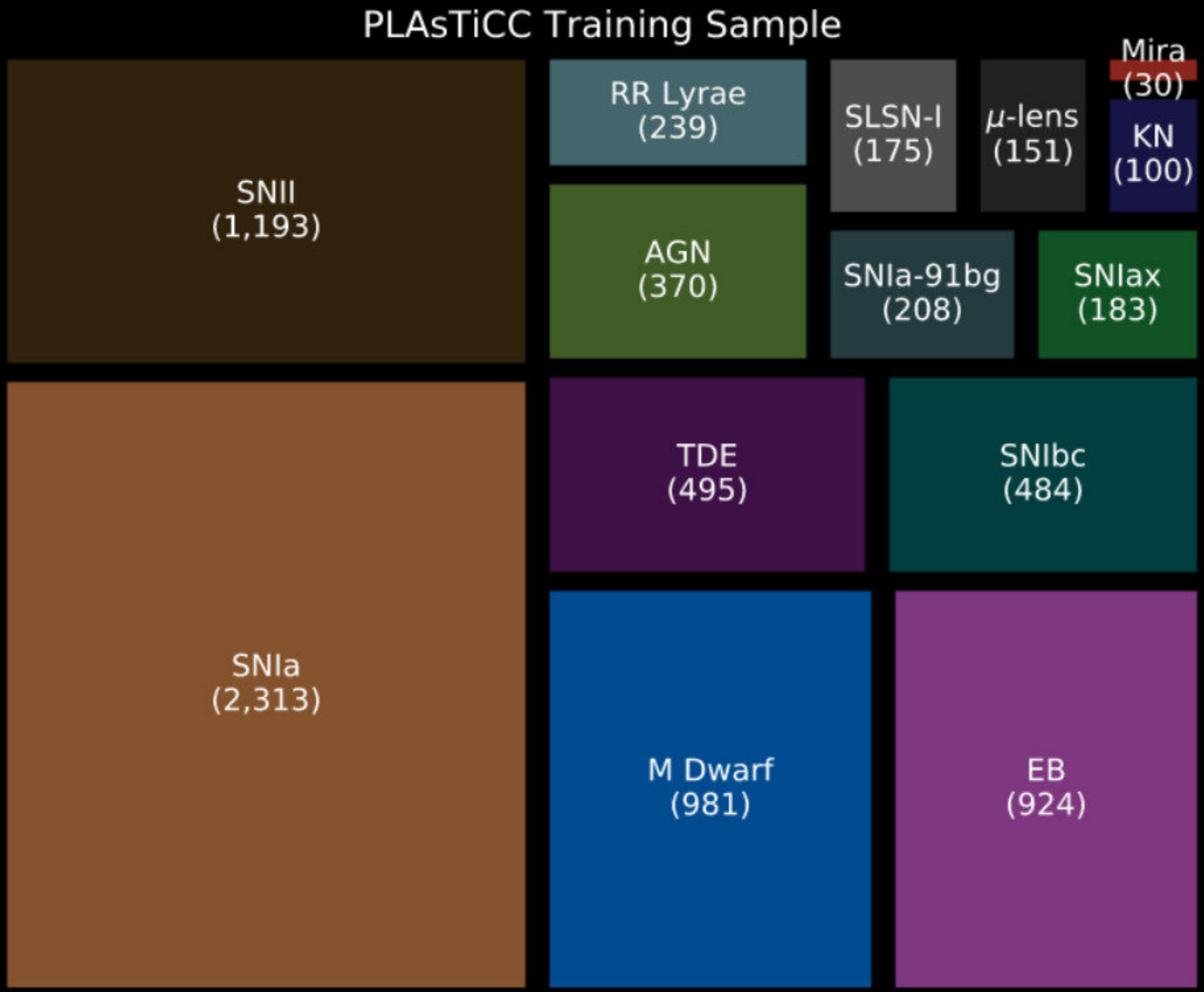
International Astrostatistics Association Award

Cheng, YST, Menard & Bruna + 20

or building effective brokers / *classifiers*



Rubin Observatory



Improving individual downstream tasks with AI will *not* revolutionize astronomy

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."

- Where are we in AI for Astronomy - What's the problem?
- What could be the AlphaFold moment for astronomy?
- How do we get there in a cost-effective way?

- Where are we in AI for Astronomy - What's the problem?
- What could be the AlphaFold moment for astronomy?
- How do we get there in a cost-effective way?

Knowledge Graph in Astronomical Research with Large Language Models: Quantifying Driving Forces in Interdisciplinary Scientific Discovery

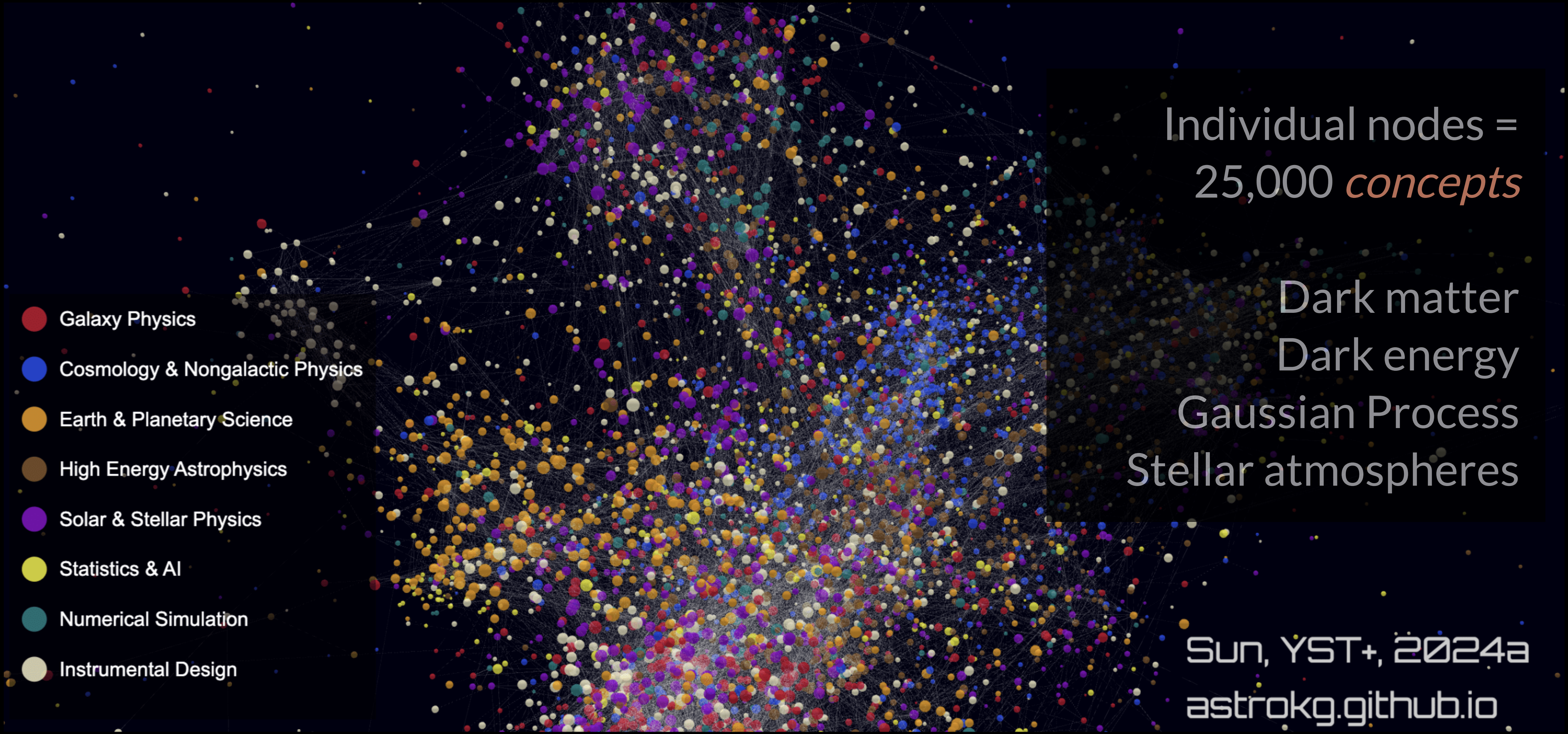
Zechang Sun[1*] , **Yuan-Sen Ting**[2,3] , **Yaobo Liang**[4] , **Nan Duan**[4] , **Song Huang**[1] , **Zheng Cai**[1]
[1]Department of Astronomy, Tsinghua University, Beijing, China
[2]Research School of Astronomy and Astrophysics, Australian National University, Canberra, Australia
[3]School of Computing, Australian National University, Canberra, Australia
[4]Microsoft Research Asia, Beijing, China
szc22@mails.tsinghua.edu.cn, yuan-sen.ting@anu.edu.au, yalia@microsoft.com,
nanduan.nlp@outlook.com, {shuang, zcai}@mail.tsing

Abstract

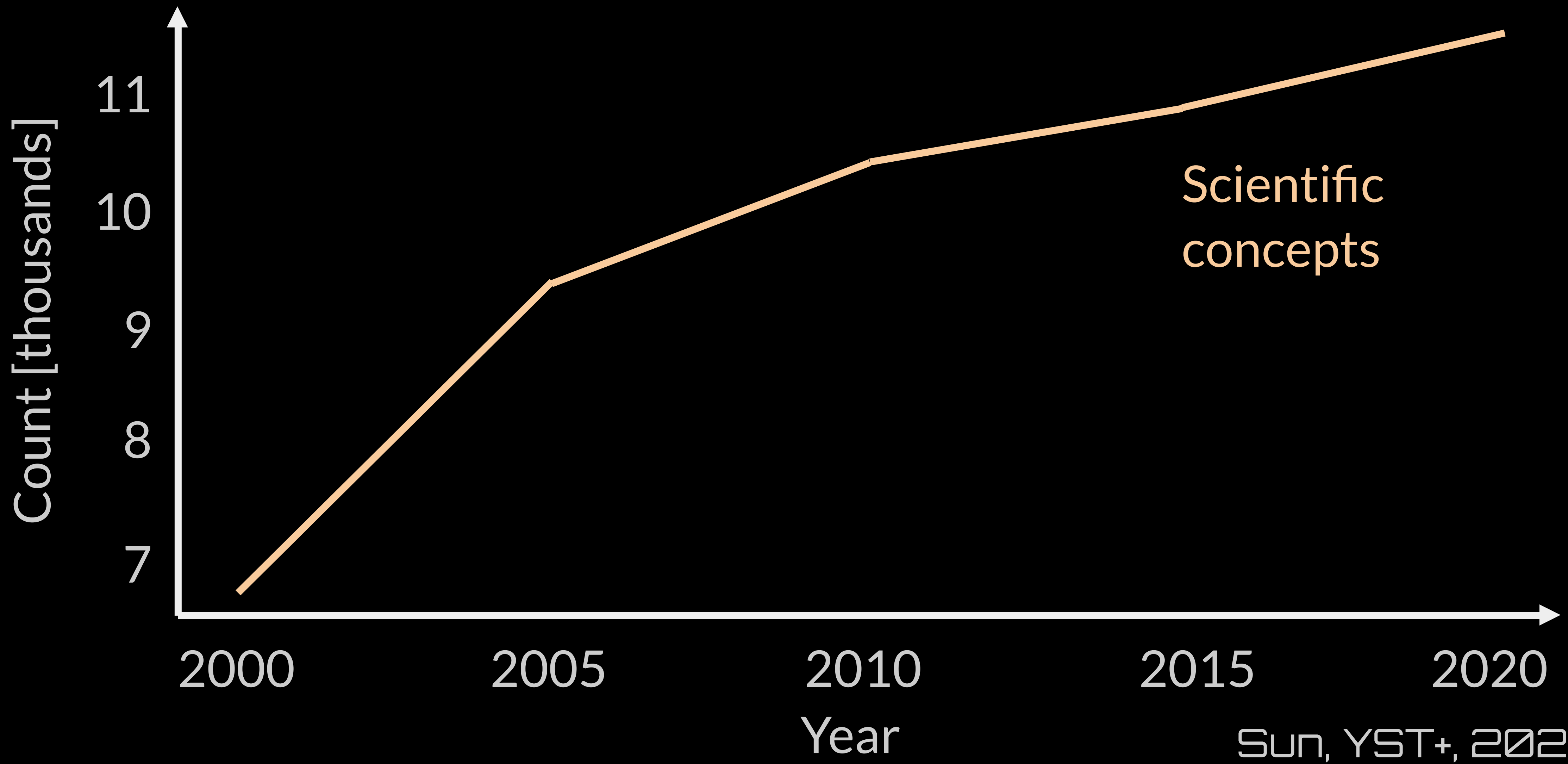
Identifying and predicting the factors that contribute to the success of interdisciplinary research is crucial for advancing scientific discovery. However, there is a lack of methods to quantify the integration of new ideas and technological advancements in astronomical research and how these new technologies drive further scientific breakthroughs.

the theoretical understa
eral relativity has drive
berg, 2008], and each
leads to new windows
the detection of gravita
al., 2016], which was
cutting-edge technologi
high-performance com
standing complex syste

Visualizing the *knowledge graph* and evolution in astronomy

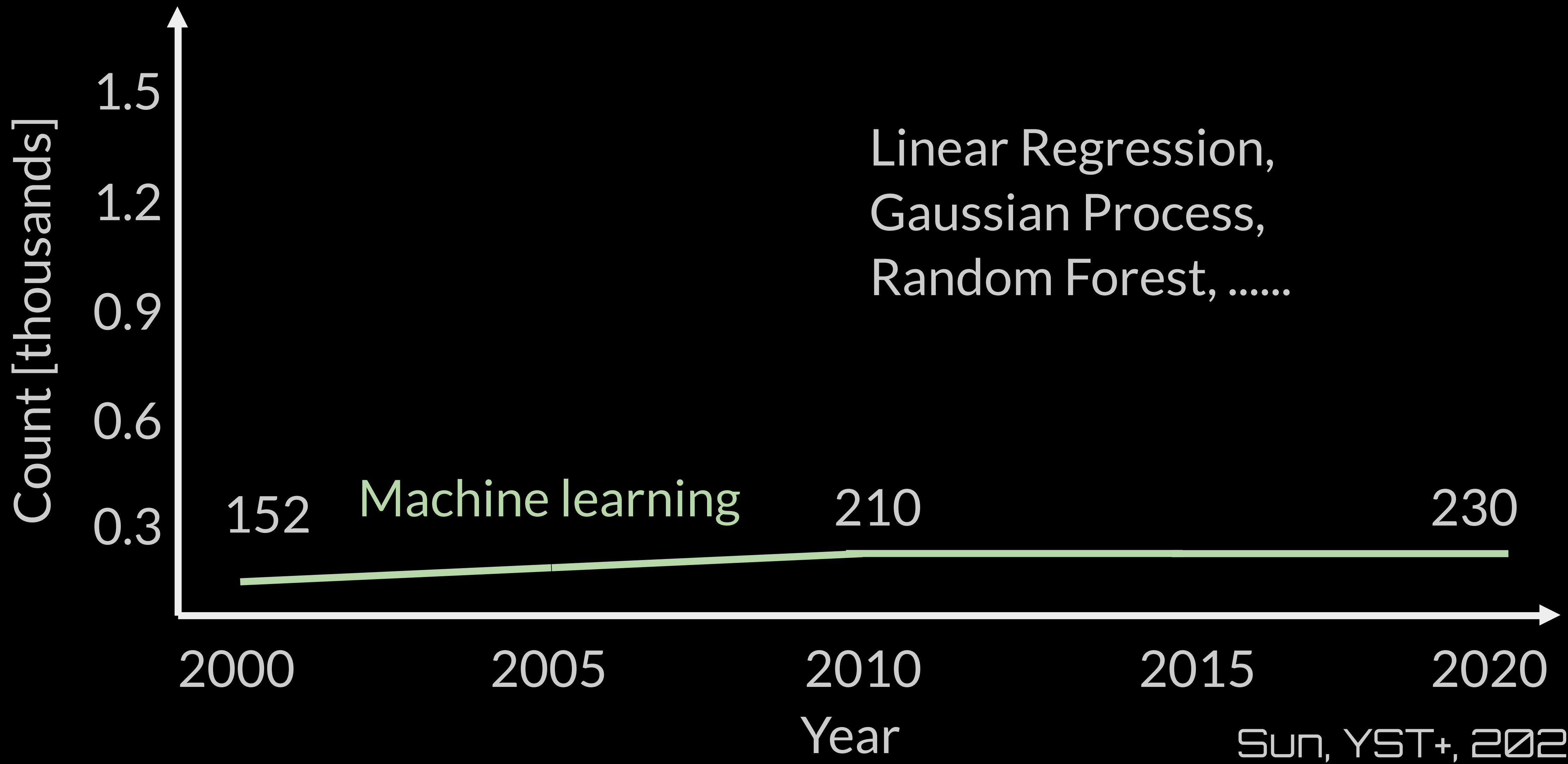


Quantifying the growth of the field -- by groups of concepts



Sun, YST+, 2024a

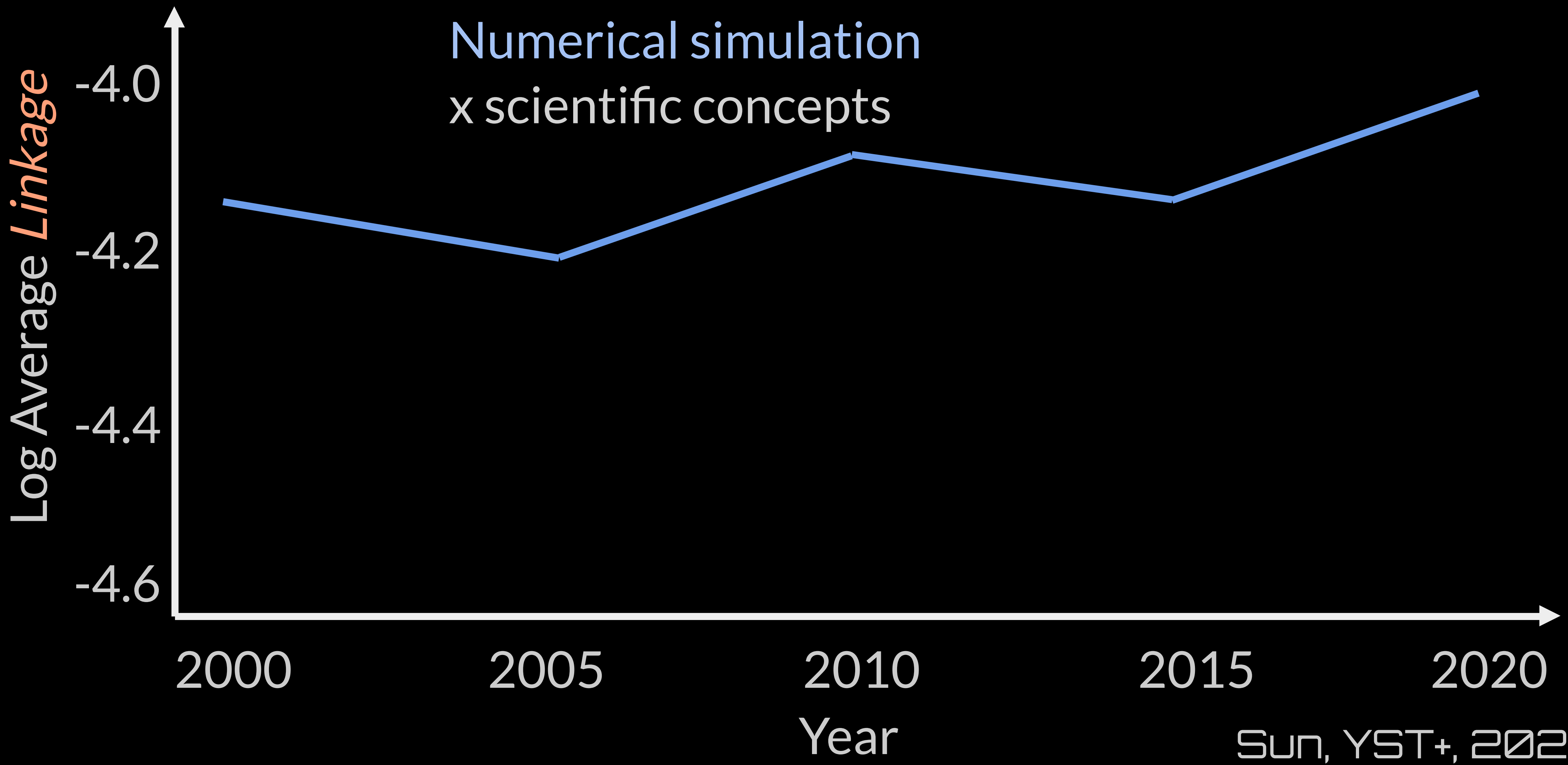
The number of ML concepts in astronomy has *not grown*



Sun, YST+, 2024a

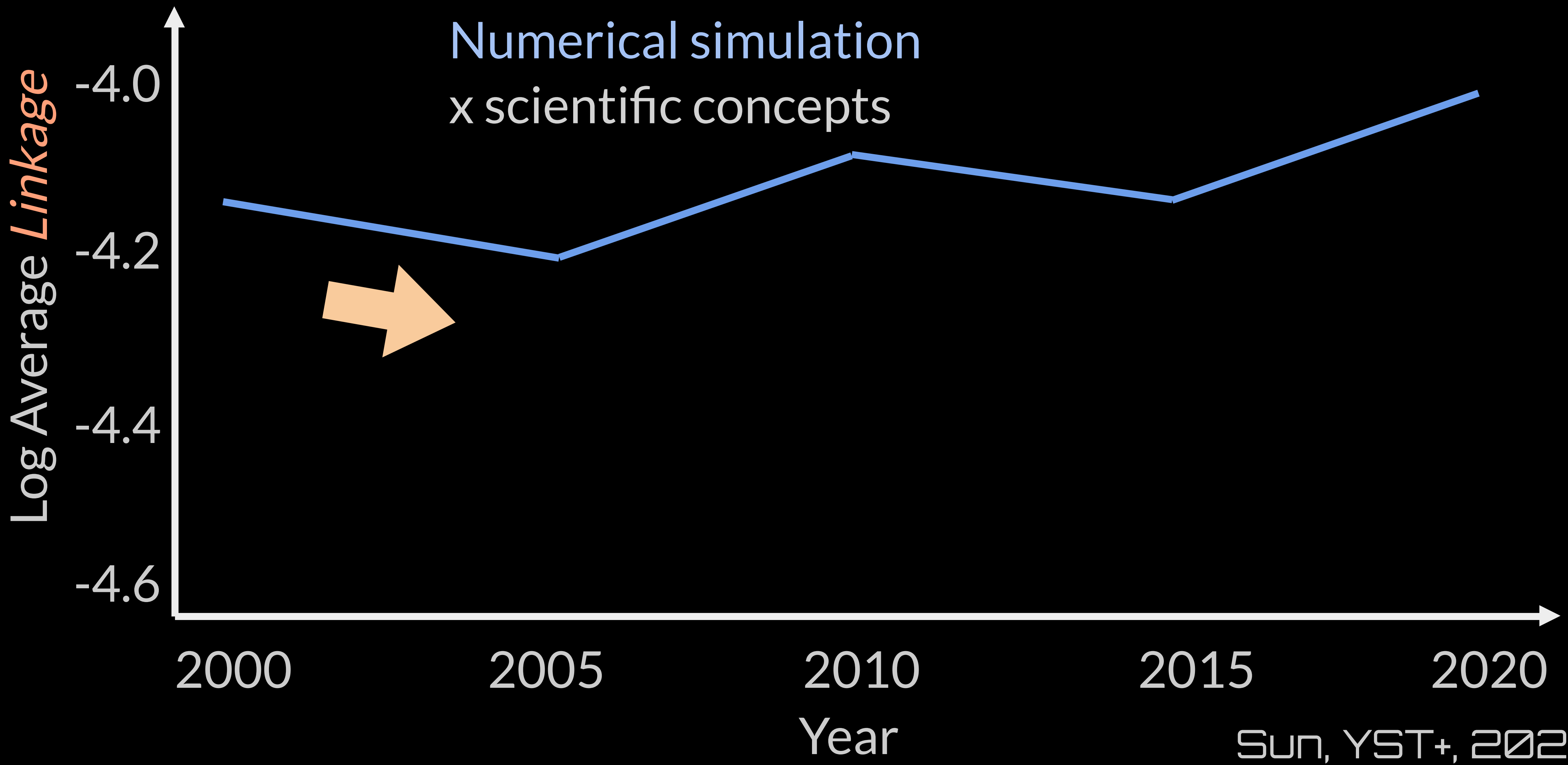
Quantifying the *cross-domain* interaction:
How technical concepts inspire scientific ones

Cross-domain linkage shows a *two-phase* evolution



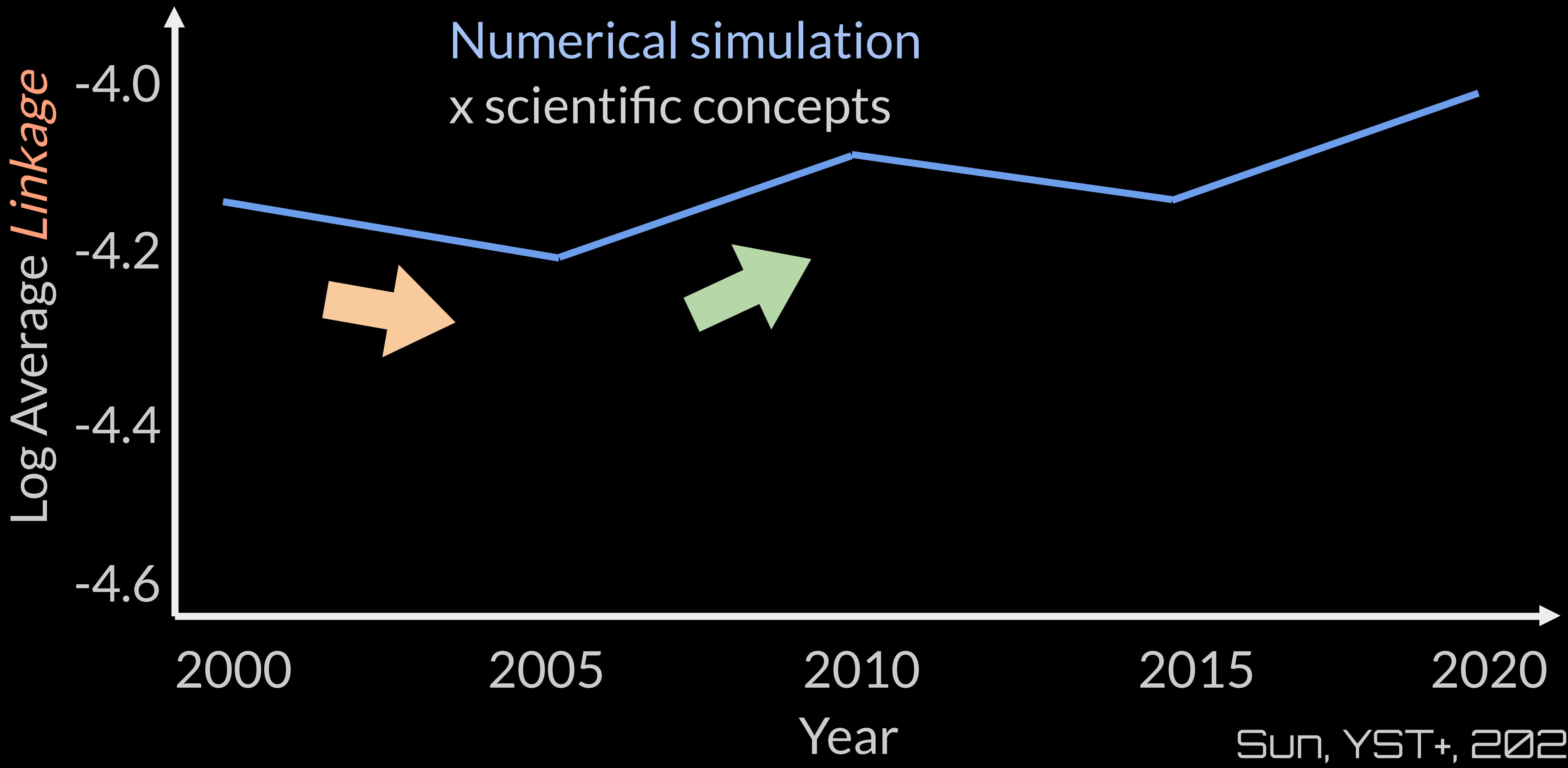
Sun, YST+, 2024a

Cross-domain linkage shows a *two-phase* evolution



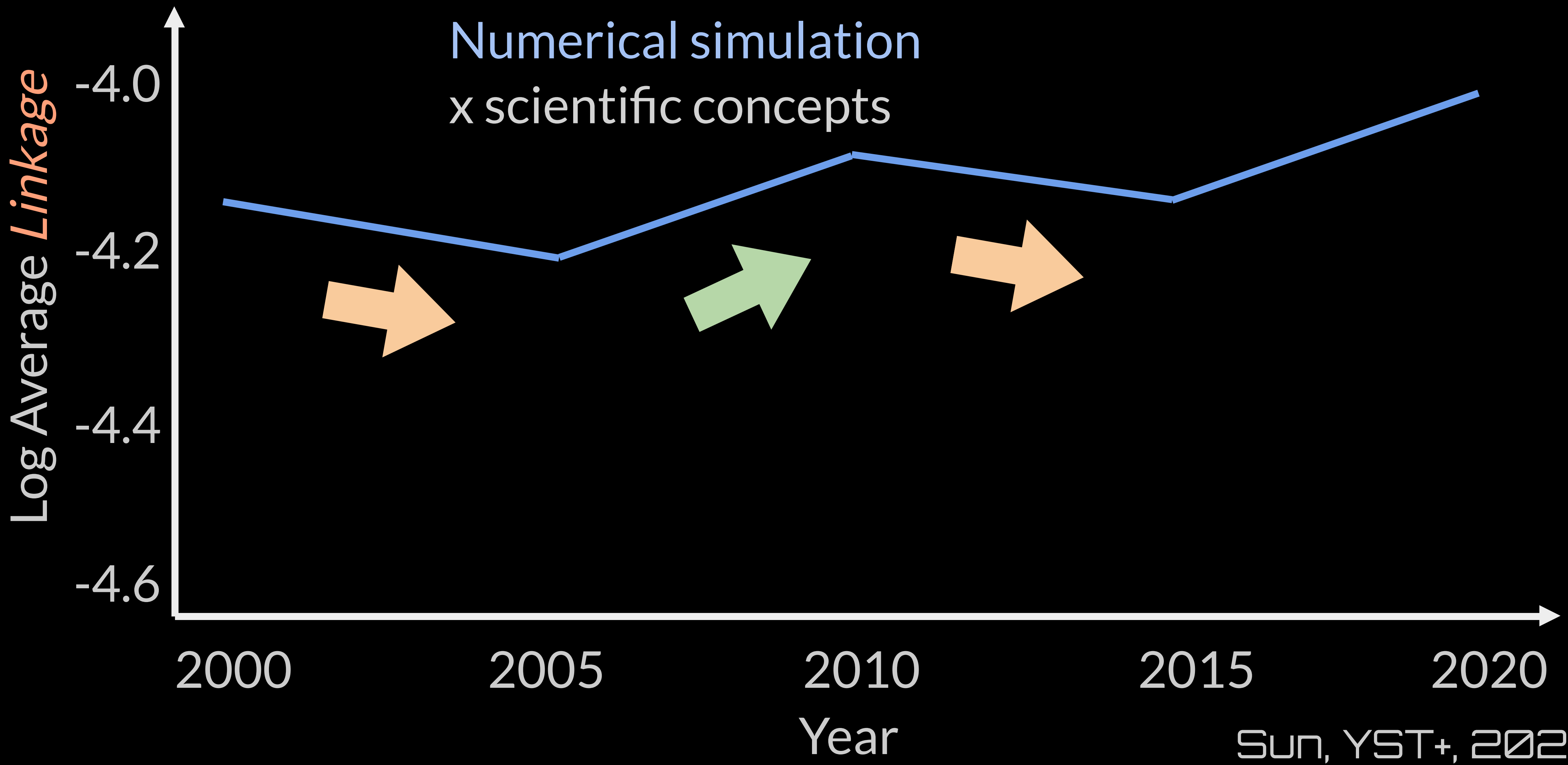
Sun, YST+, 2024a

Cross-domain linkage shows a *two-phase* evolution



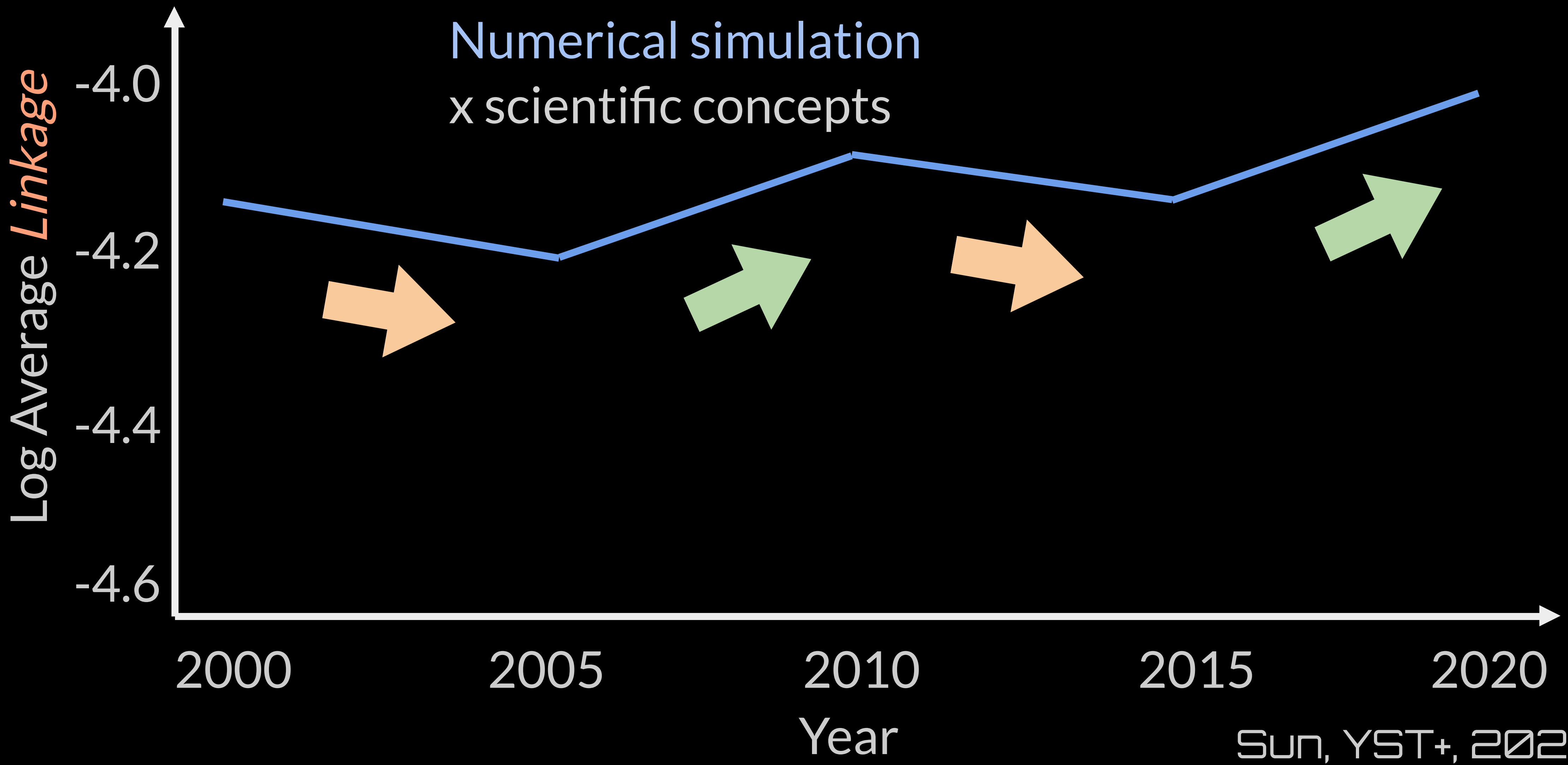
Sun, YST+, 2024a

Cross-domain linkage shows a *two-phase* evolution



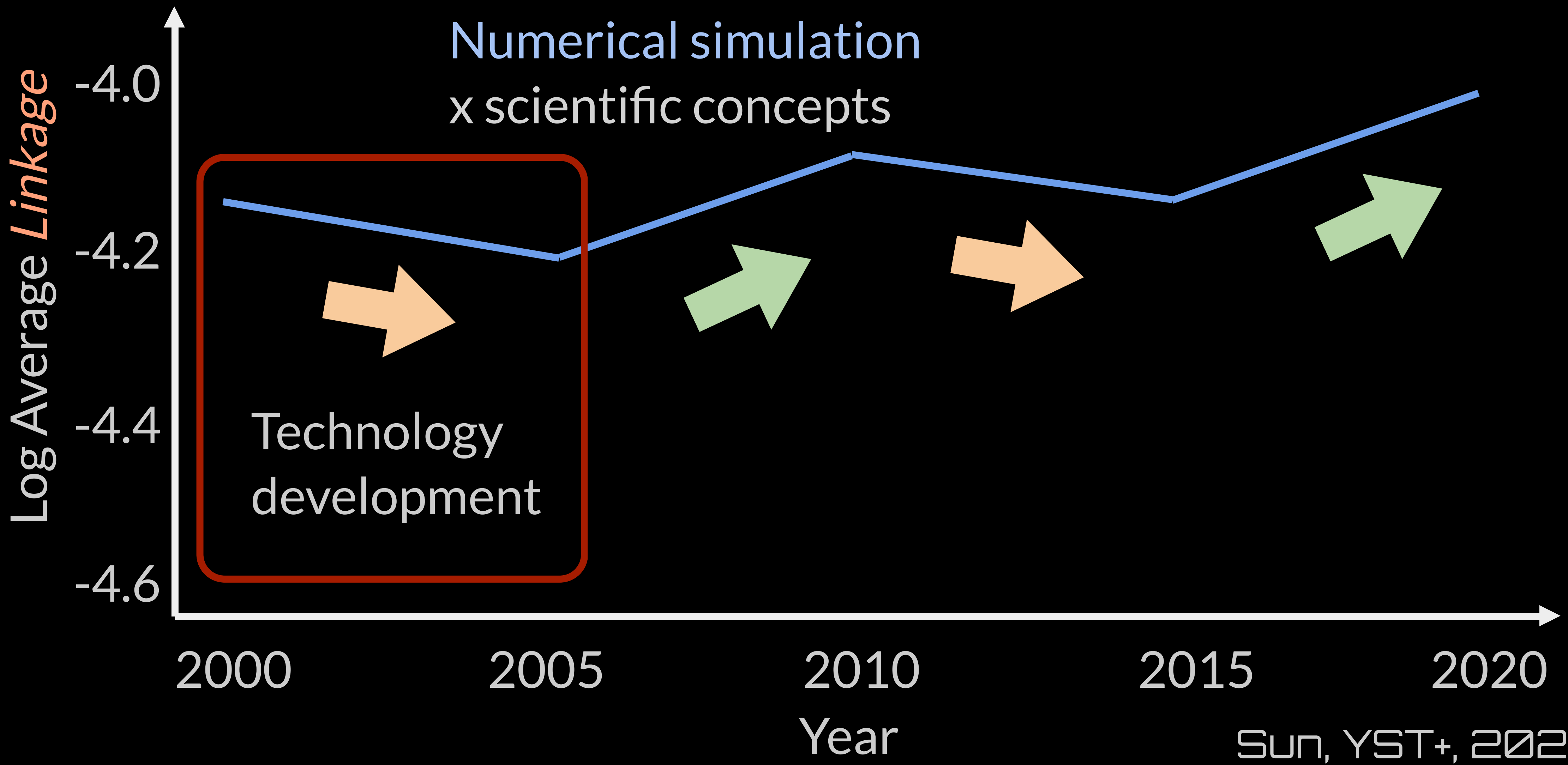
Sun, YST+, 2024a

Cross-domain linkage shows a *two-phase* evolution

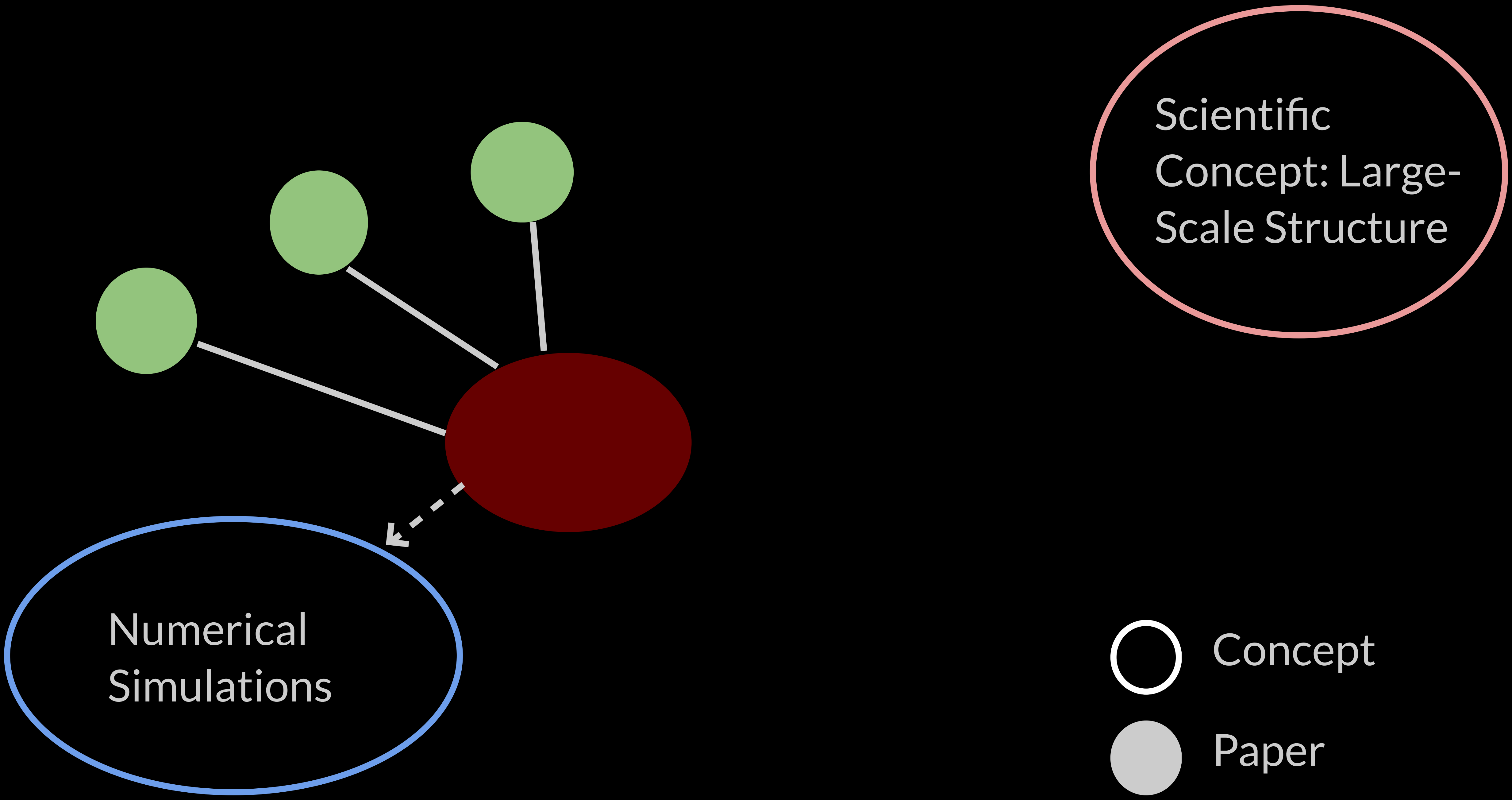


Sun, YST+, 2024a

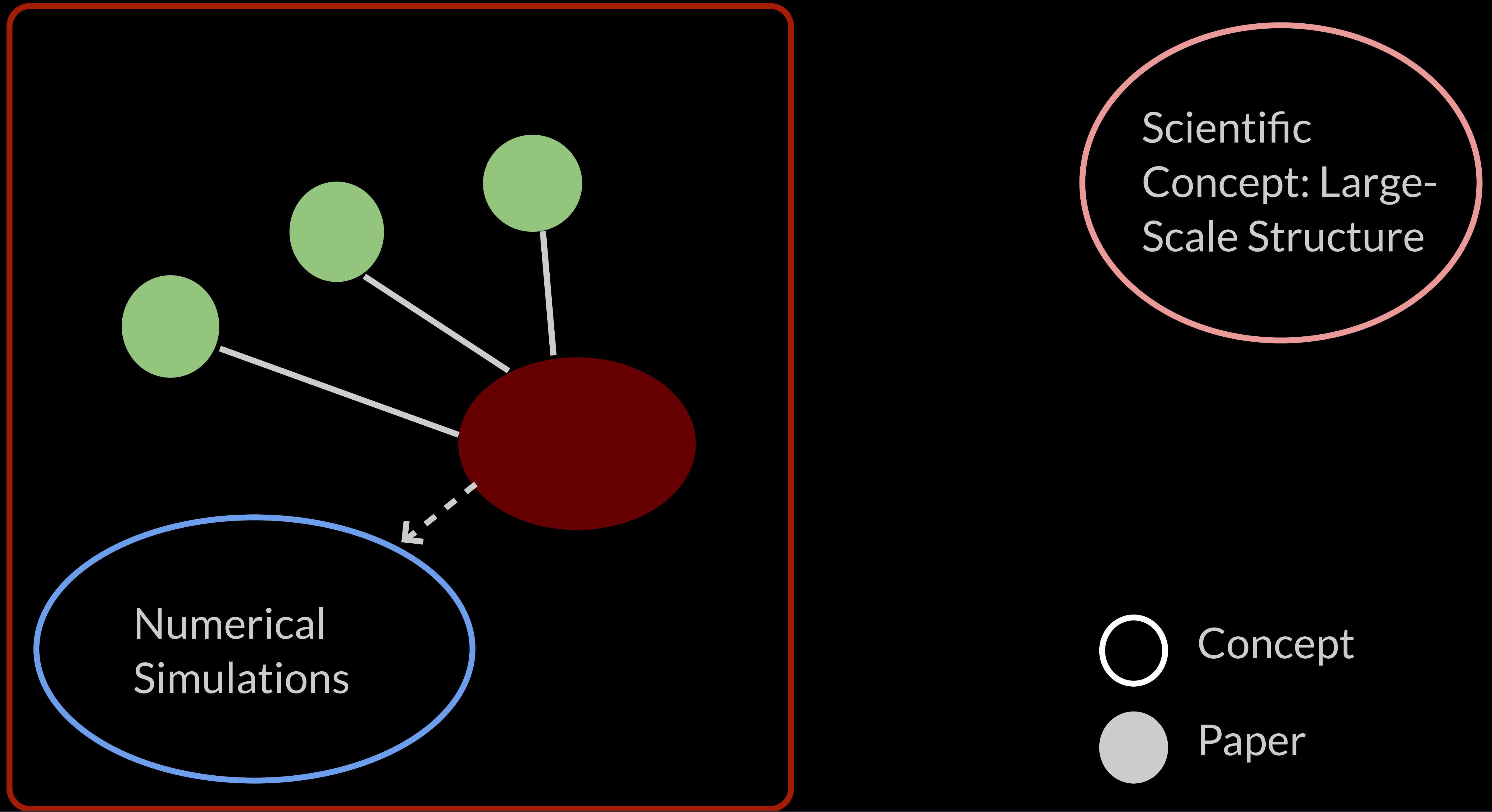
Cross-domain linkage shows a *two-phase* evolution

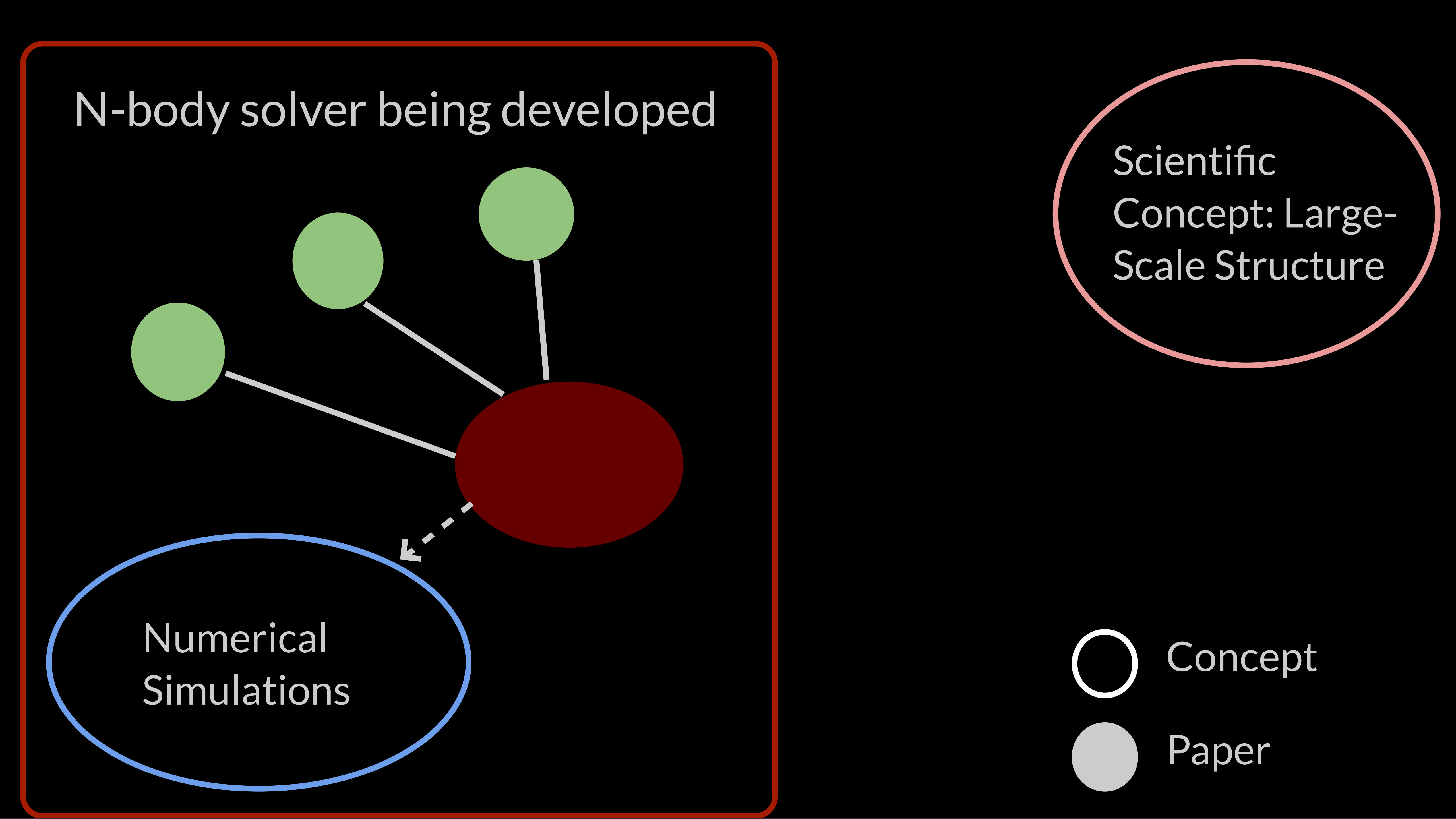


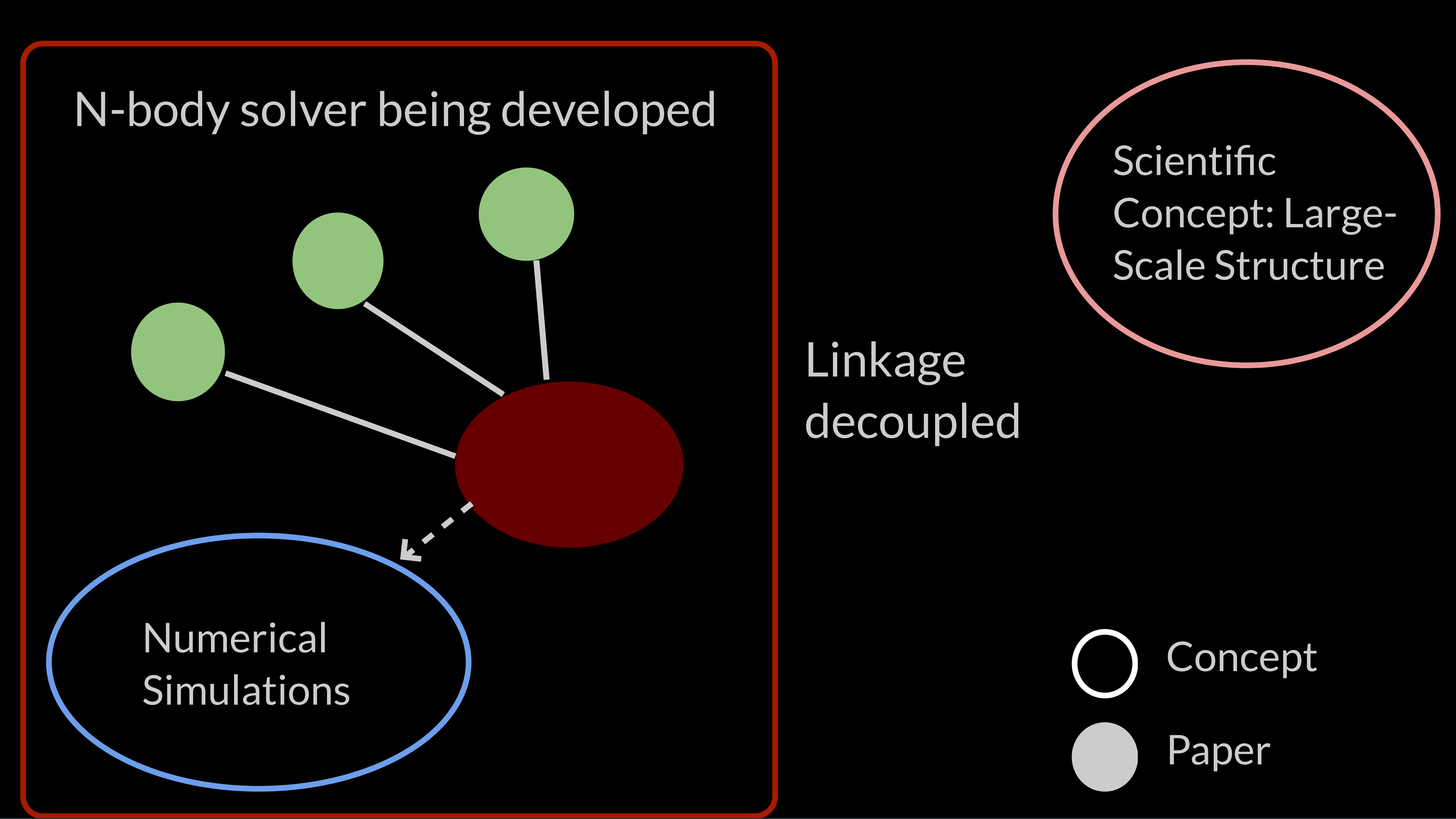
Sun, YST+, 2024a



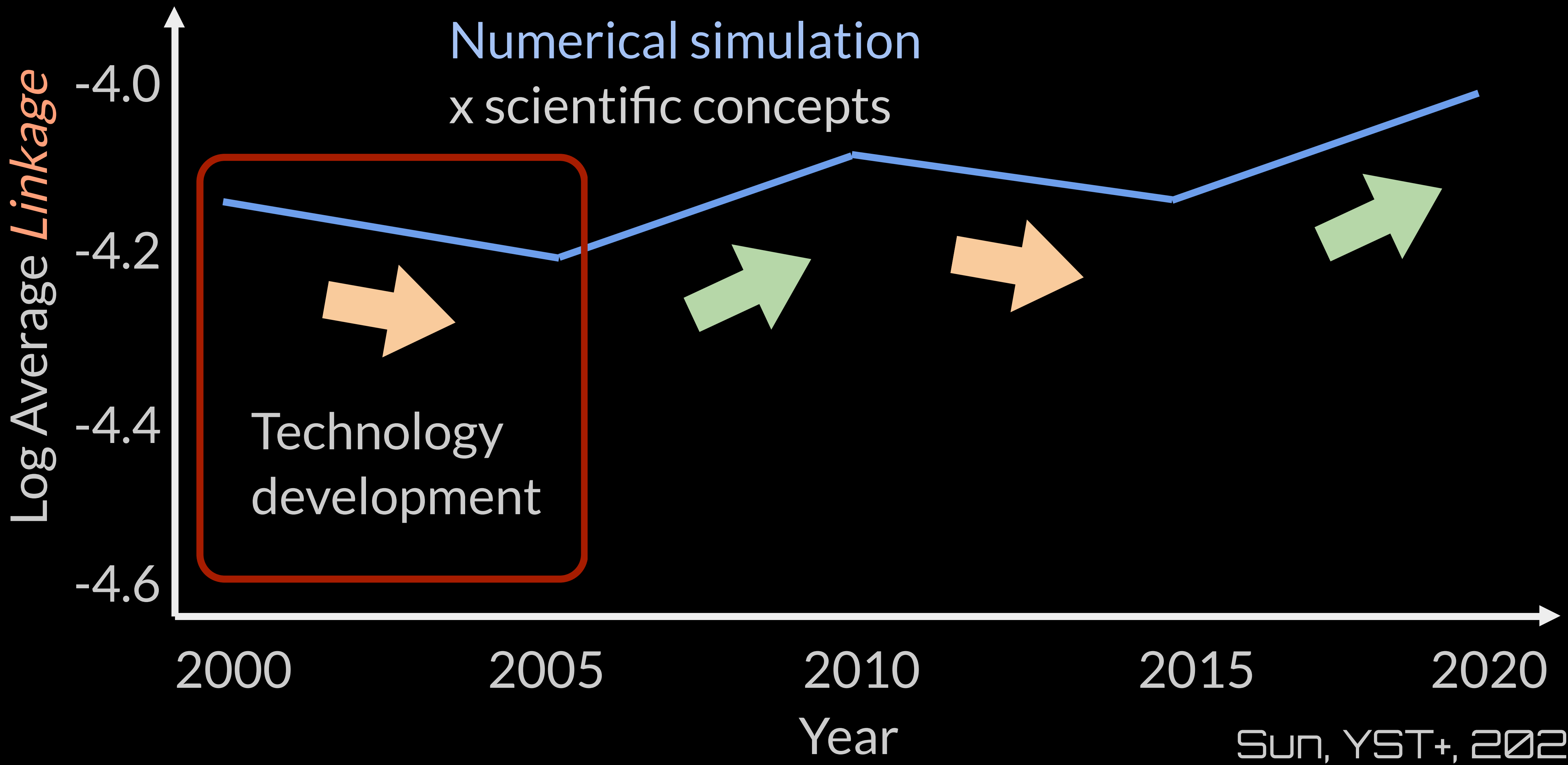
Scientific Concept: Large-Scale Structure
Numerical Simulations
Concept
Paper


Scientific Concept: Large-Scale Structure
Numerical Simulations
Concept
Paper


N-body solver being developed
Numerical Simulations
Scientific Concept: Large-Scale Structure
Concept
Paper

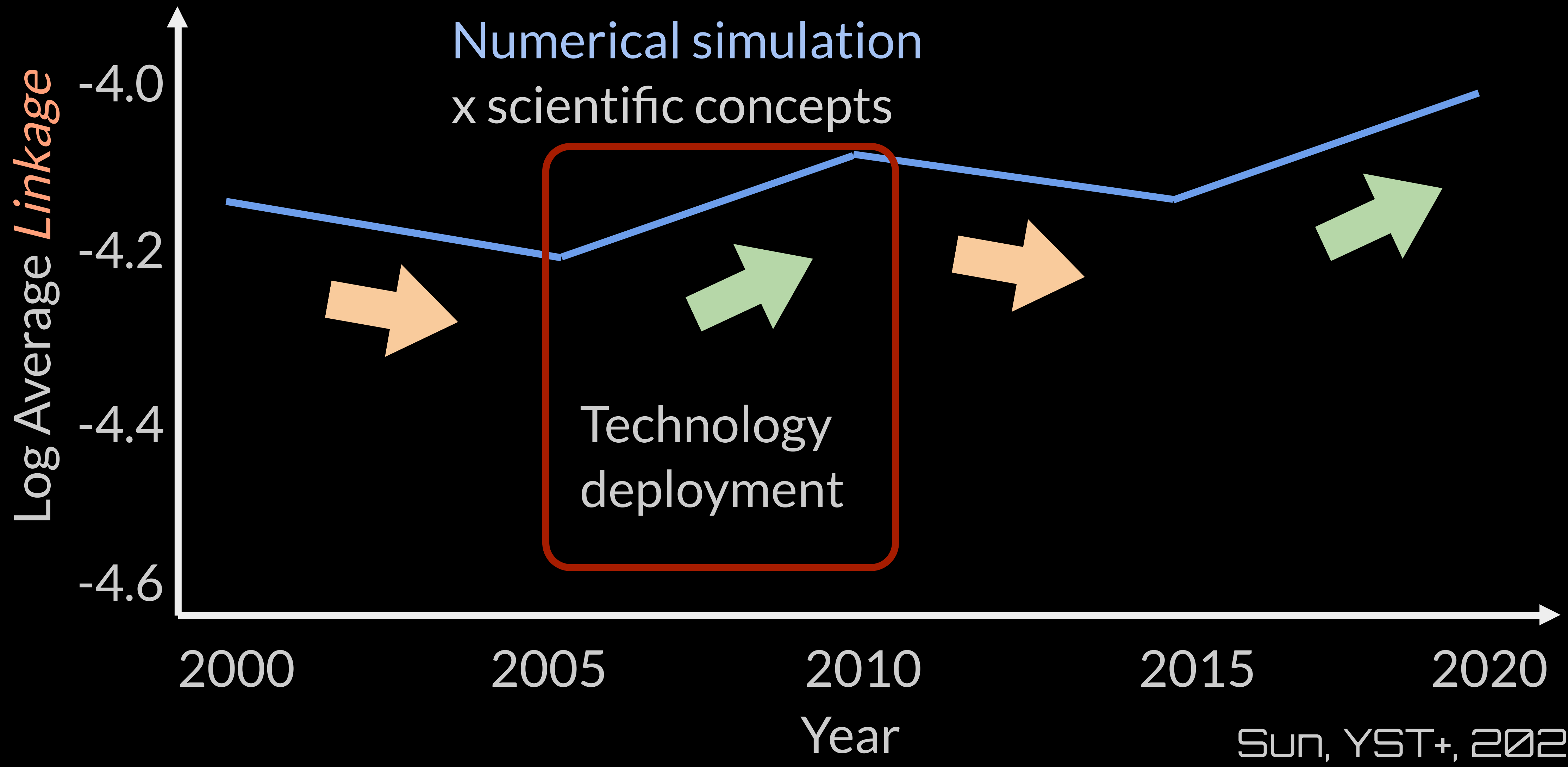

N-body solver being developed
Linkage decoupled
Scientific Concept: Large-Scale Structure
Numerical Simulations
Concept
Paper

Cross-domain linkage shows a *two-phase* evolution

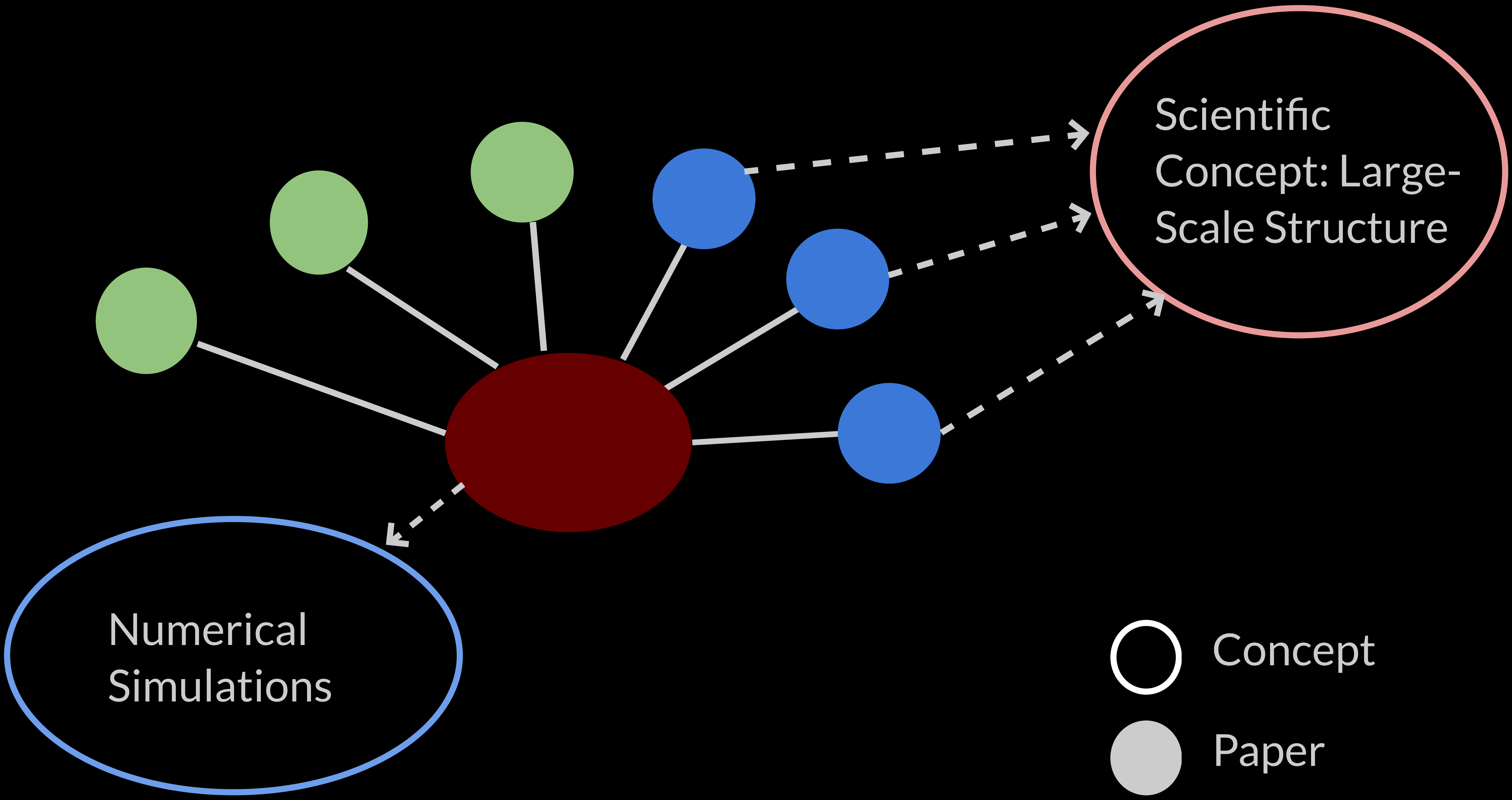


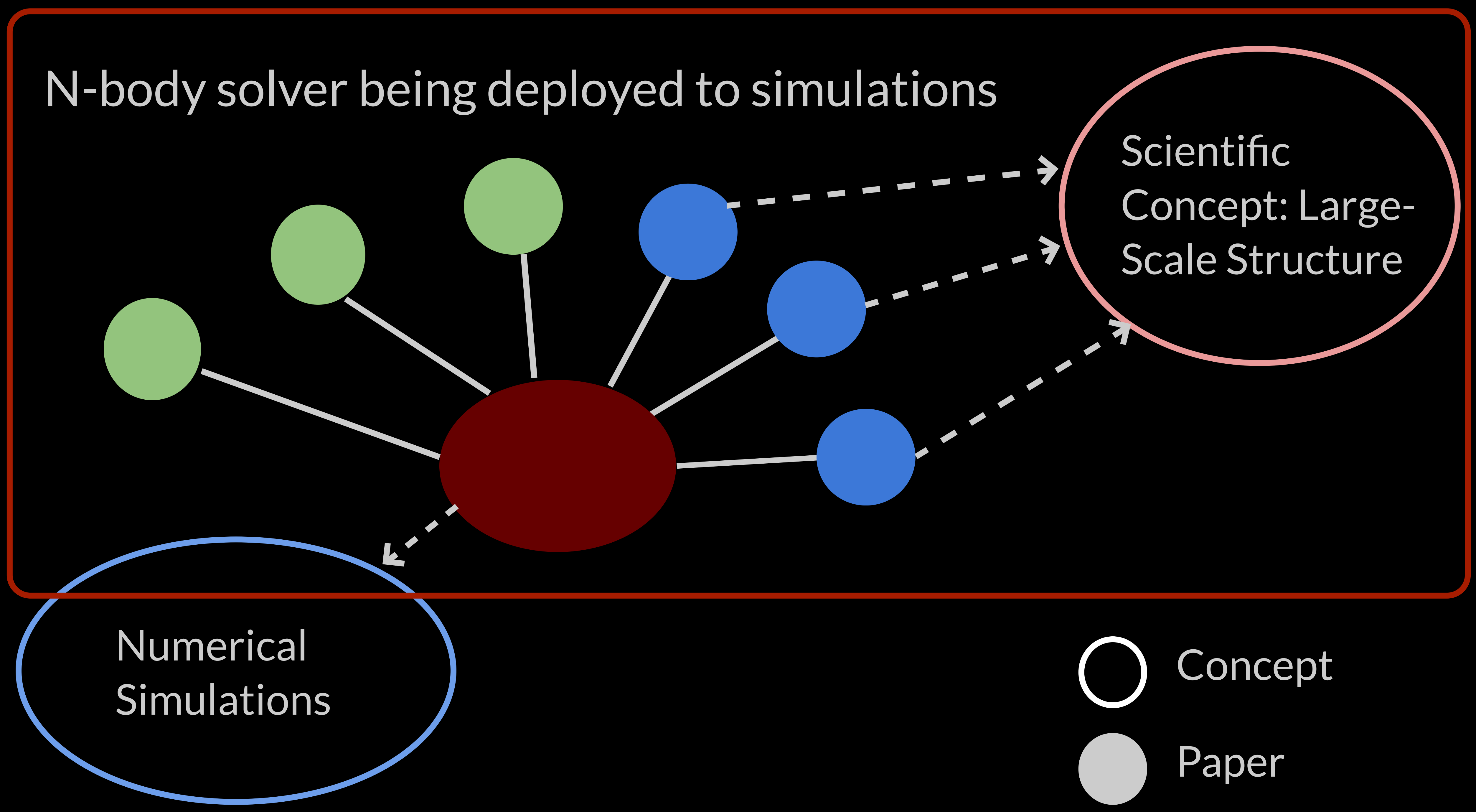
Sun, YST+, 2024a

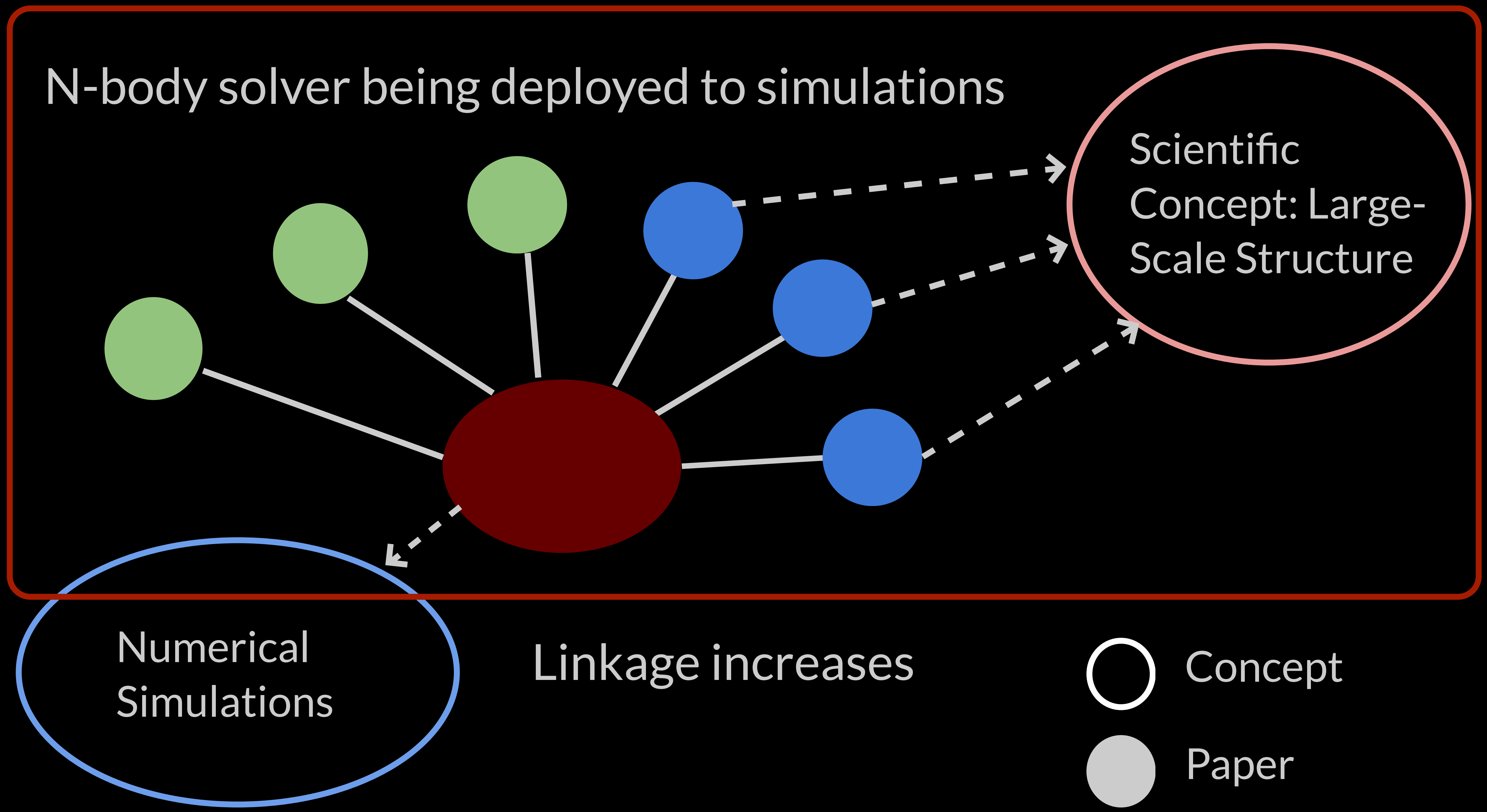
Cross-domain linkage shows a *two-phase* evolution



Sun, YST+, 2024a

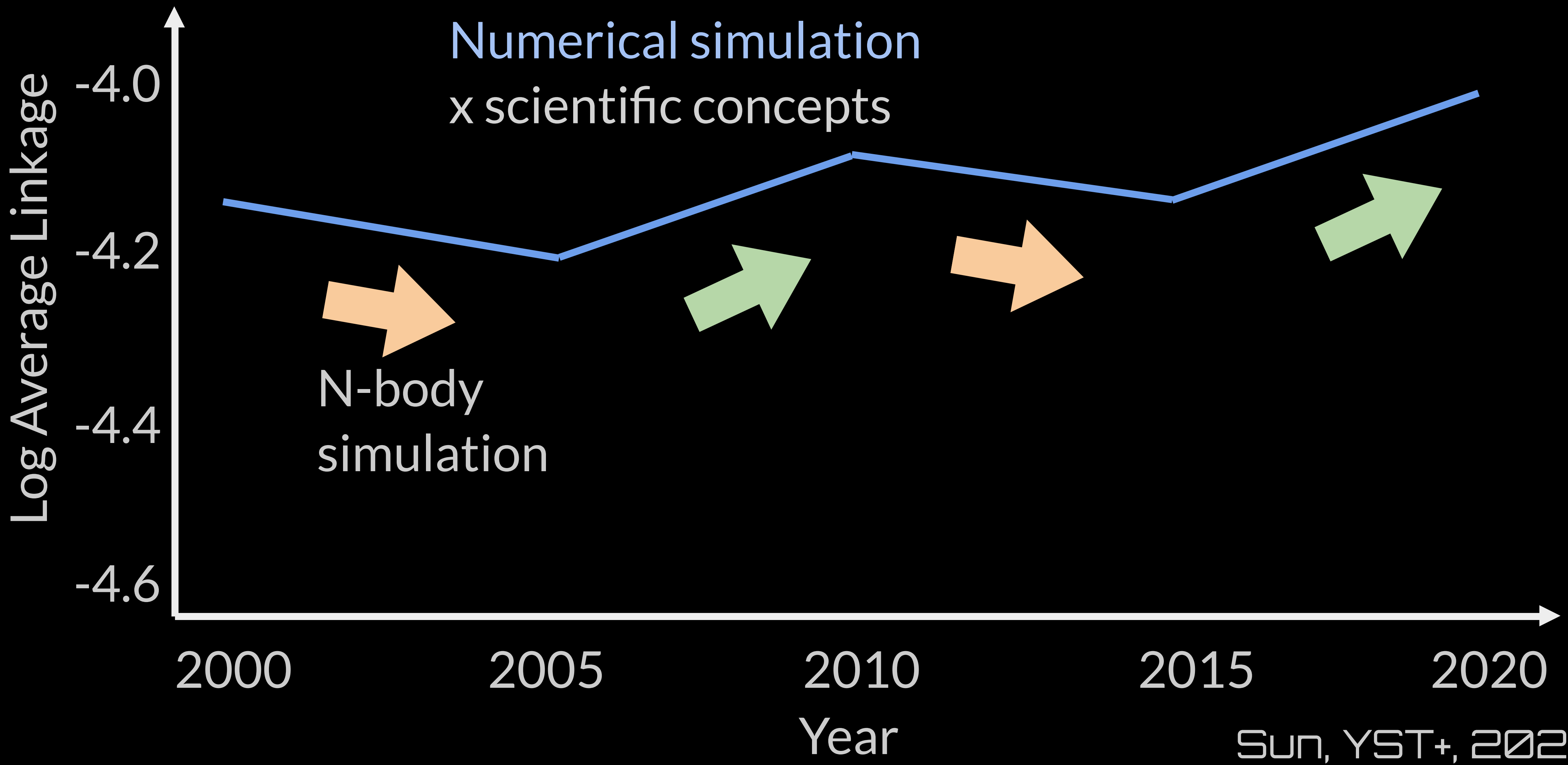

Scientific Concept: Large-Scale Structure
Numerical Simulations
Concept
Paper


N-body solver being deployed to simulations
Scientific Concept: Large-Scale Structure
Numerical Simulations
Concept
Paper



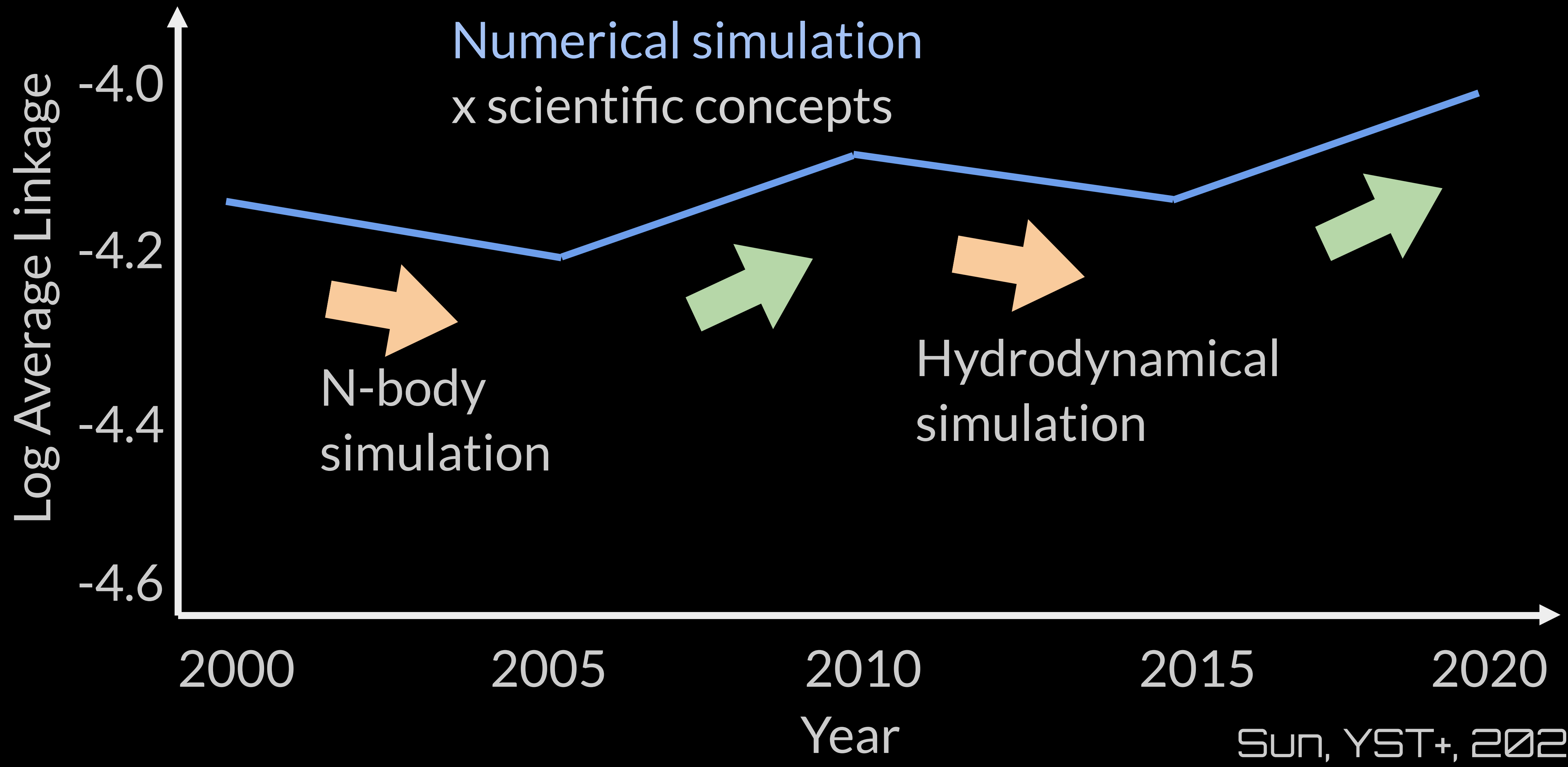
N-body solver being deployed to simulations
Scientific Concept: Large-Scale Structure
Numerical Simulations
Linkage increases
Concept
Paper

Cross-domain linkage shows a *two-phase* evolution



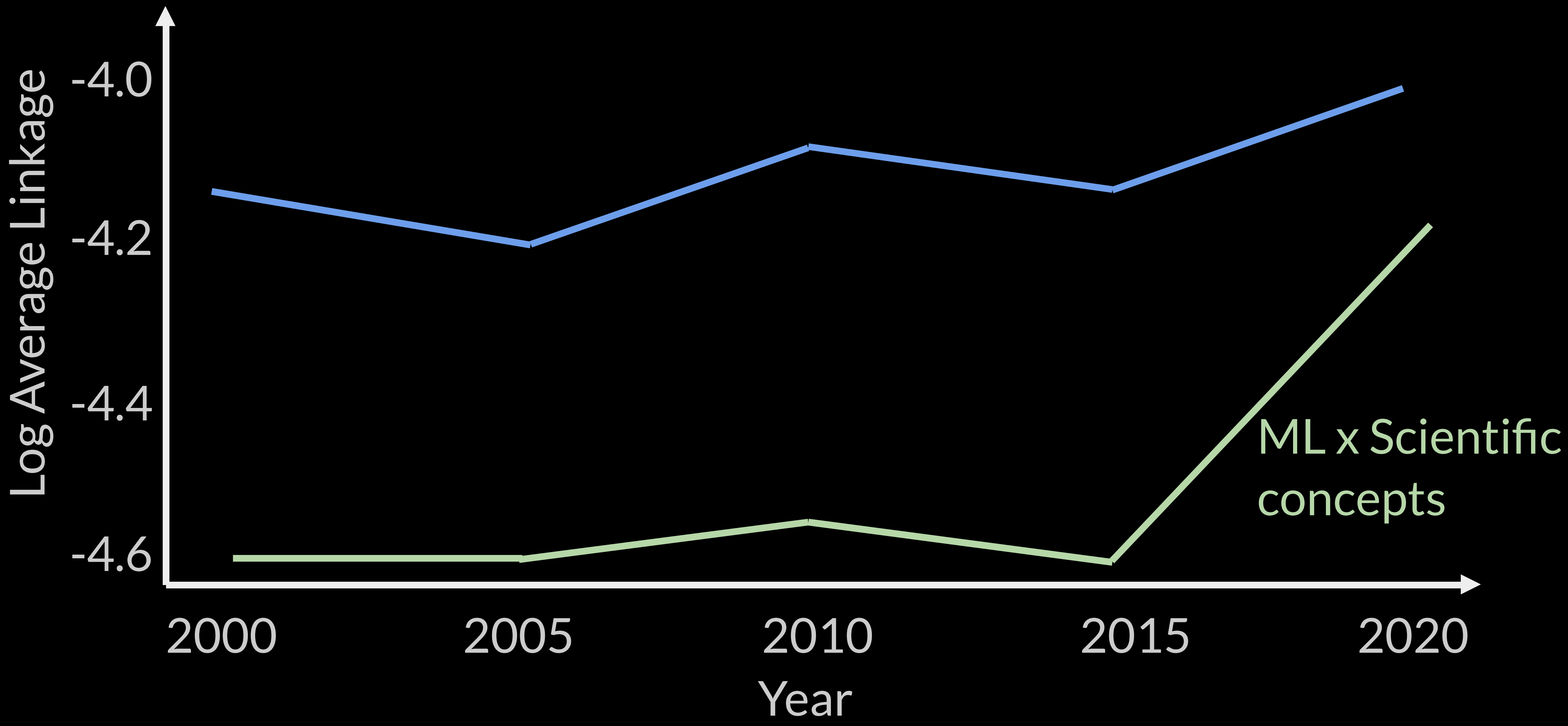
Sun, YST+, 2024a

Cross-domain linkage shows a *two-phase* evolution



Sun, YST+, 2024a

Interest in AI x Astronomy *outpaces* technological development



Machine learning is becoming *increasingly integrated* into astronomy,

However, the field has been *slow to incorporate new concepts* and techniques.

Astronomy is *not* biology

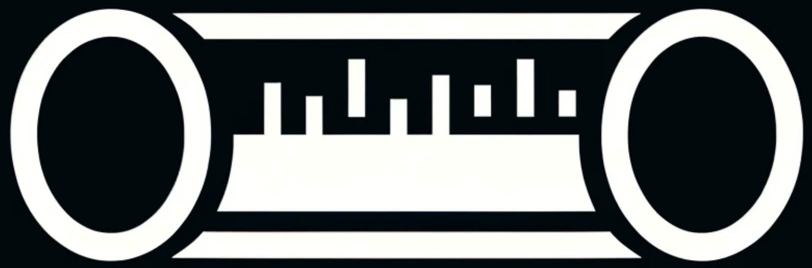
Data / Observation

Theory / Hypothesis

Astronomy is *not* biology

Data /
Observation

Theory /
Hypothesis

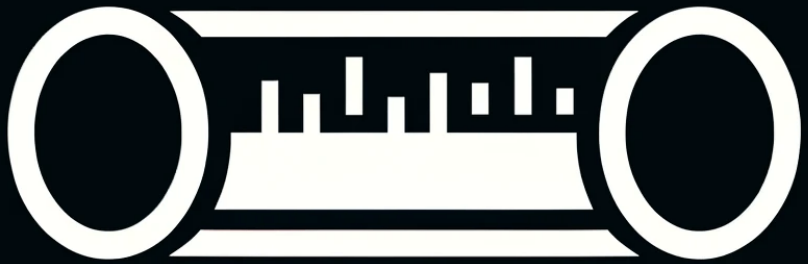


Analysis
Pipelines

Astronomy is *not* biology

Data / Observation

Theory / Hypothesis



Analysis Pipelines

True

False

Biology faced a fundamental *bottleneck*

Data / Observation

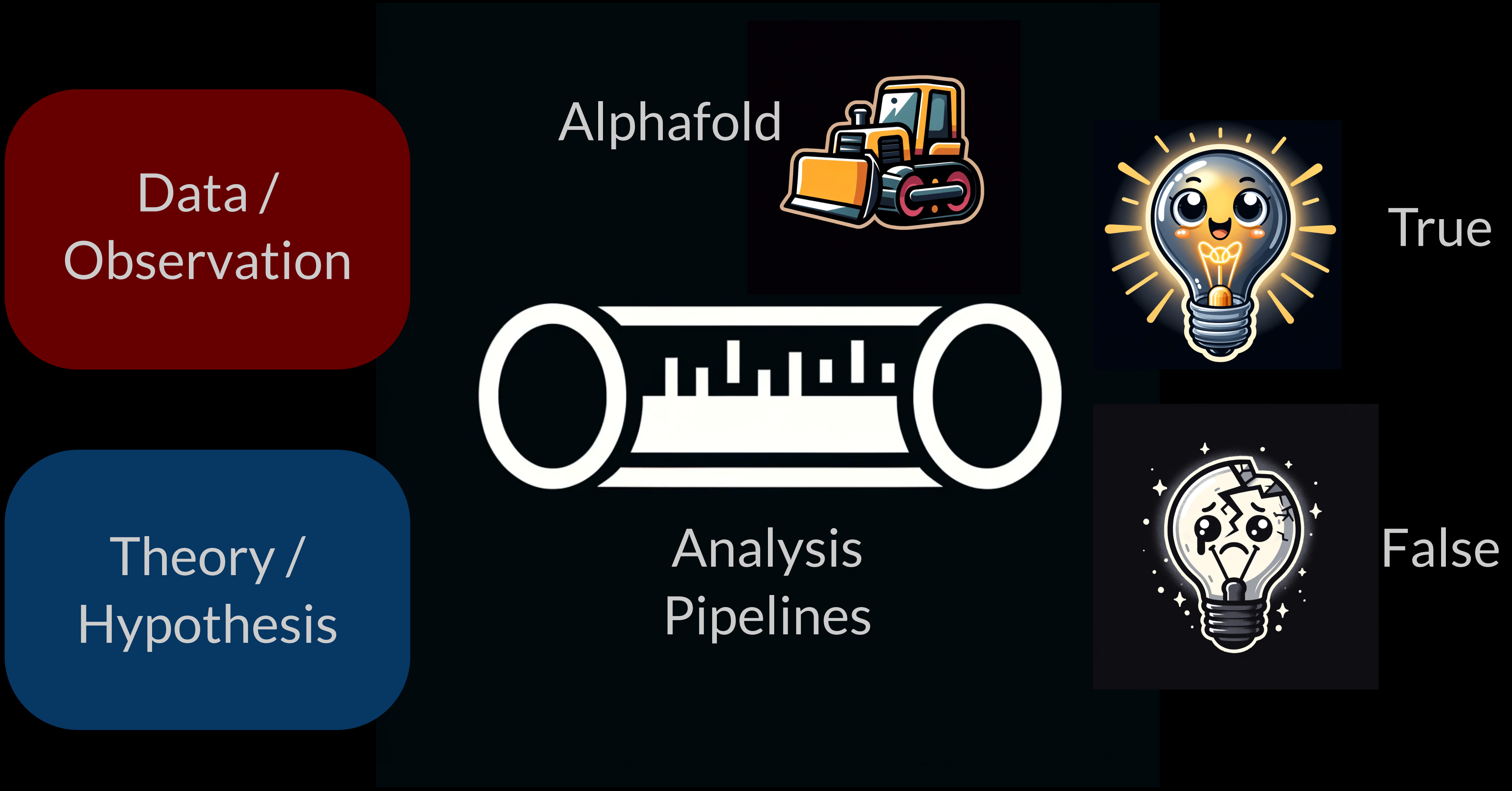
Theory / Hypothesis

Analysis Pipelines

True

False

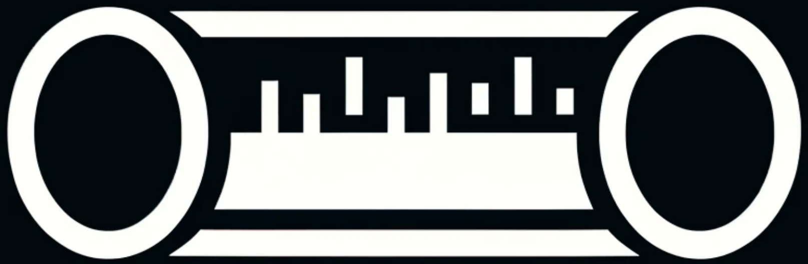
Biology faced a fundamental *bottleneck*



Astronomy already has a successful *standard model*

Data / Observation

Theory / Hypothesis



LamdaCDM

True

False

- Where are we in AI for Astronomy - What's the problem?
- What could be the AlphaFold moment for astronomy?
- How do we get there in a cost-effective way?

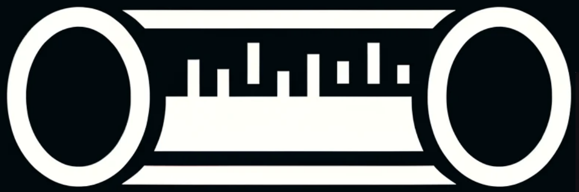
- Where are we in AI for Astronomy - What's the problem?
- What could be the AlphaFold moment for astronomy?
- How do we get there in a cost-effective way?

Toward an *AI Astronomer*

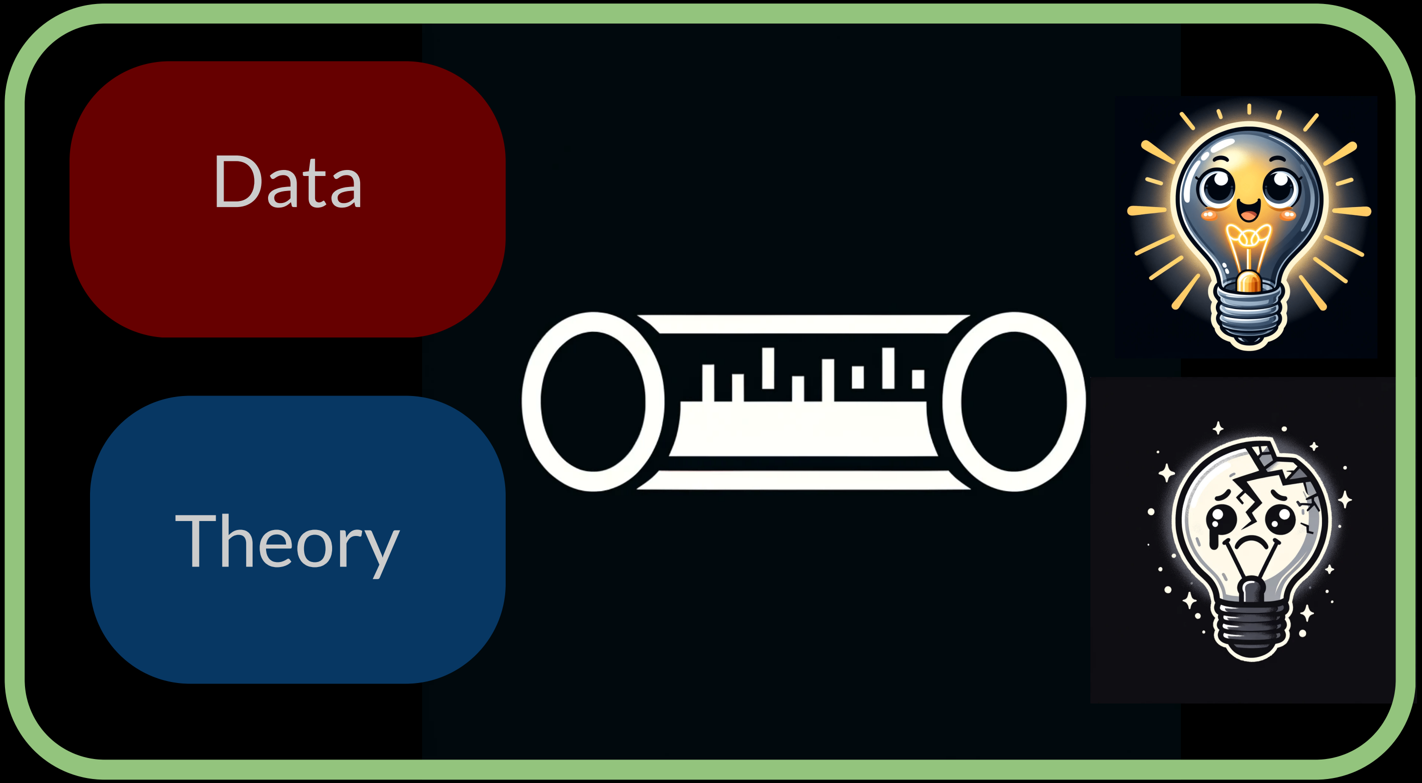
Research is essentially a *reinforcement learning* process

Data

Theory



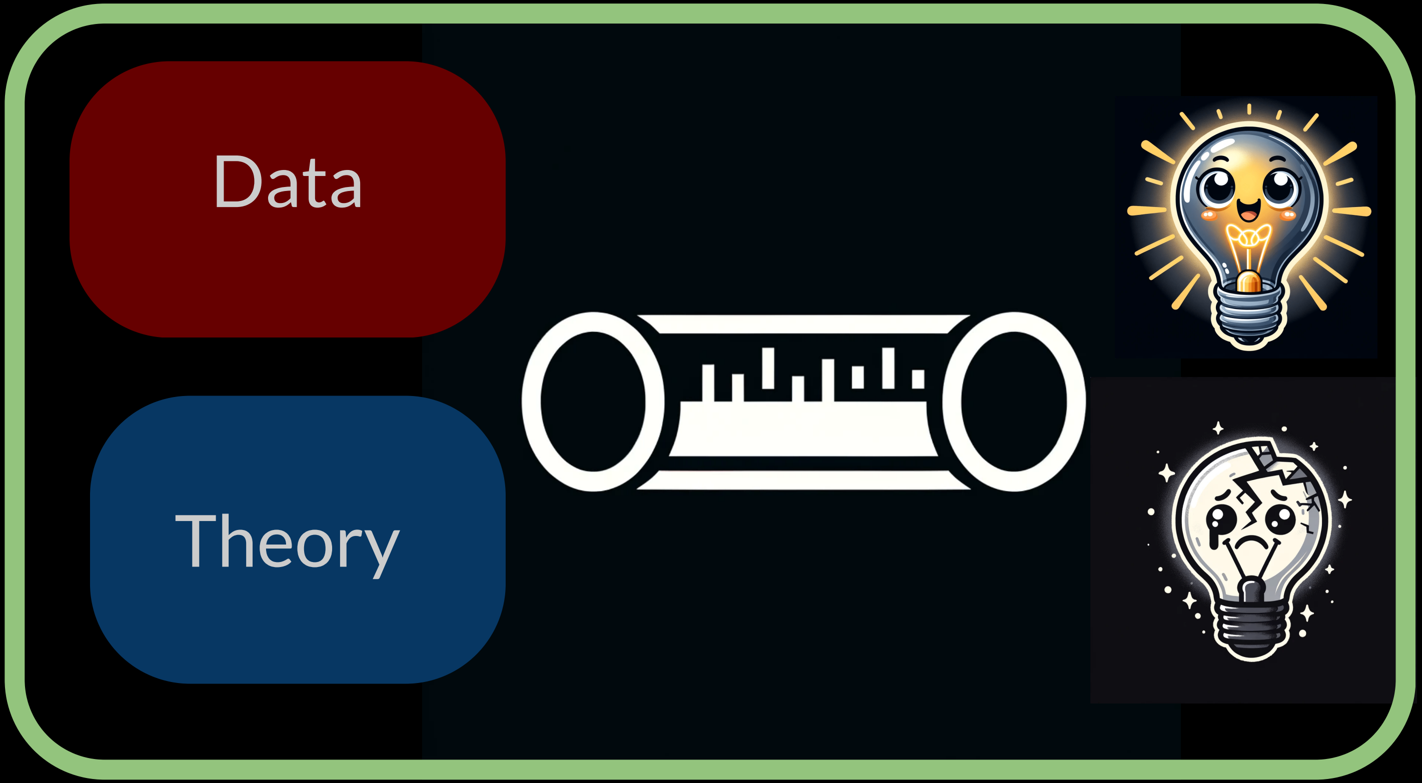
Research is essentially a *reinforcement learning* process



State of the research

Research is essentially a *reinforcement learning* process

Evaluate the state

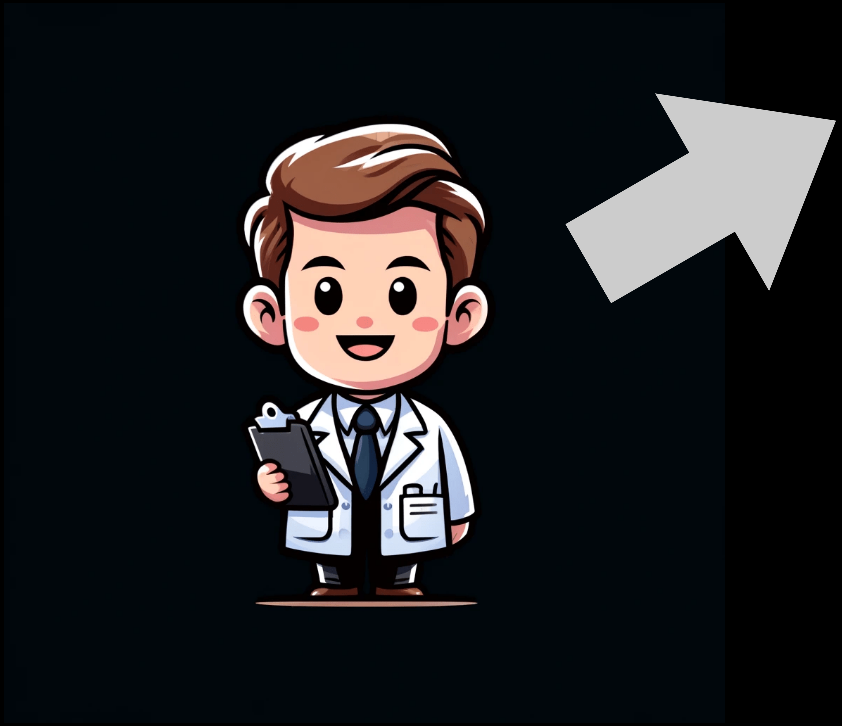


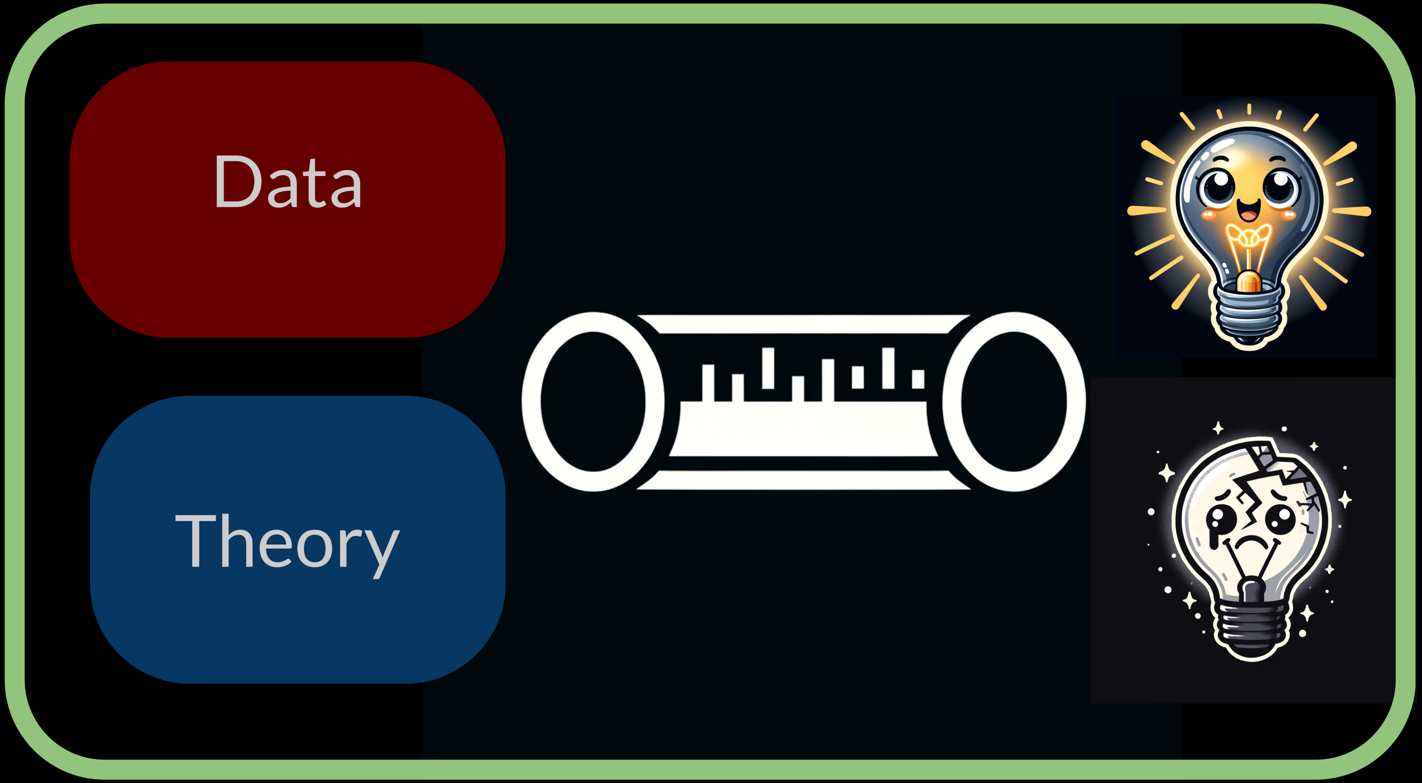
State of the research

Research is essentially a *reinforcement learning* process

Evaluate the state

propose new theory or data



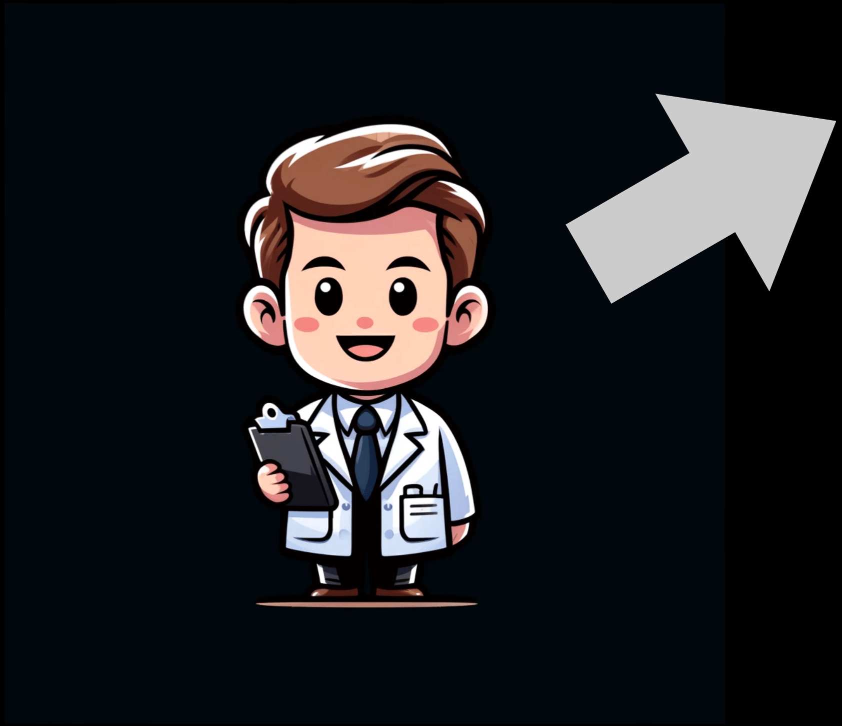


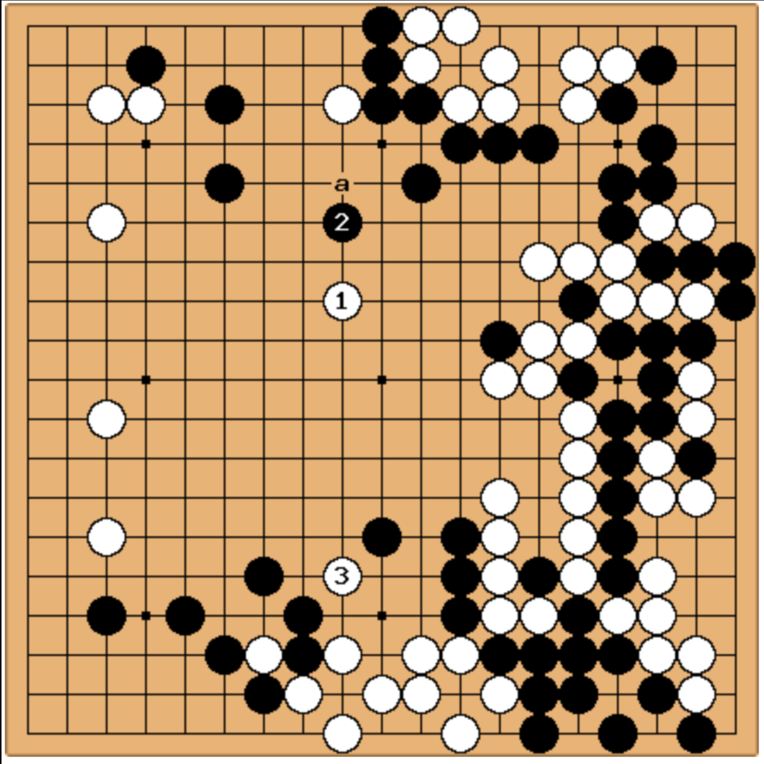
State of the research

Research is essentially a *reinforcement learning* process

Evaluate the state

propose new theory or data

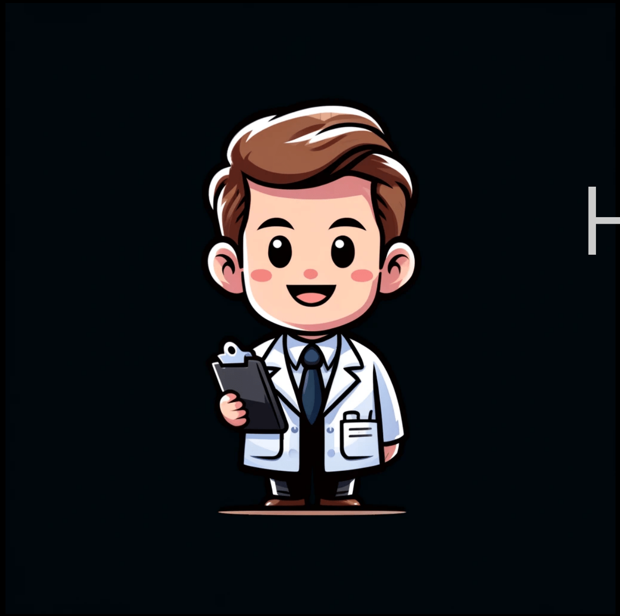




State of the research

But most *"state" of research* cannot be mathematically defined

But most *"state" of research* cannot be mathematically defined



Human *"intuition"* + experience

Perform "reinforcement learning" *(policy learning)*
without a reward function

Perform "reinforcement learning" *(policy learning)* without a reward function

Enabling LLM agents to *learn through real-world exploration* and interaction.

Automating Astronomical Research with Large Language Model Agents: A Case Study with Galaxy Spectral Energy Distribution Fitting

Zechang Sun [1] Yuan-Sen Ting [2 3] Yaobo Liang [4] Nan Duan [4] Song Huang [1] Zheng Cai [1]

Abstract

The application of AI has generated significant interest in astrophysics; however, most studies have been limited to using AI for specific predefined tasks. In contrast, real-life astronomical research involves a complex series of reasoning processes that researchers develop through years of training and experience. In this study, we explore the use of large language model-based collaboration agents to distill knowledge and experience by interacting with actual astronomical questions and

Bubeck et al., 2023). T
has led to a myriad of
perform tasks beyond
including applications
robotics (Wang et al.,
and mathematics (Az
there has been a growi
models to design and c
both in chemistry and
2023; Lei et al., 2024;

Sun, YST+, submitted

Automating Astronomical Research with Large Language Model Agents: A Case Study with Galaxy Spectral Energy Distribution Fitting

Zechang Sun [1] Yuan-Sen Ting [2 3] Yaobo Liang [4] Nan Duan [4] Song Huang [1] Zheng Cai [1]

Abstract

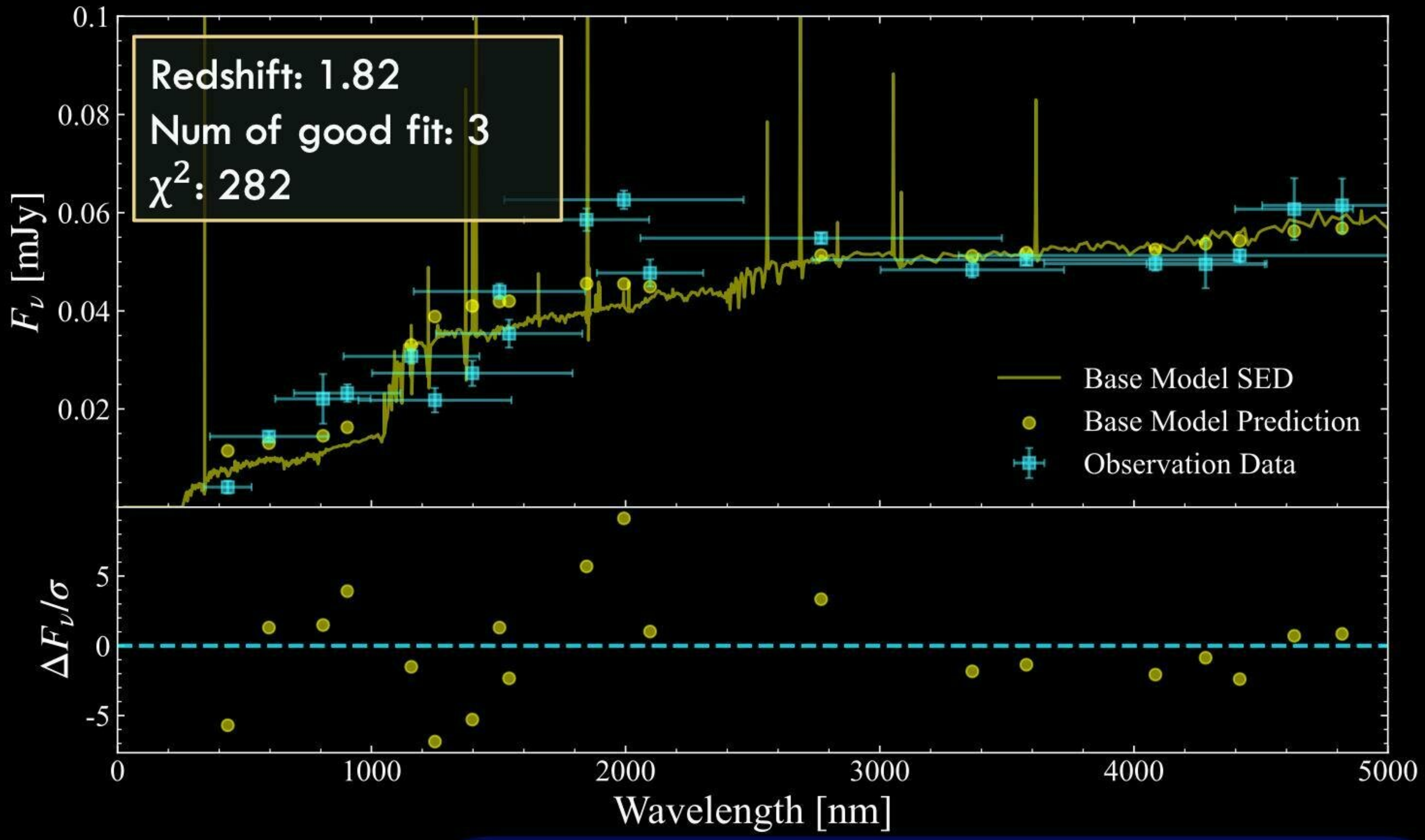
The application of AI has generated significant interest in astrophysics; however, most studies have been limited to using AI for specific predefined tasks. In contrast, real-life astronomical research involves a complex series of reasoning processes that researchers develop through years of training and experience. In this study, we explore the use of large language model-based collaboration agents to distill knowledge and experience by interacting with actual astronomical questions and

Bubeck et al., 2023). T
has led to a myriad of
perform tasks beyond
including applications
robotics (Wang et al.,
and mathematics (Az
there has been a growi
models to design and c
both in chemistry and
2023; Lei et al., 2024;

Sun, YST+, submitted

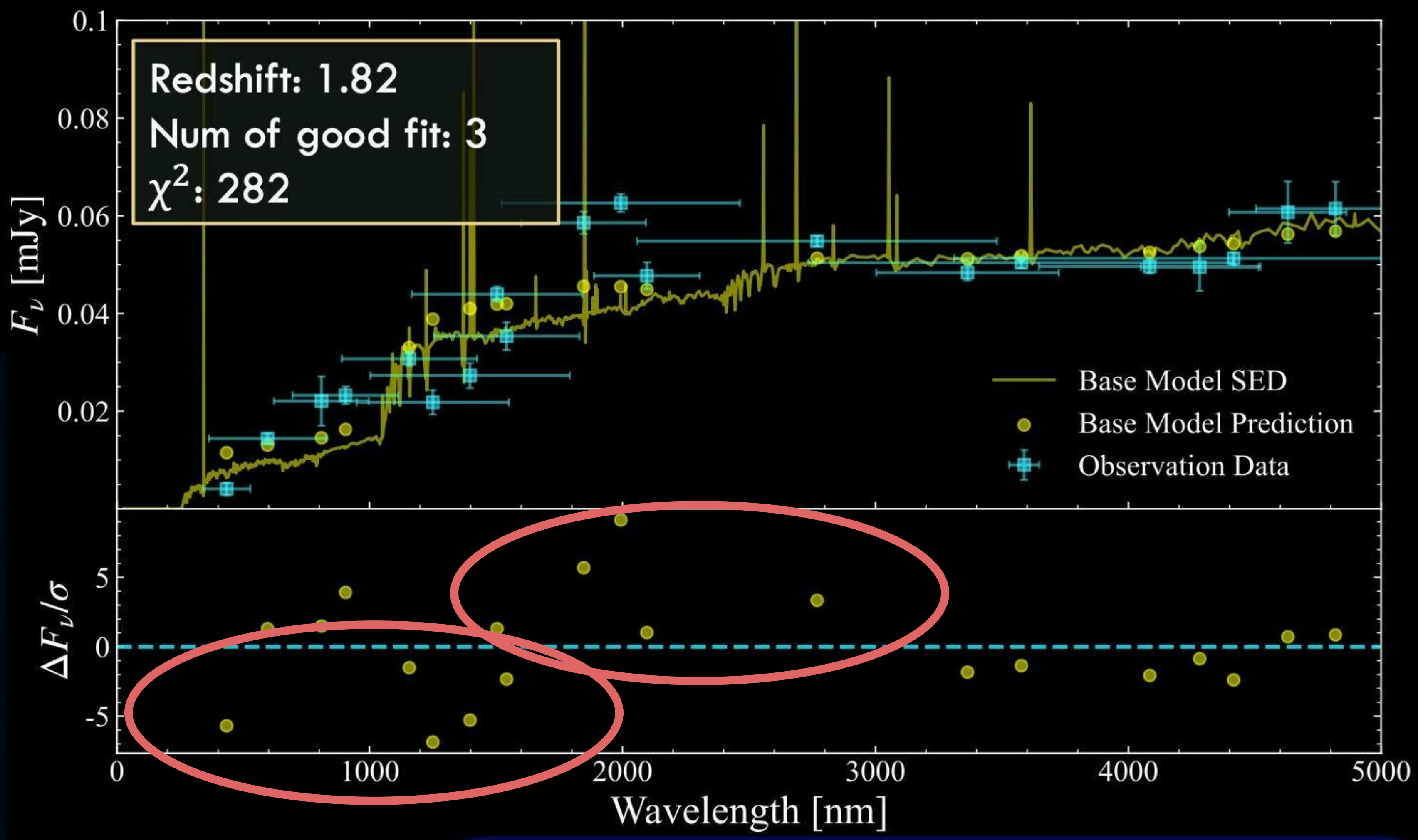
Action space is vast and transcends mathematical formalism

A default fit with an SED model

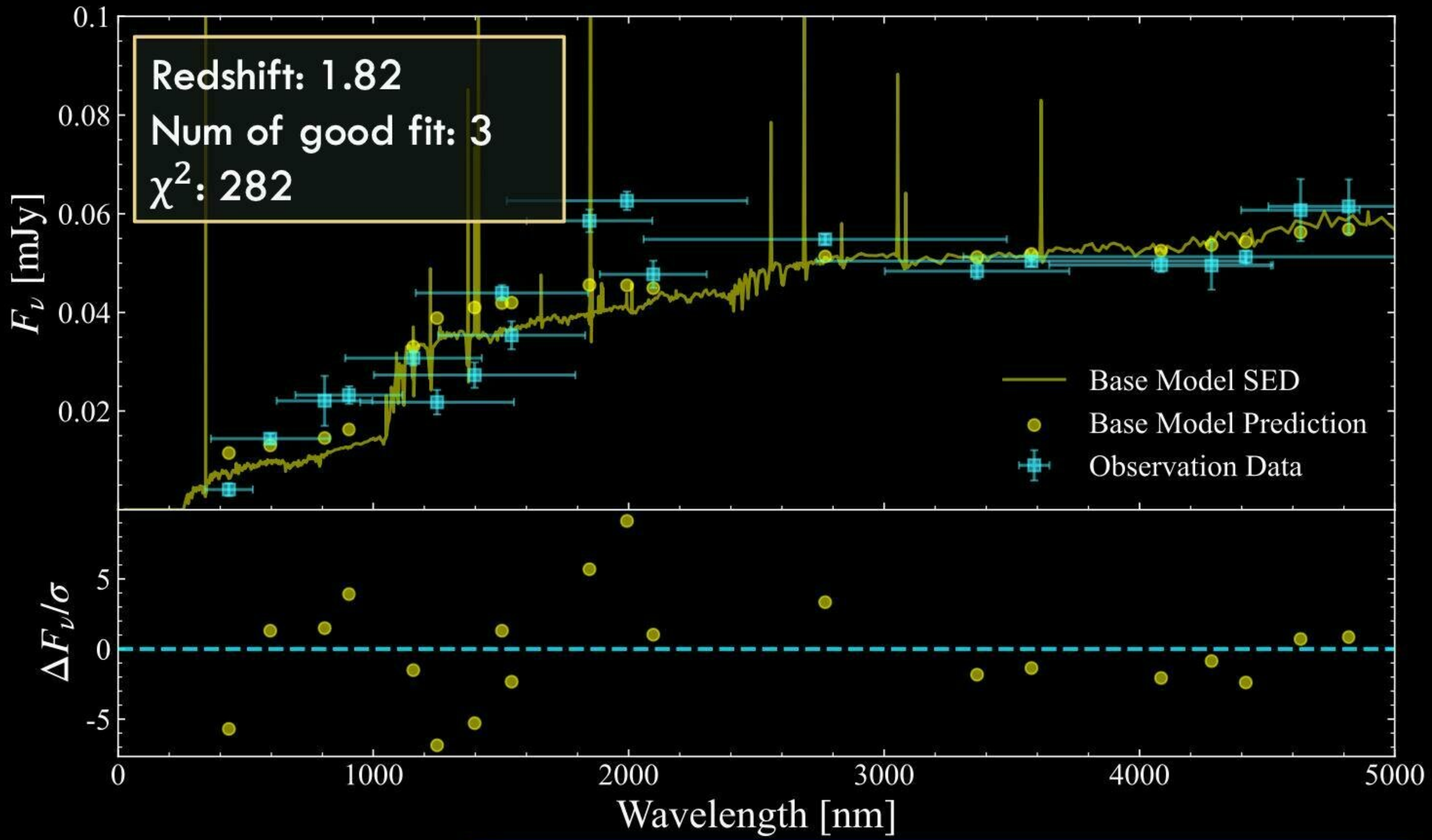


Action space is vast and transcends mathematical formalism

A default fit with an SED model

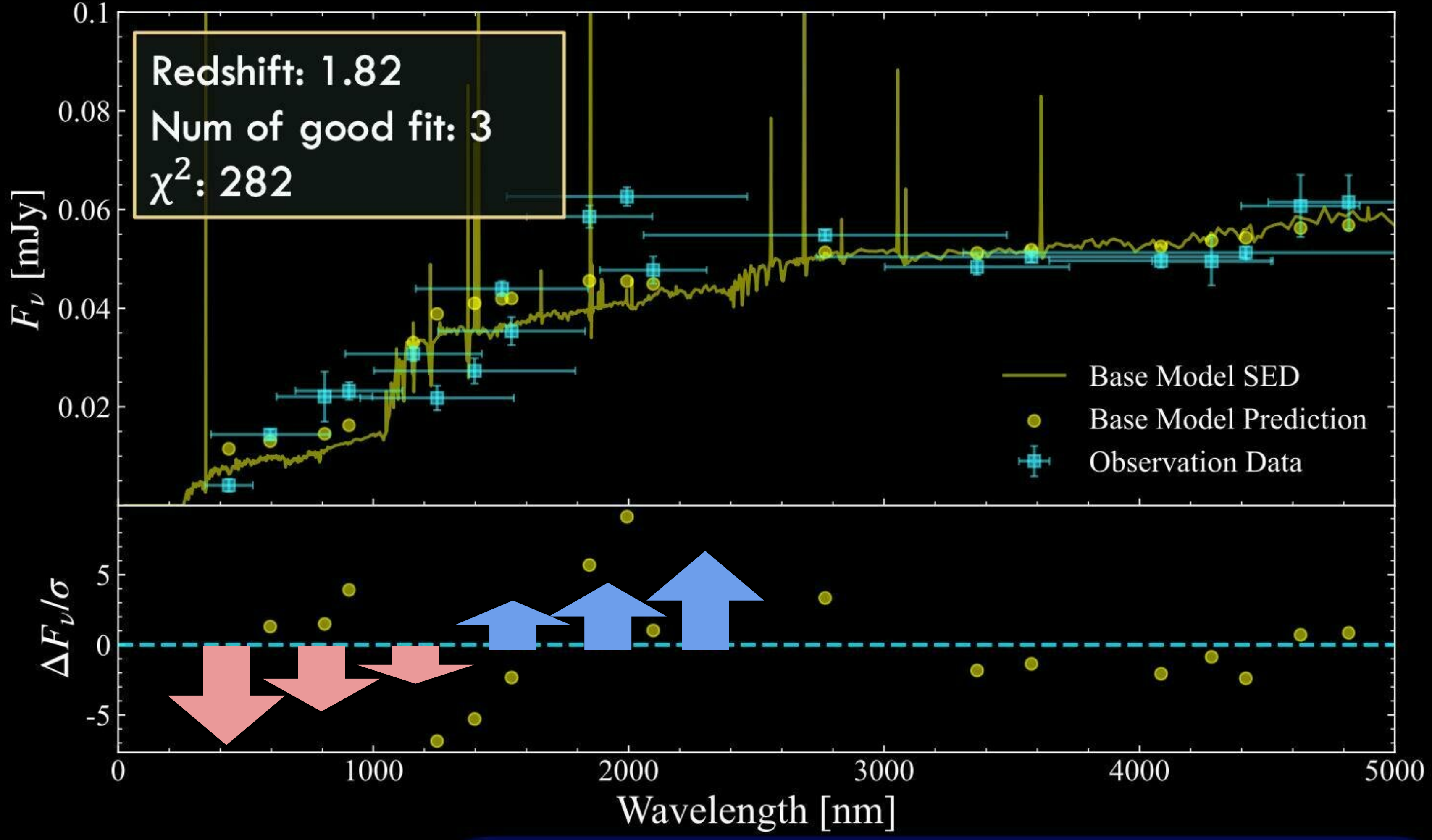


Action space is vast and transcends mathematical formalism



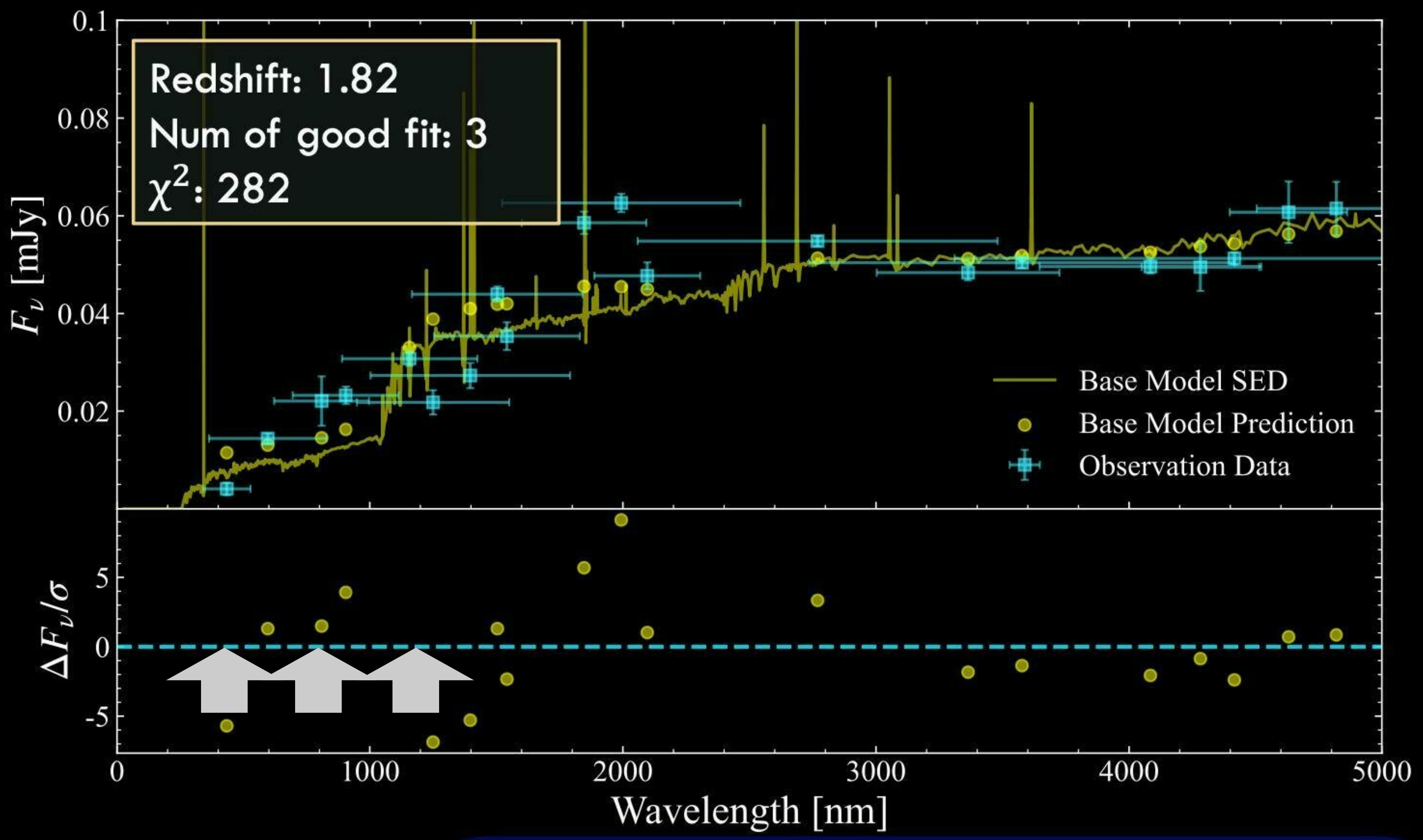
Action space is vast and transcends mathematical formalism

Extinction model ?

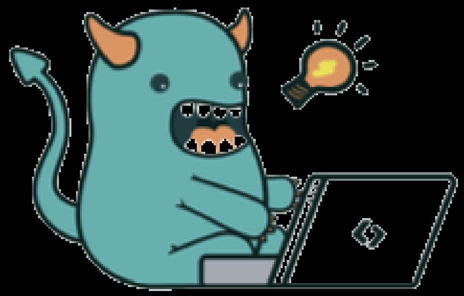


Action space is vast and transcends mathematical formalism

Young stellar population?



Introducing *Mephisto**



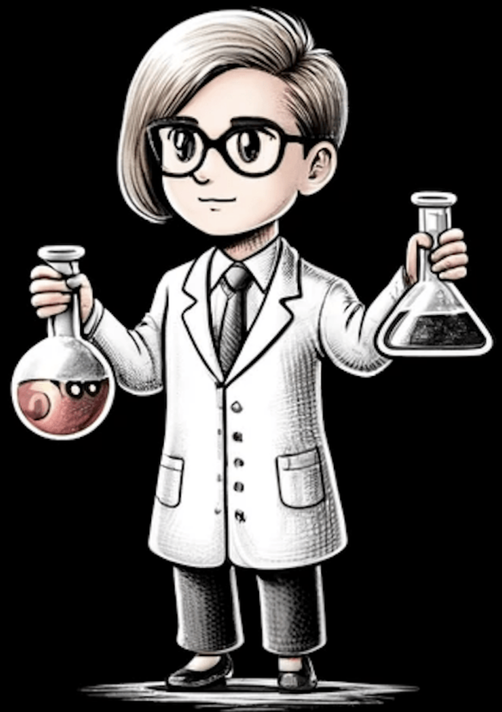
* In the classic tale of Faust, Mephisto is a demon who tempts the scholar Faust with *knowledge* and power in exchange for his soul.

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

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

Proposing actions

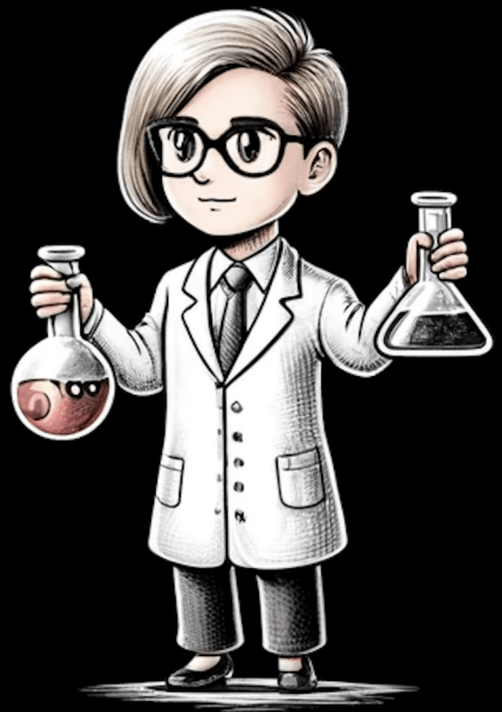
A *collaboration* of multiple AI agents (LLM models)



Proposing actions   Execute actions

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

Proposing actions

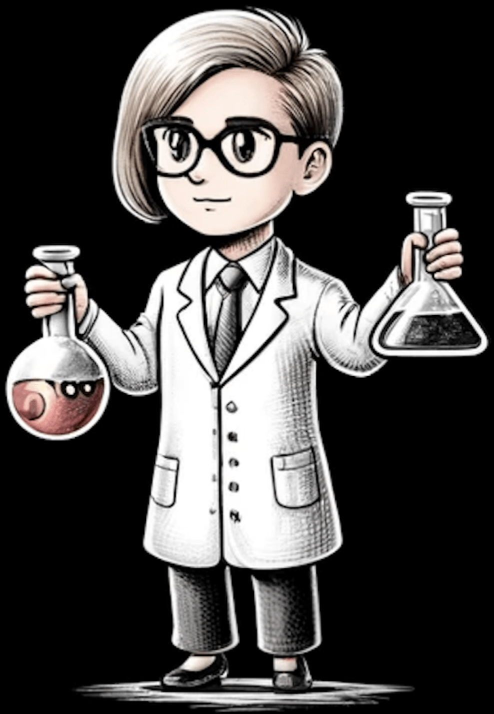


Execute actions

State evolution

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

Proposing actions



Execute actions

State evolution

Knowledge distillation

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

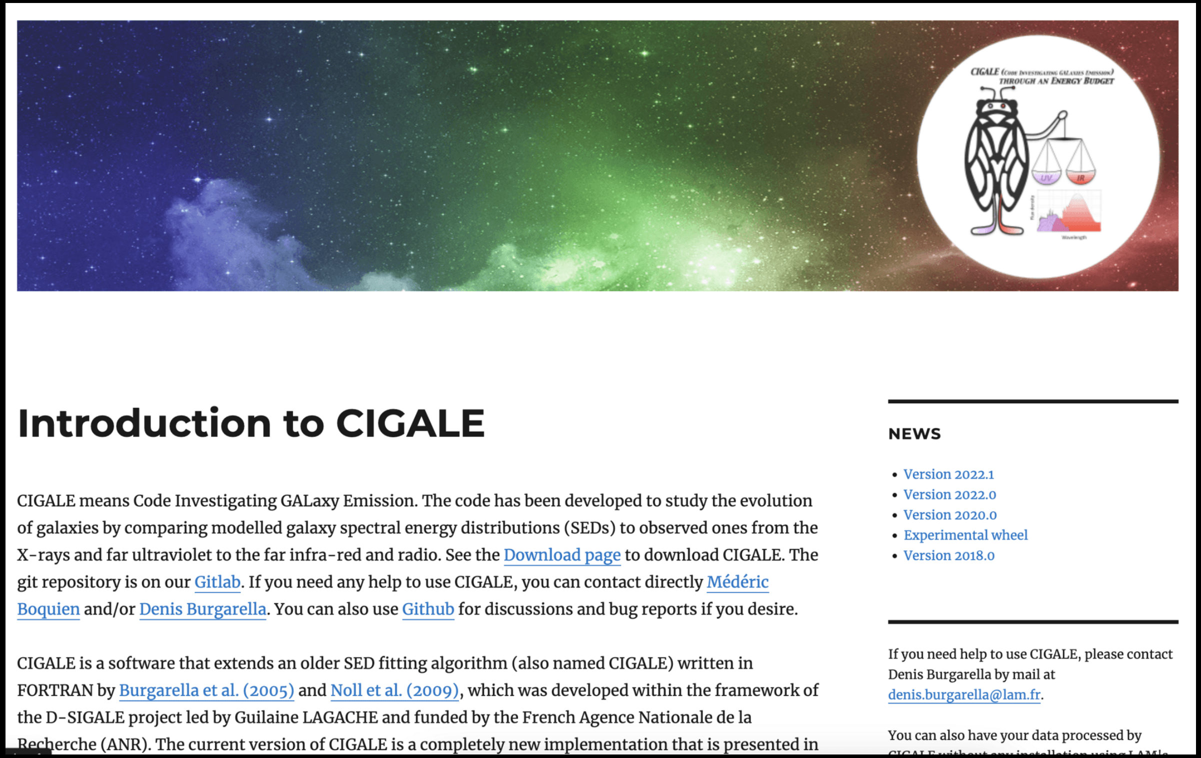
Proposing actions

Execute actions

State evolution

Knowledge distillation

A case study: Fitting galaxy spectral energy distributions



Introduction to CIGALE

CIGALE means Code Investigating GALaxy Emission. The code has been developed to study the evolution of galaxies by comparing modelled galaxy spectral energy distributions (SEDs) to observed ones from the X-rays and far ultraviolet to the far infra-red and radio. See the Download page to download CIGALE. The git repository is on our Gitlab. If you need any help to use CIGALE, you can contact directly Médéric Boquien and/or Denis Burgarella. You can also use Github for discussions and bug reports if you desire.

CIGALE is a software that extends an older SED fitting algorithm (also named CIGALE) written in FORTRAN by Burgarella et al. (2005) and Noll et al. (2009), which was developed within the framework of the D-SIGALE project led by Guilaine LAGACHE and funded by the French Agence Nationale de la Recherche (ANR). The current version of CIGALE is a completely new implementation that is presented in

NEWS

- Version 2022.1
- Version 2022.0
- Version 2020.0
- Experimental wheel
- Version 2018.0

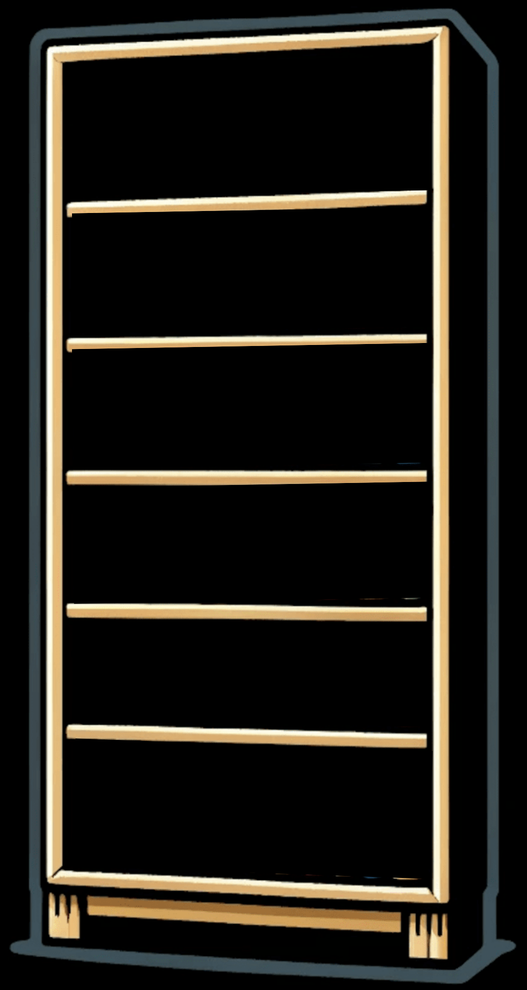
If you need help to use CIGALE, please contact Denis Burgarella by mail at denis.burgarella@lam.fr.

You can also have your data processed by

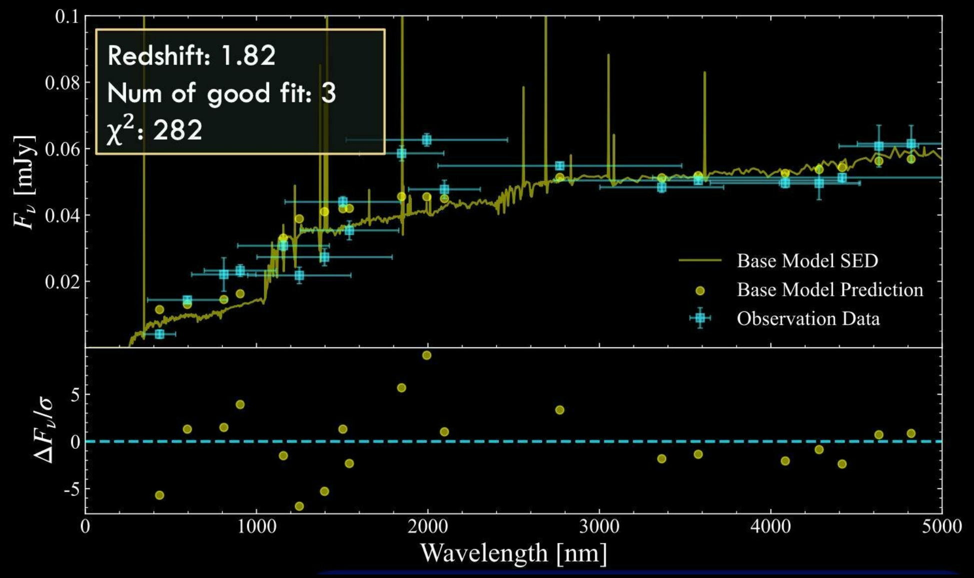
CIGALE SED Fitting Codes

JWST's JADES Survey

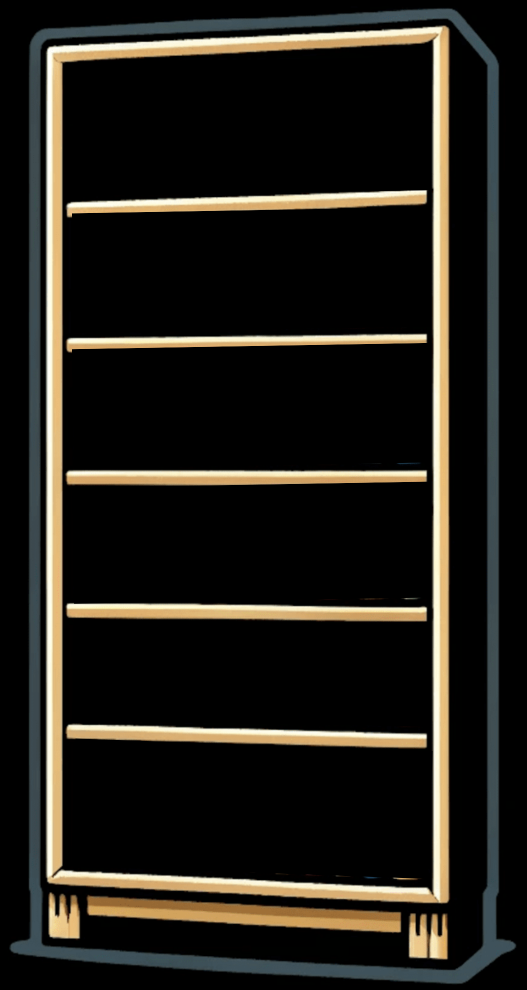
Enabling AI to collect *"knowledge"* through exploration



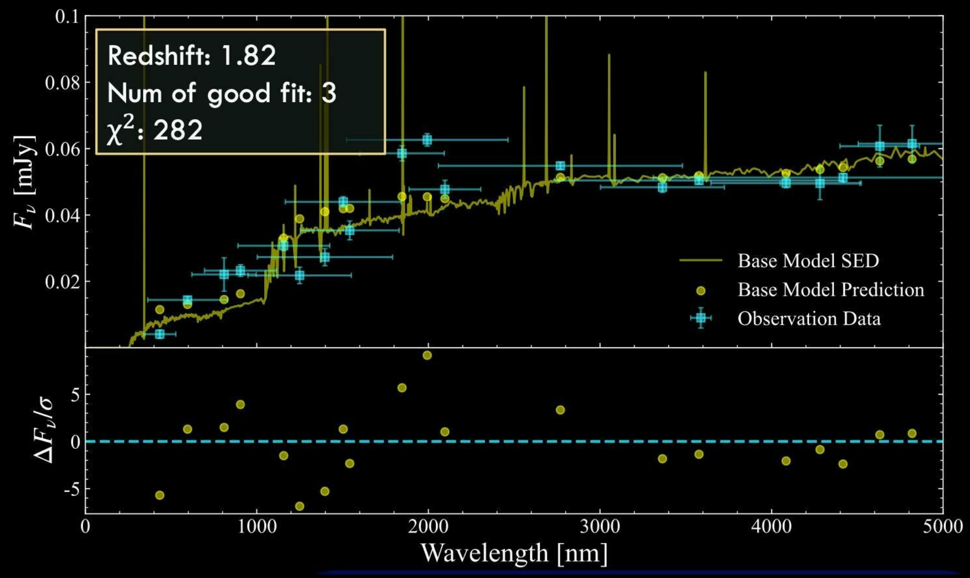
Knowledge base



Enabling AI to collect *"knowledge"* through exploration

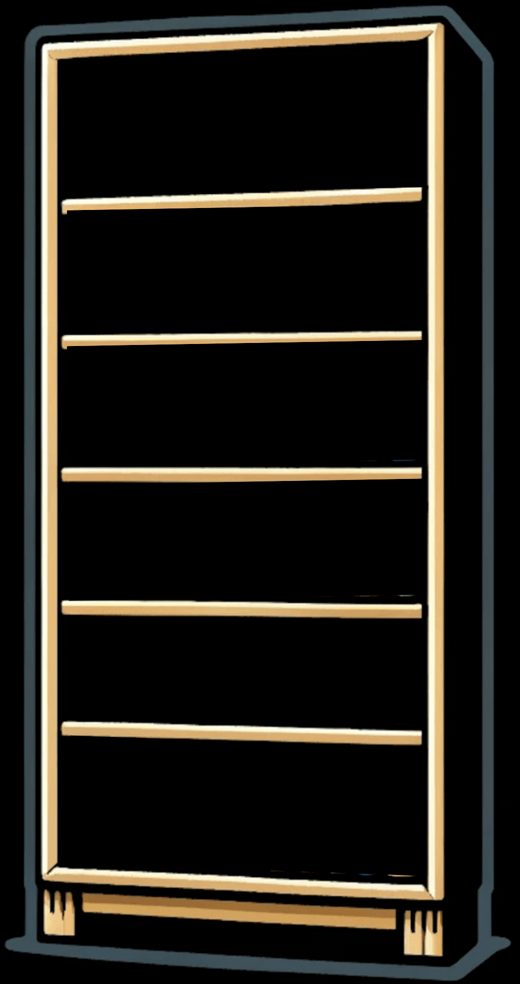


Knowledge base

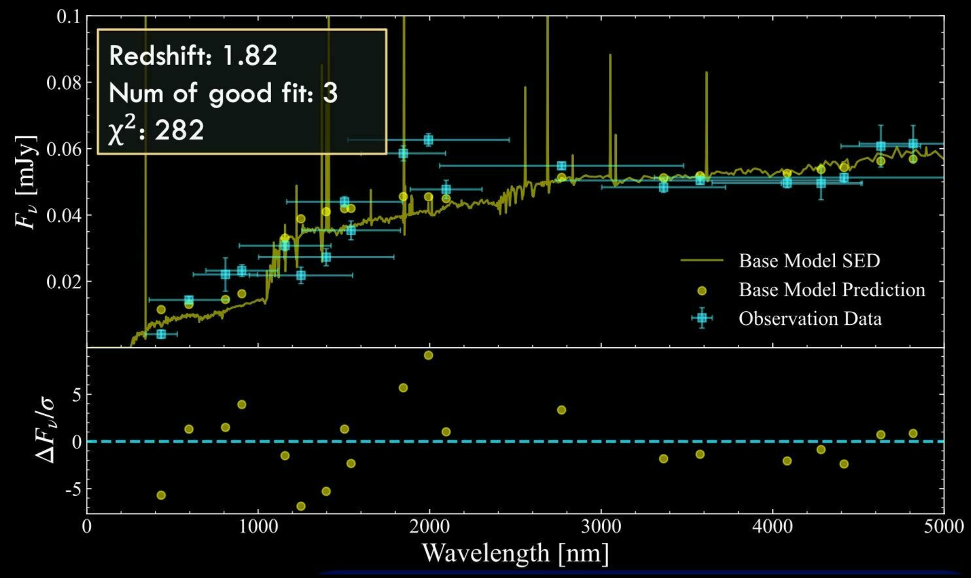


Proposing Actions - e.g., different physical models / parameter range

Enabling AI to collect *"knowledge"* through exploration



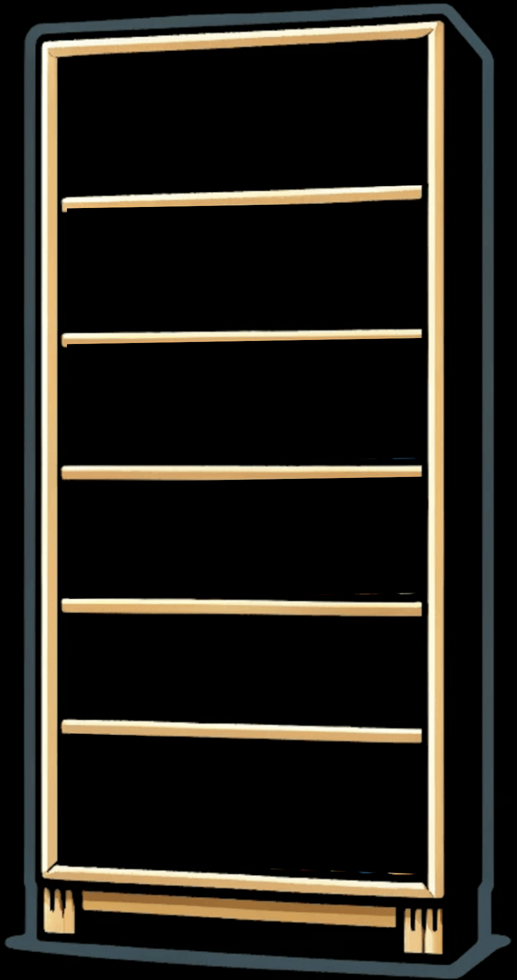
Knowledge base



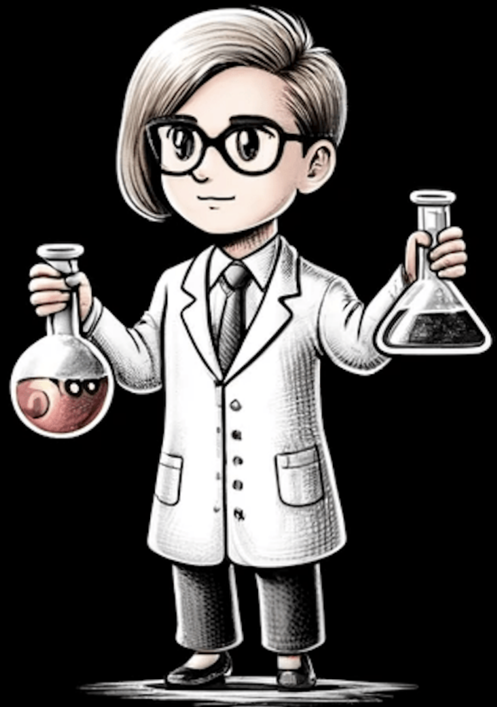
1 2 3 4

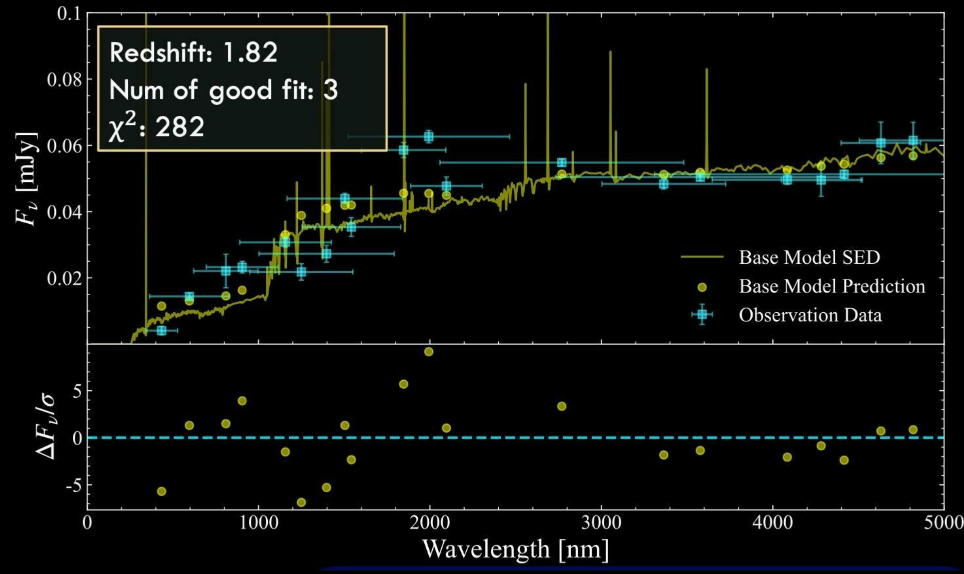
Proposing Actions - e.g., different physical models / parameter range

Enabling AI to collect *"knowledge"* through exploration



Knowledge base

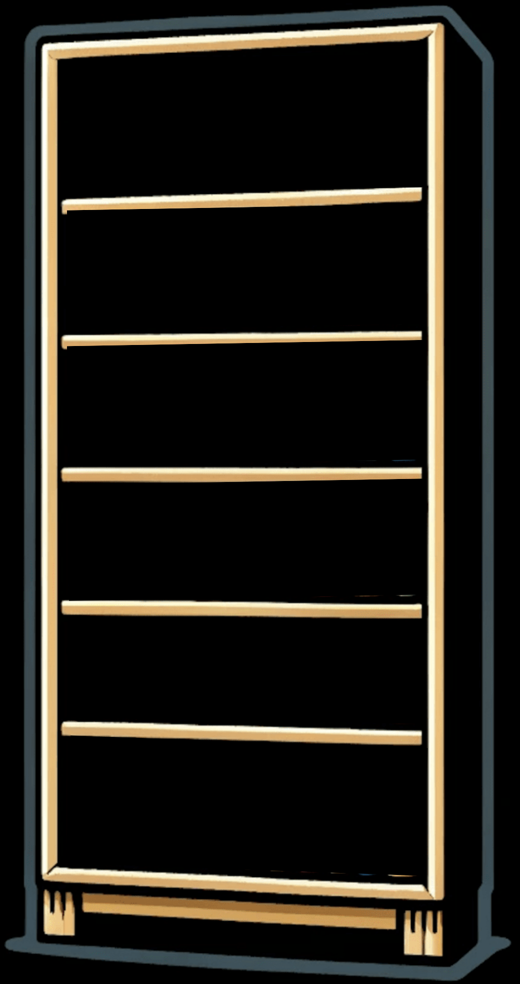




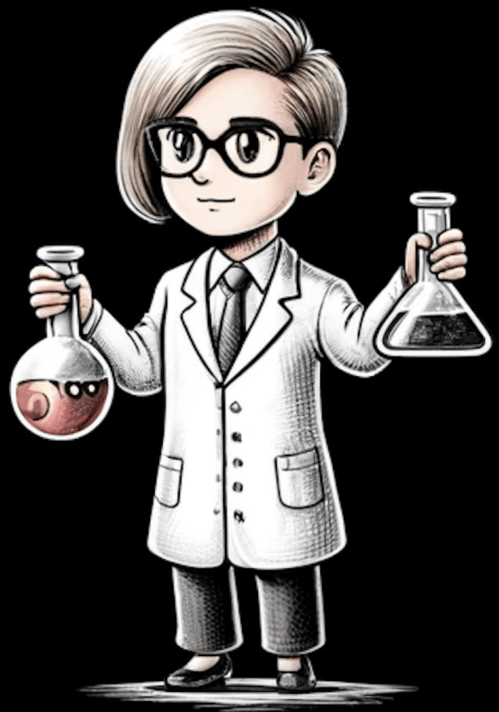
1 2 3 4

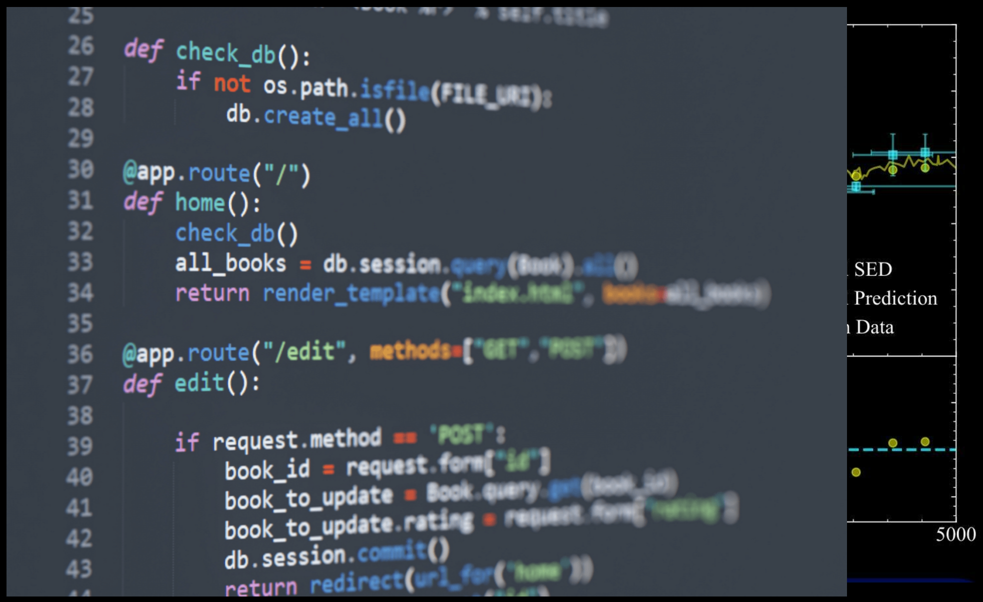
Execute Actions - write configuration files, run the codes, automously

Enabling AI to collect *"knowledge"* through exploration



Knowledge base

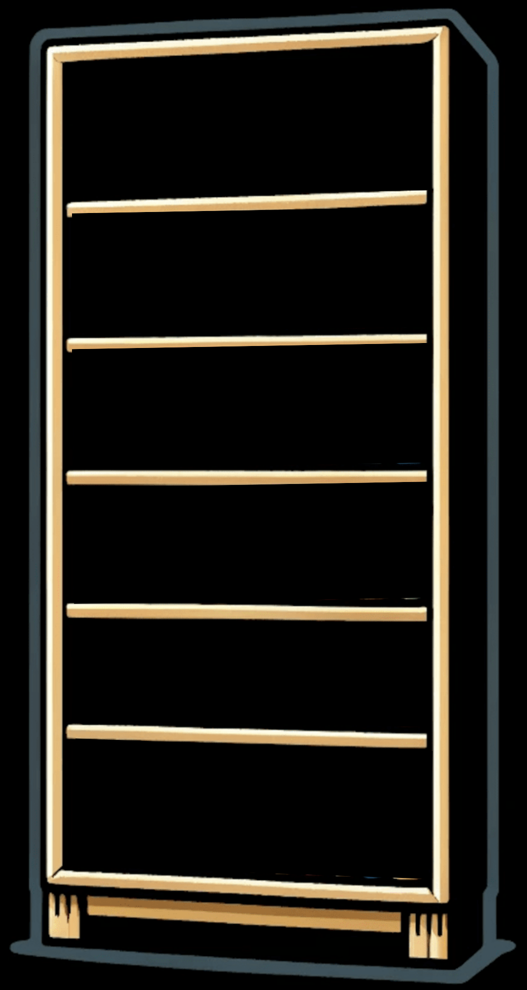




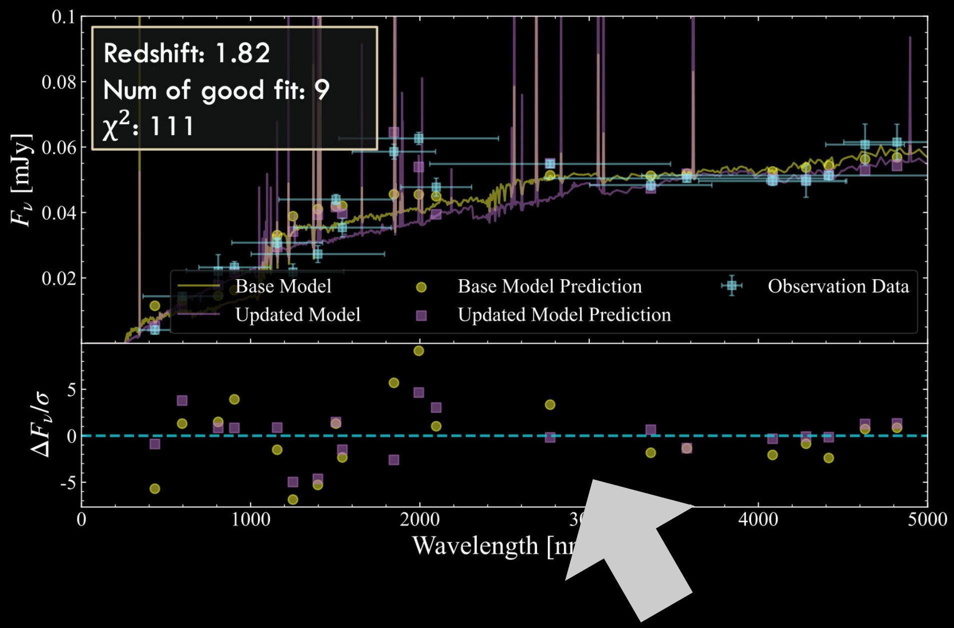
1 2 3 4

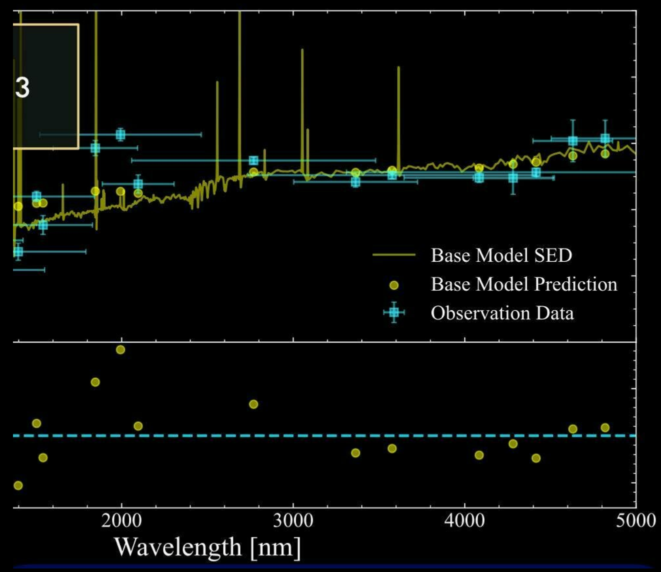
Execute Actions - write configuration files, run the codes, automously

Enabling AI to collect *"knowledge"* through exploration

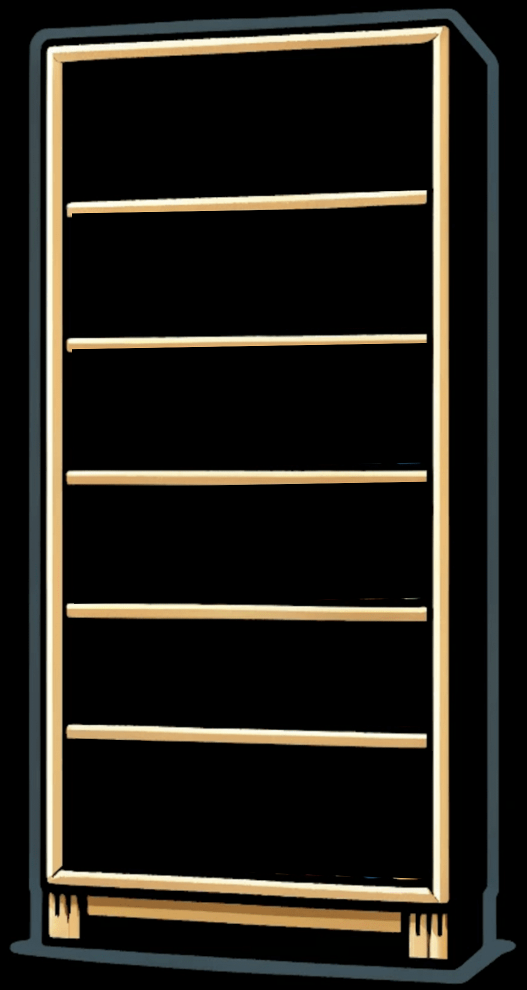


Knowledge base

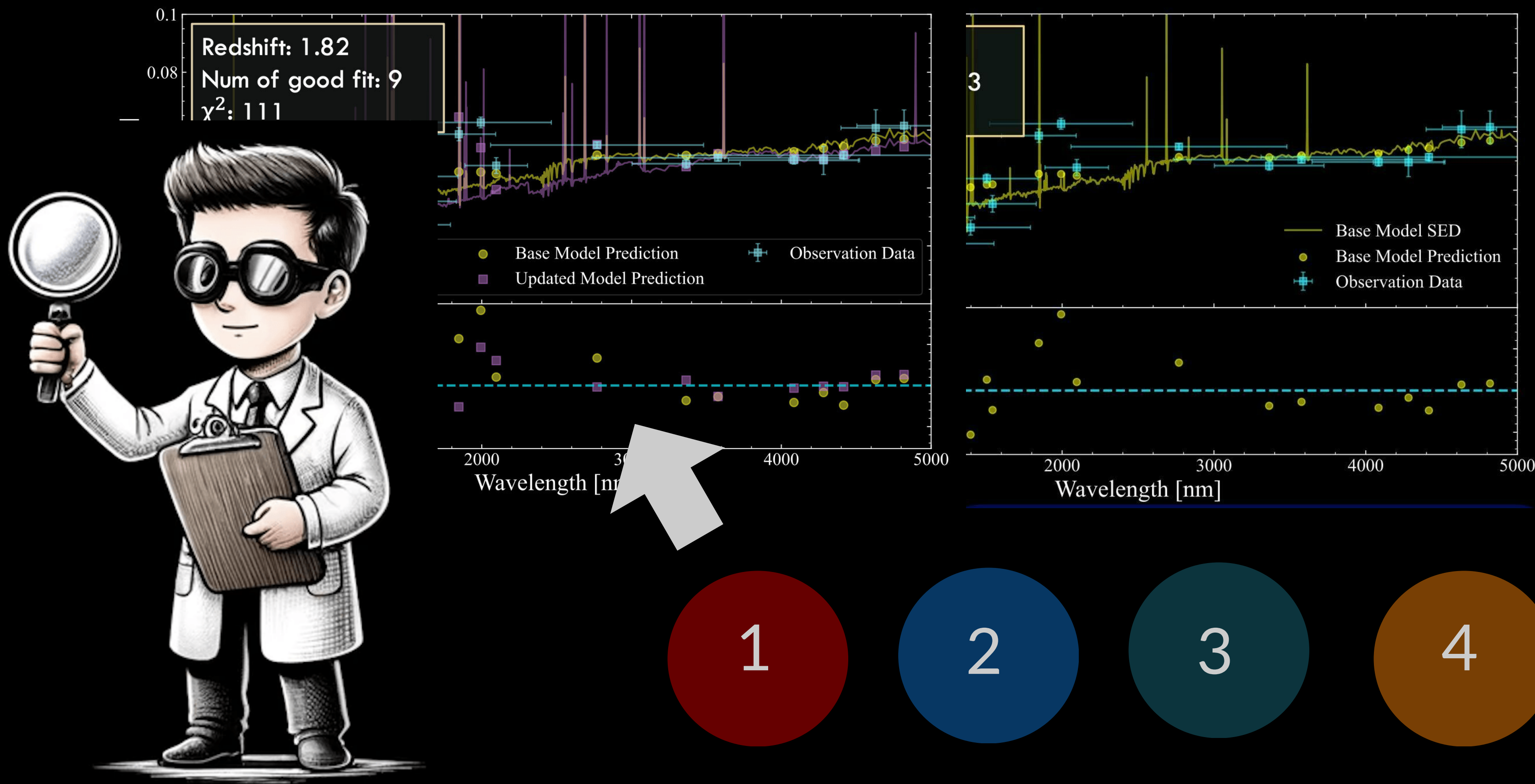




Enabling AI to collect *"knowledge"* through exploration



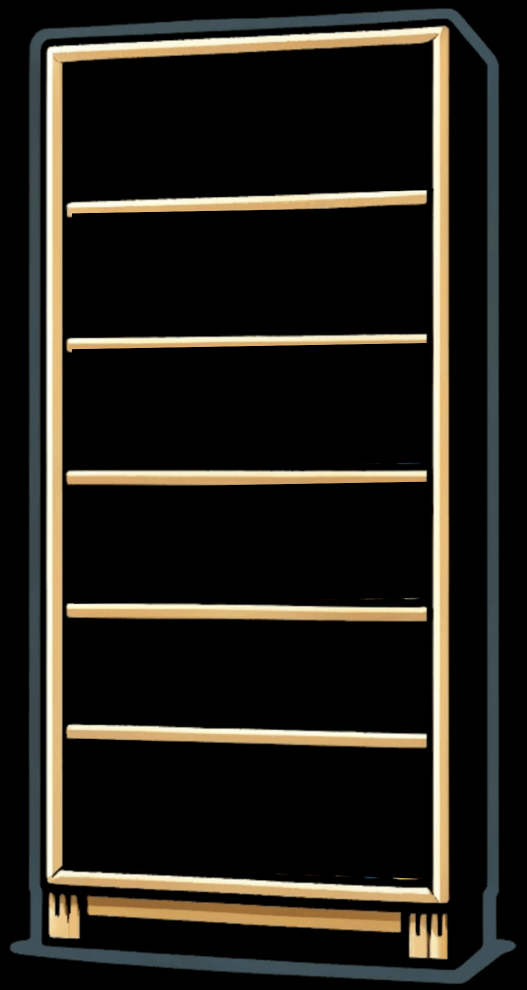
Knowledge base



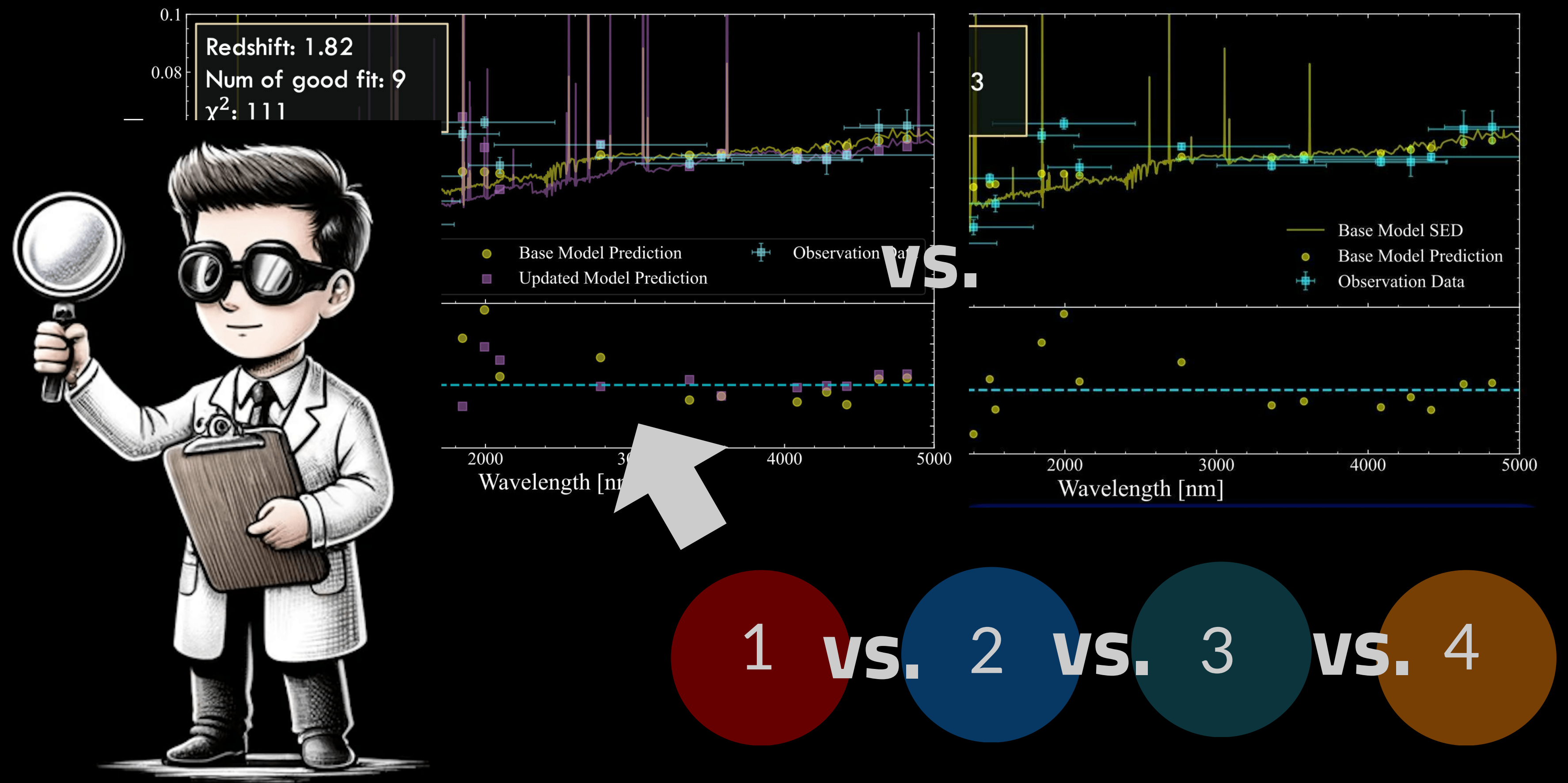
1 2 3 4

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

Enabling AI to collect *"knowledge"* through exploration

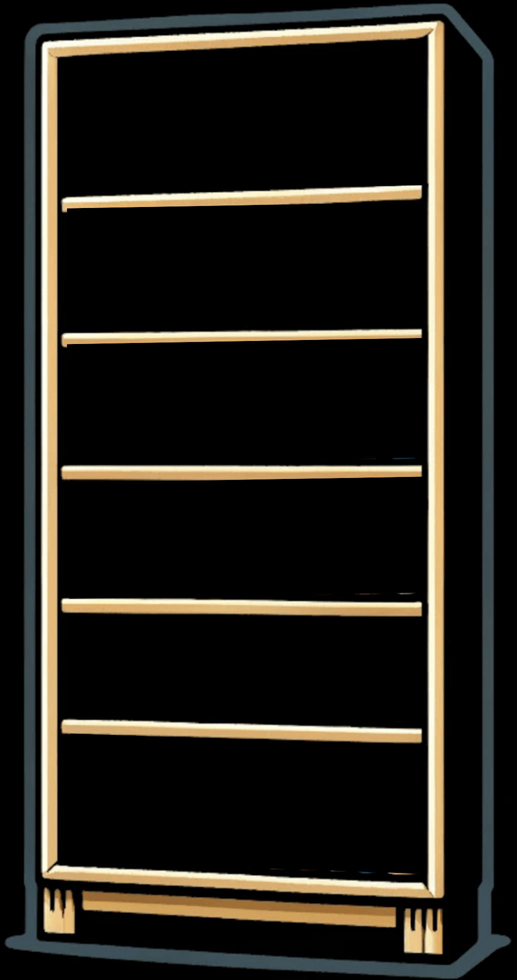


Knowledge base

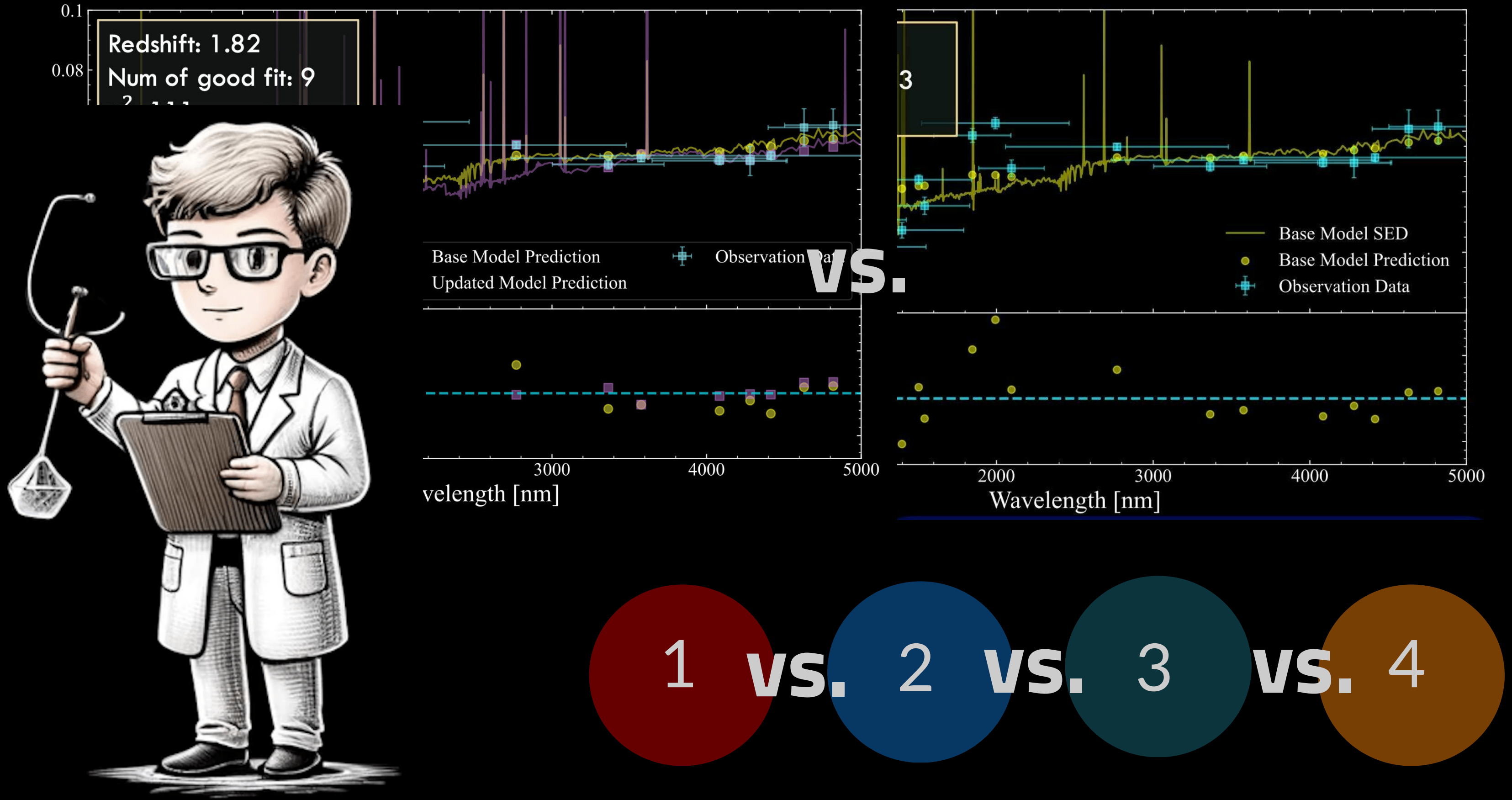


State Evaluation - evaluate the results (beyond a single error metric)

Enabling AI to collect *"knowledge"* through exploration

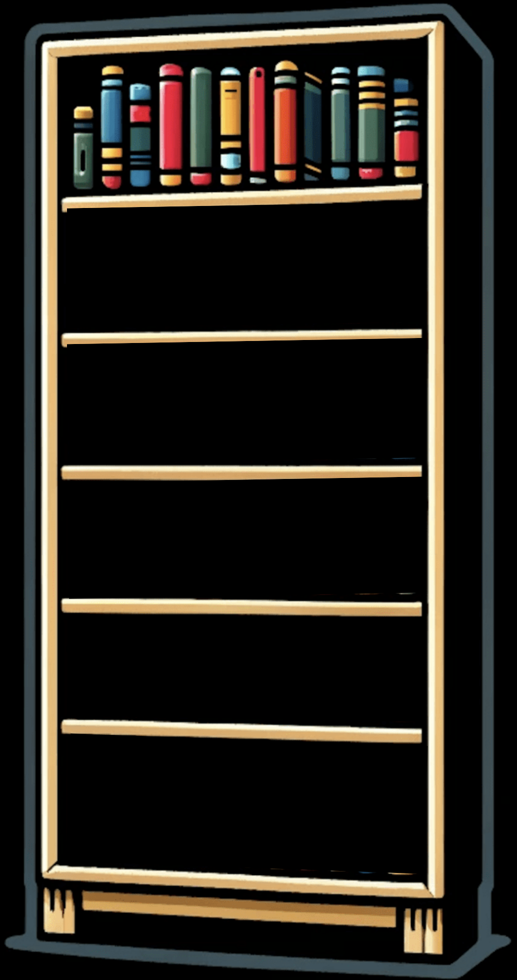


Knowledge base

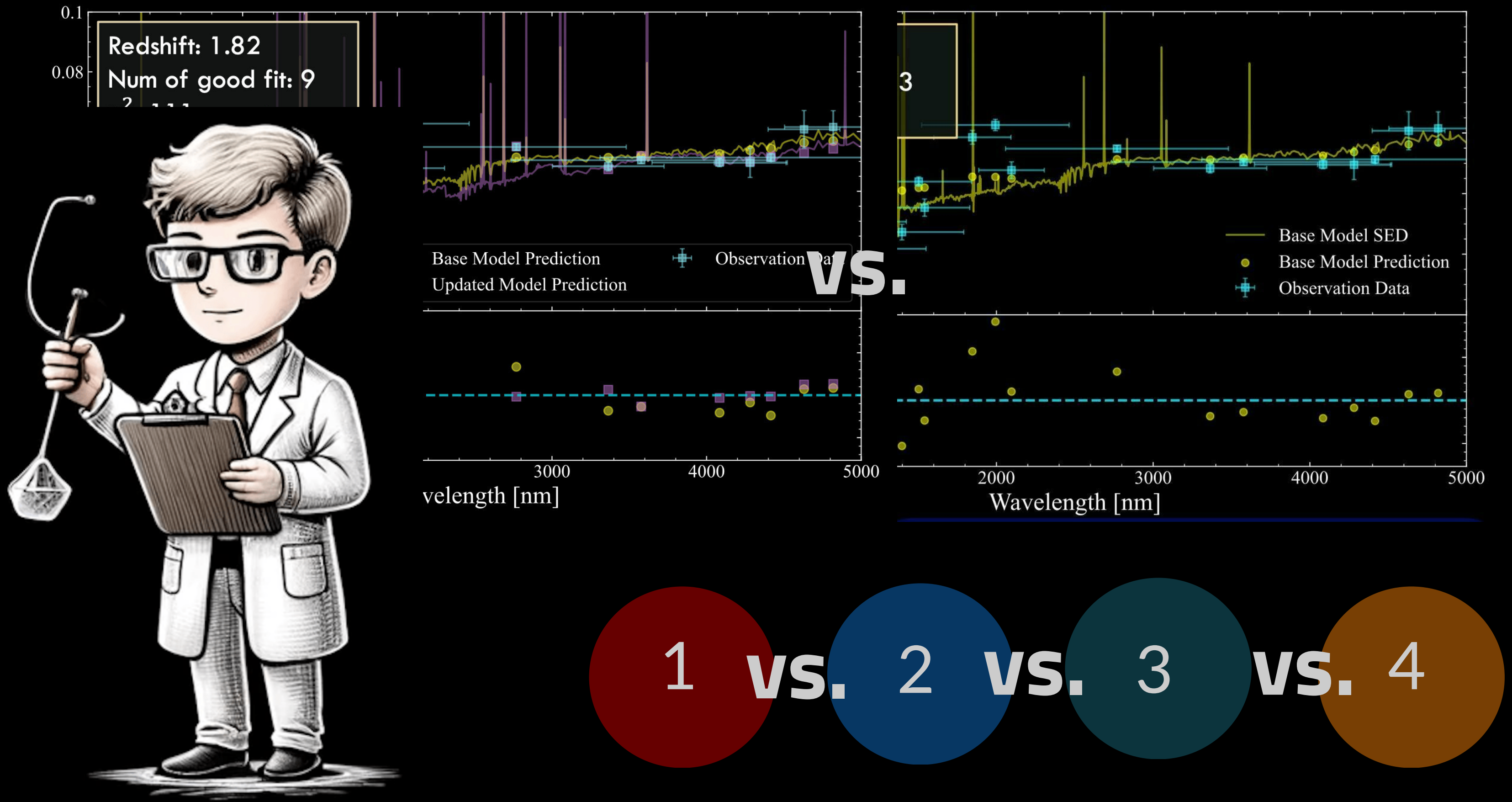


Knowledge Distillation - summarise useful actions given the previous state

Enabling AI to collect *"knowledge"* through exploration



Knowledge base



Knowledge Distillation - summarise useful actions given the previous state

Example of learned "knowledge"

Example of learned "knowledge"

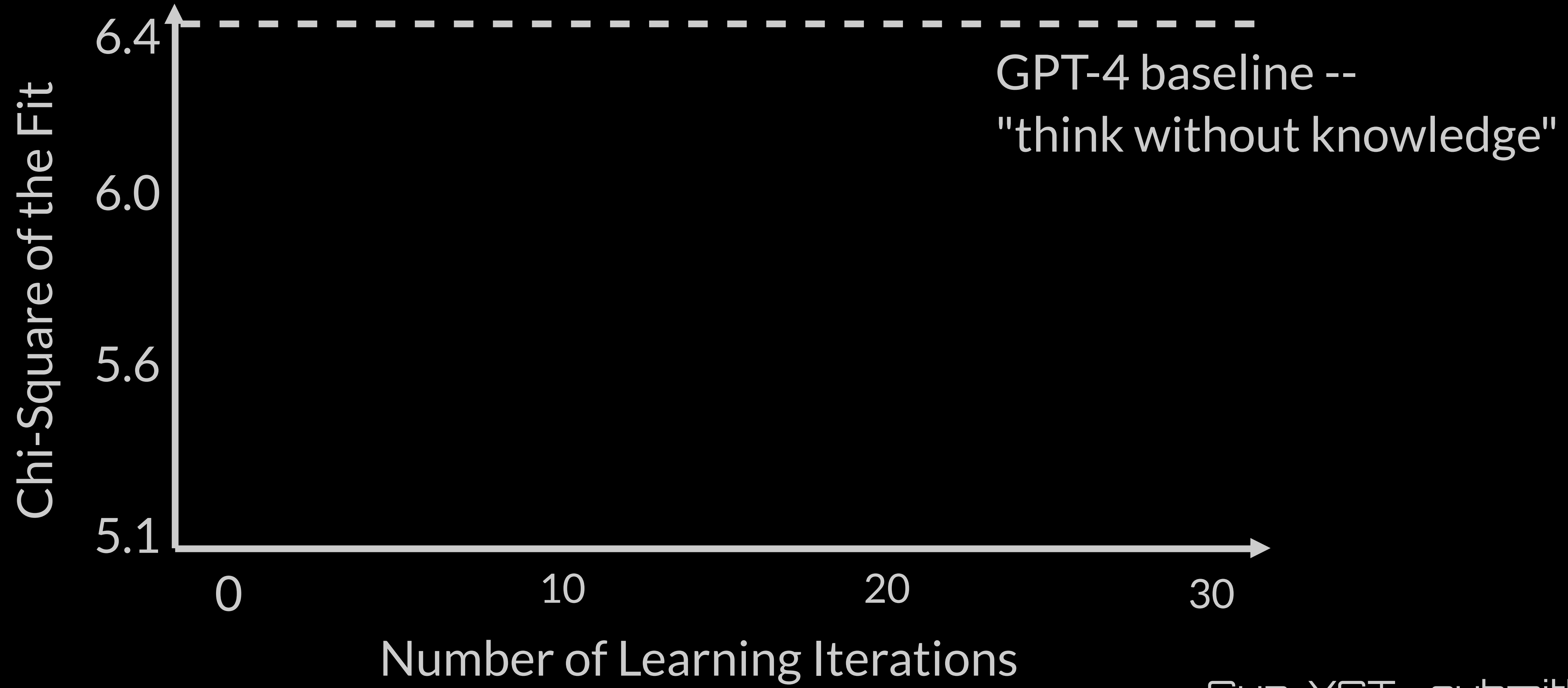
" If the fit is *overestimated in the UV and optical* bands,

Example of learned "knowledge"

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

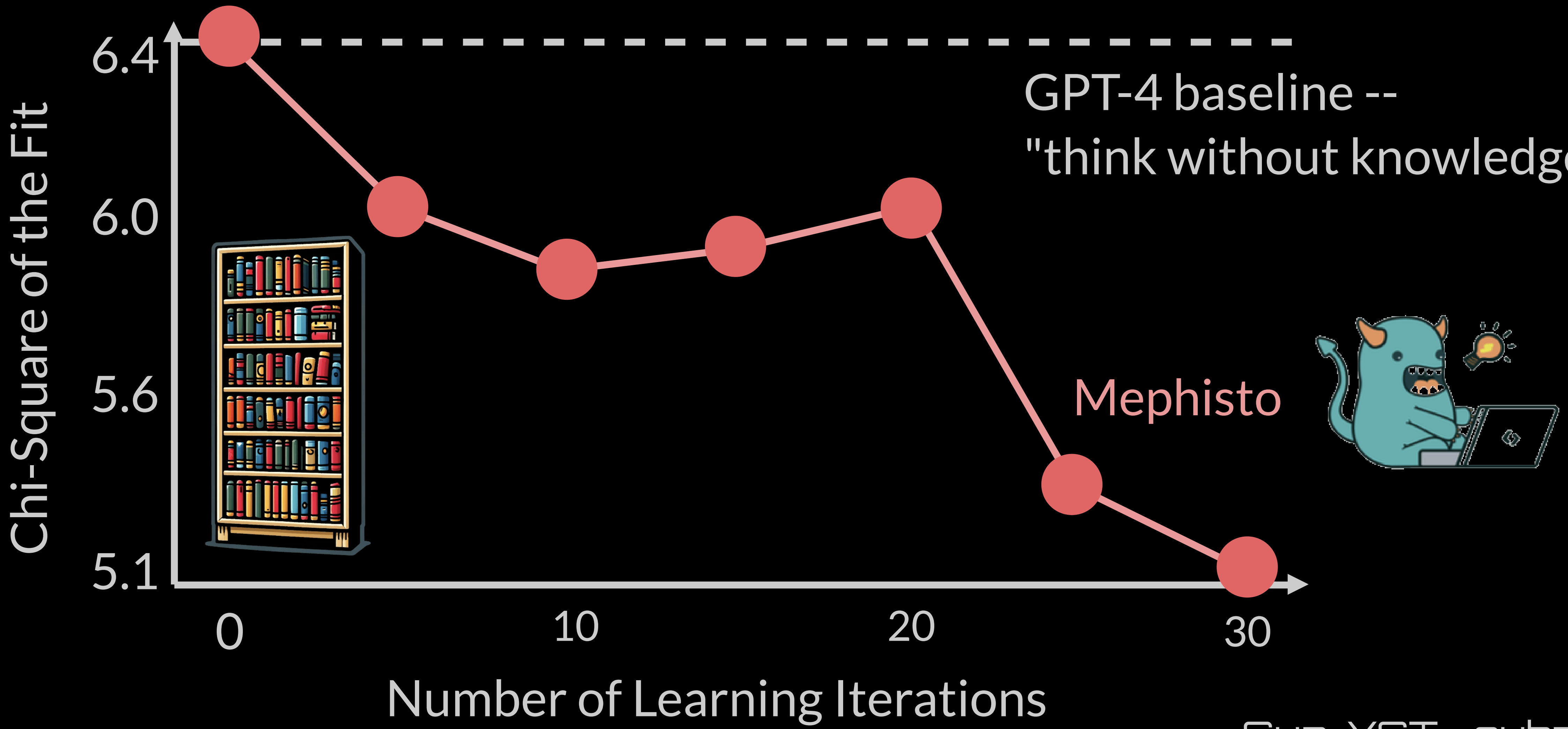
increasing the E_BV_lines parameter may lead to a better fit by accounting for more *dust attenuation* in these bands. "

Reinforcement learning with LLM *outperforms* native LLMs



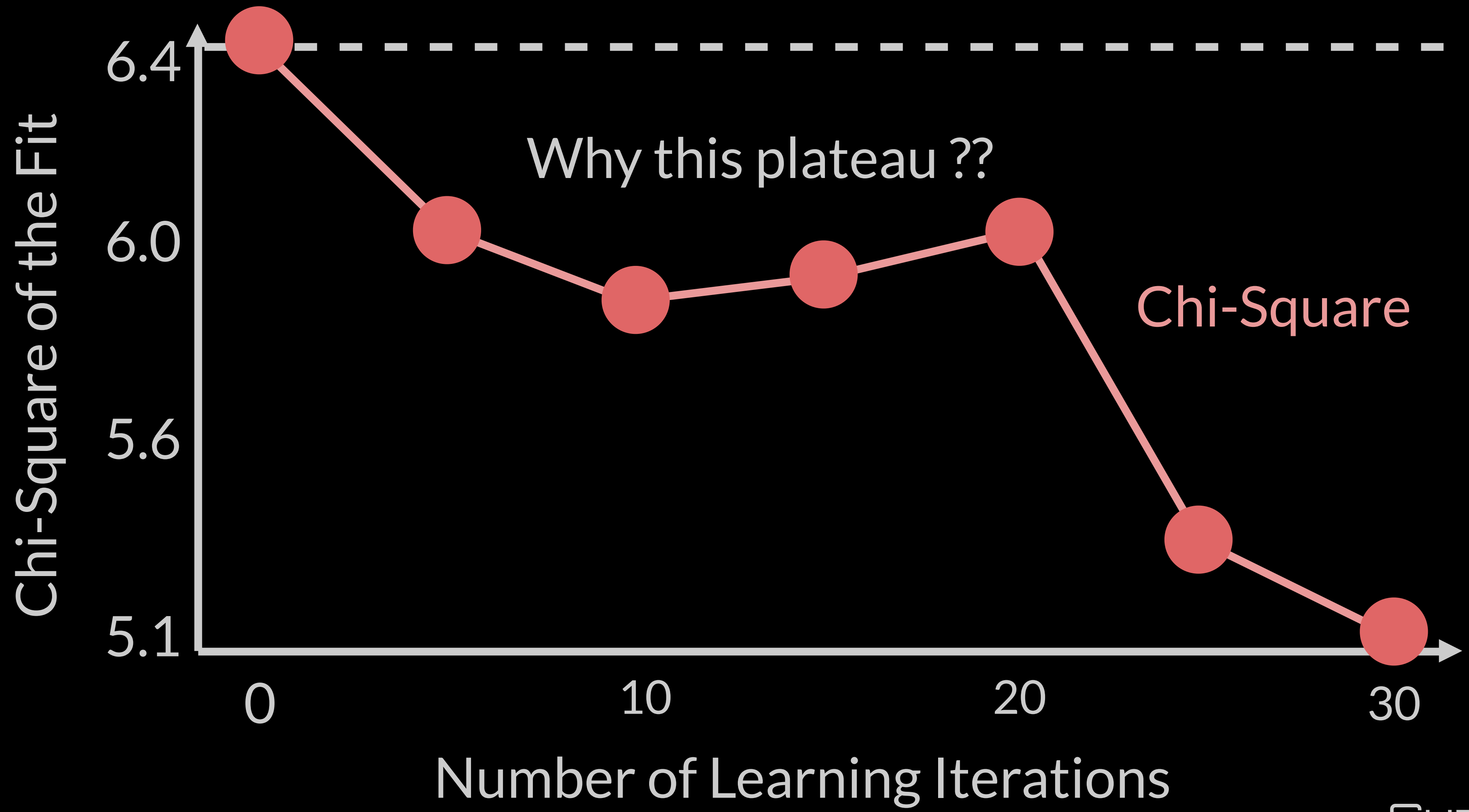
Sun, YST+, submitted

Reinforcement learning with LLM *outperforms* native LLMs



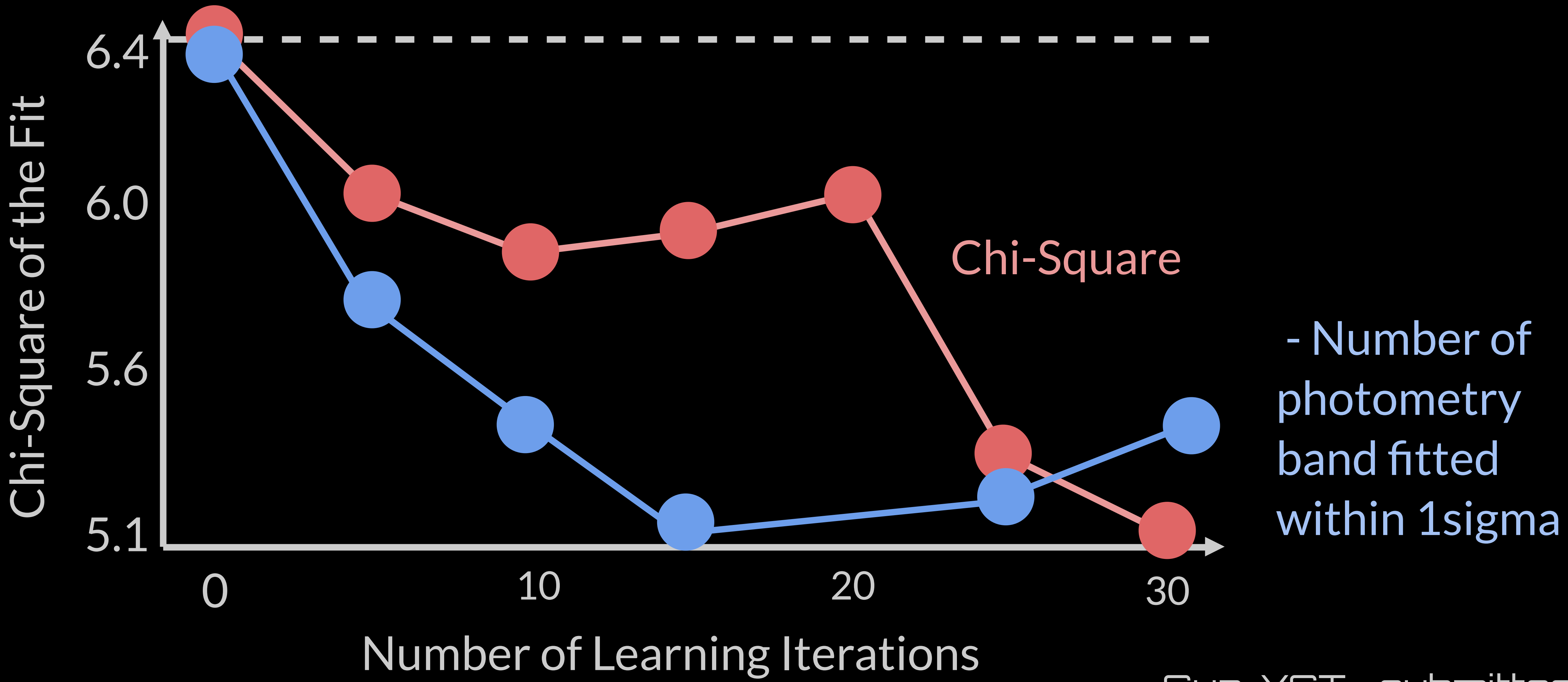
Sun, YST+, submitted

Reinforcement learning with LLM *outperforms* native LLMs



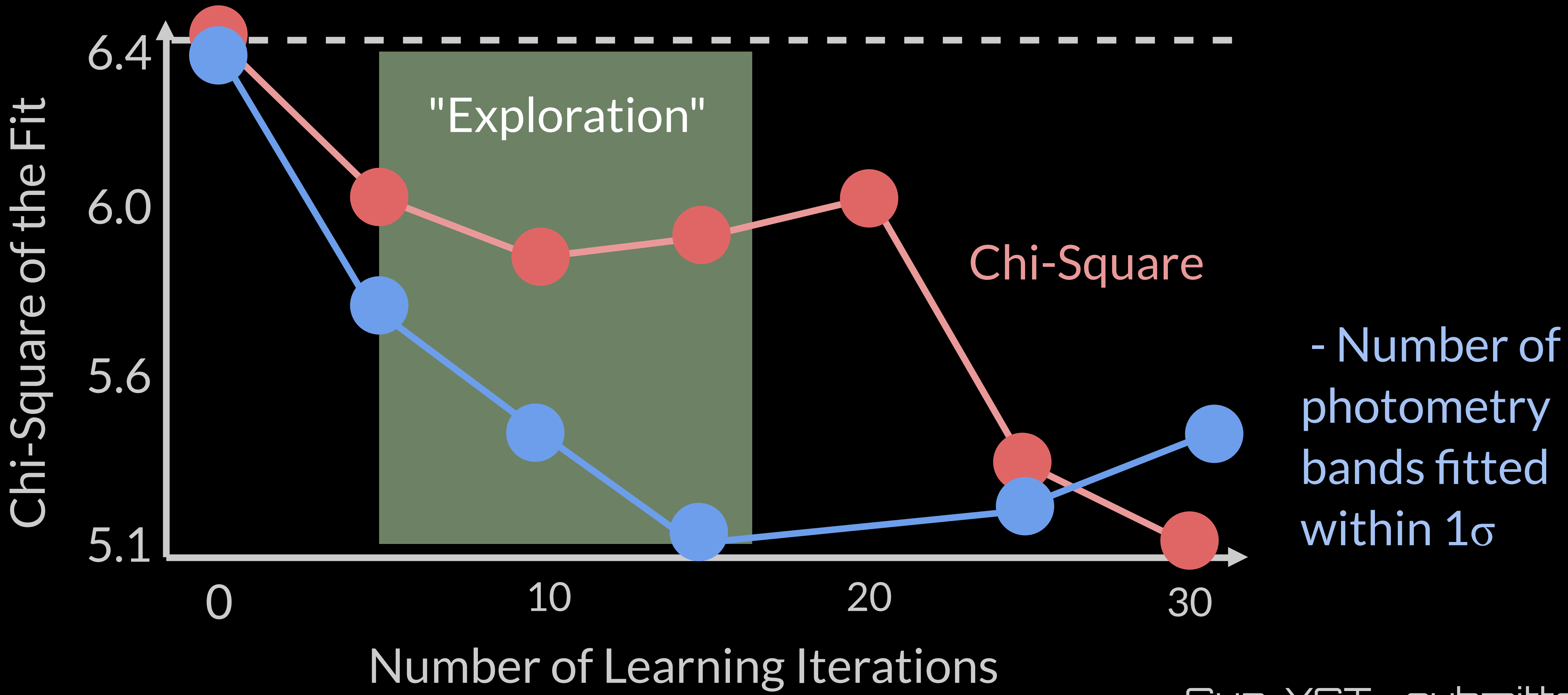
Sun, YST+, submitted

Reinforcement learning with LLM *outperforms* native LLMs



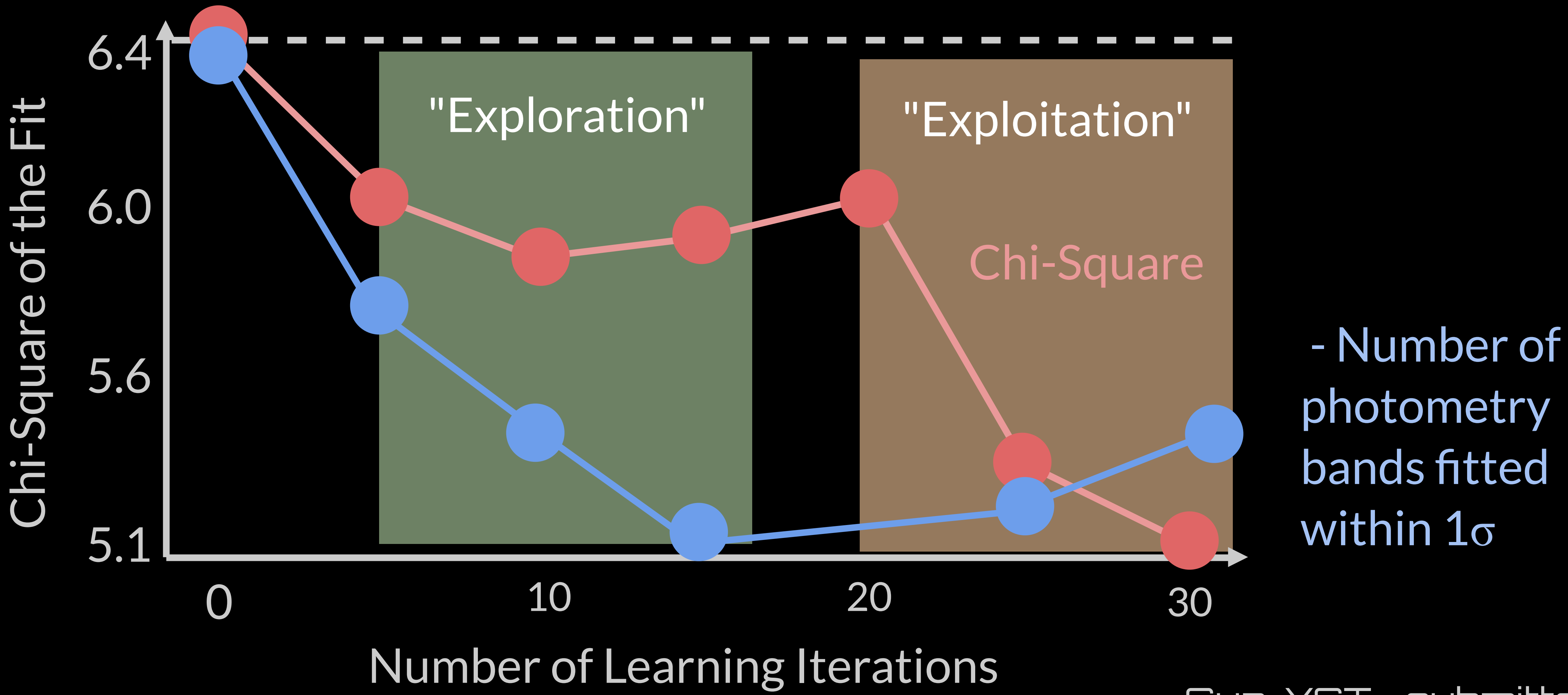
Sun, YST+, submitted

Reinforcement learning with LLM *outperforms* native LLMs



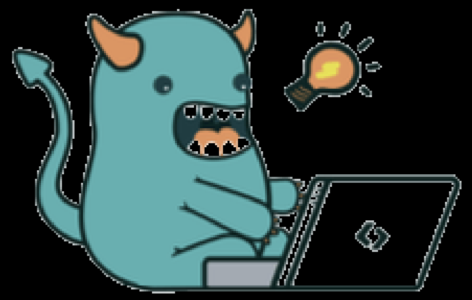
Sun, YST+, submitted

Reinforcement learning with LLM *outperforms* native LLMs

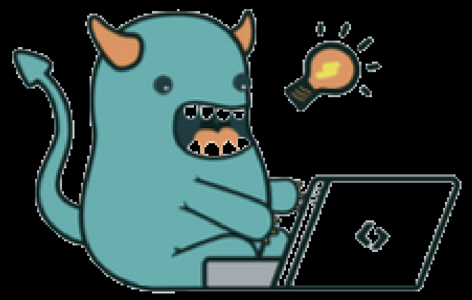


Sun, YST+, submitted

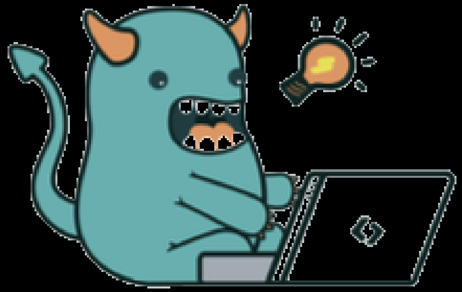
While we are far from AGI, LLMs are prolific enough to understand *basic causality through actions*



Reason about all astronomical objects in the cosmos



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

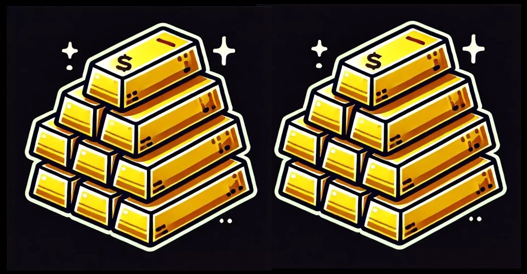


Provided that we have *a capable model* that can generate inference quickly and *cost efficiently*....

capable model **vs.** cost efficiency

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

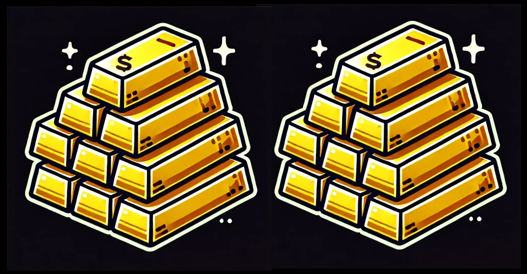
capable model **vs.** cost efficiency



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

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

capable model **vs.** cost efficiency



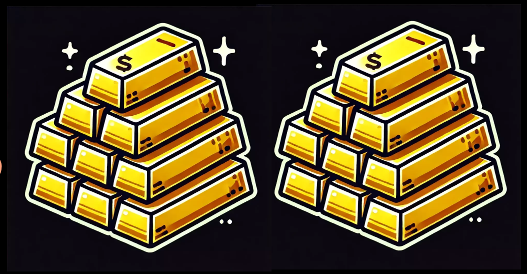
e.g., GPT-4o (this study)

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

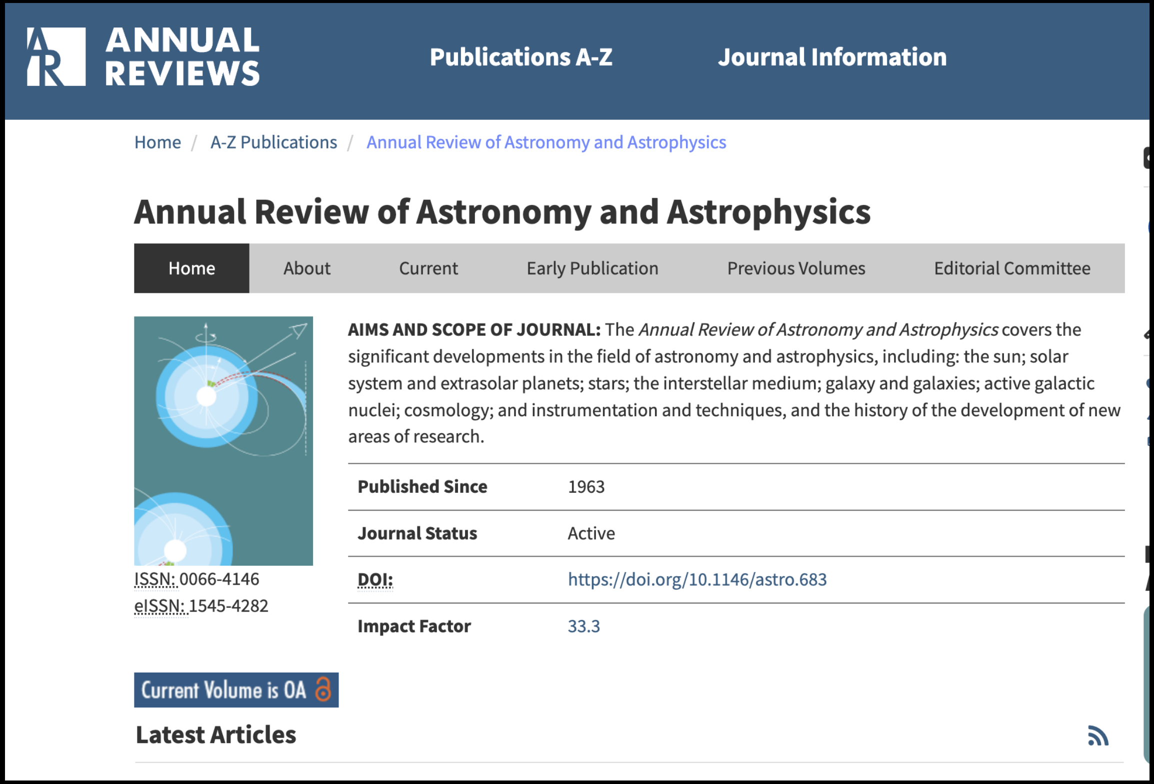
e.g., *Roman* Space Telescope, *Euclid* Space Telescope

~ approximately the build cost

- Where are we in AI for Astronomy - What's the problem?
- What could be the AlphaFold moment for astronomy?
- How do we get there in a cost-effective way?

- Where are we in AI for Astronomy - What's the problem?
- What could be the AlphaFold moment for astronomy?
- How do we get there in a cost-effective way?

The curation of an astronomy *benchmark* dataset



(i.e. as part of the Aurora-GPT effort with the Argonne National Laboratory)

But how do we know which model is better?

WHO WINS THE ASTRONOMY JEOPARDY!

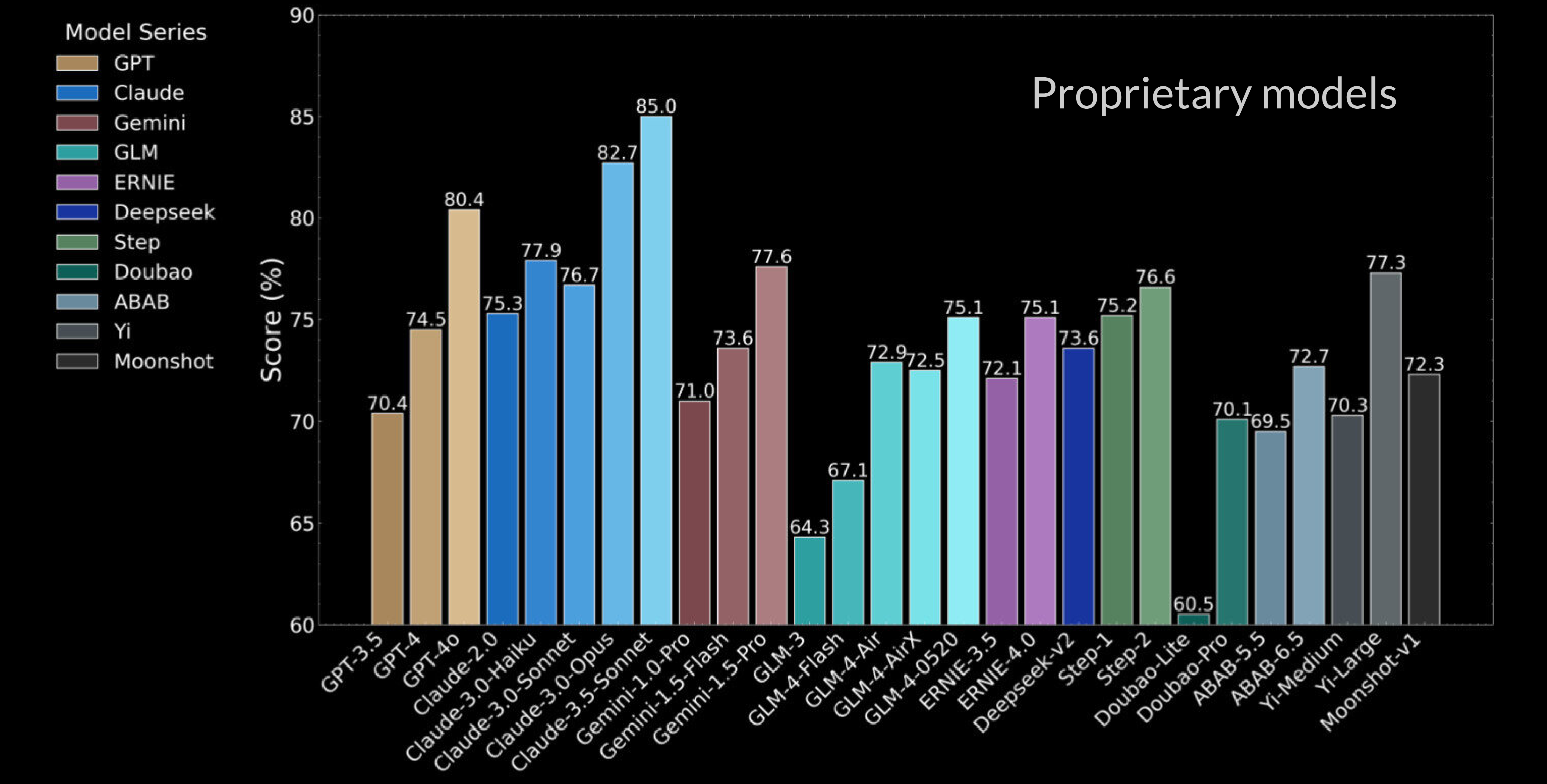
YUAN-SEN TING, ET AL.

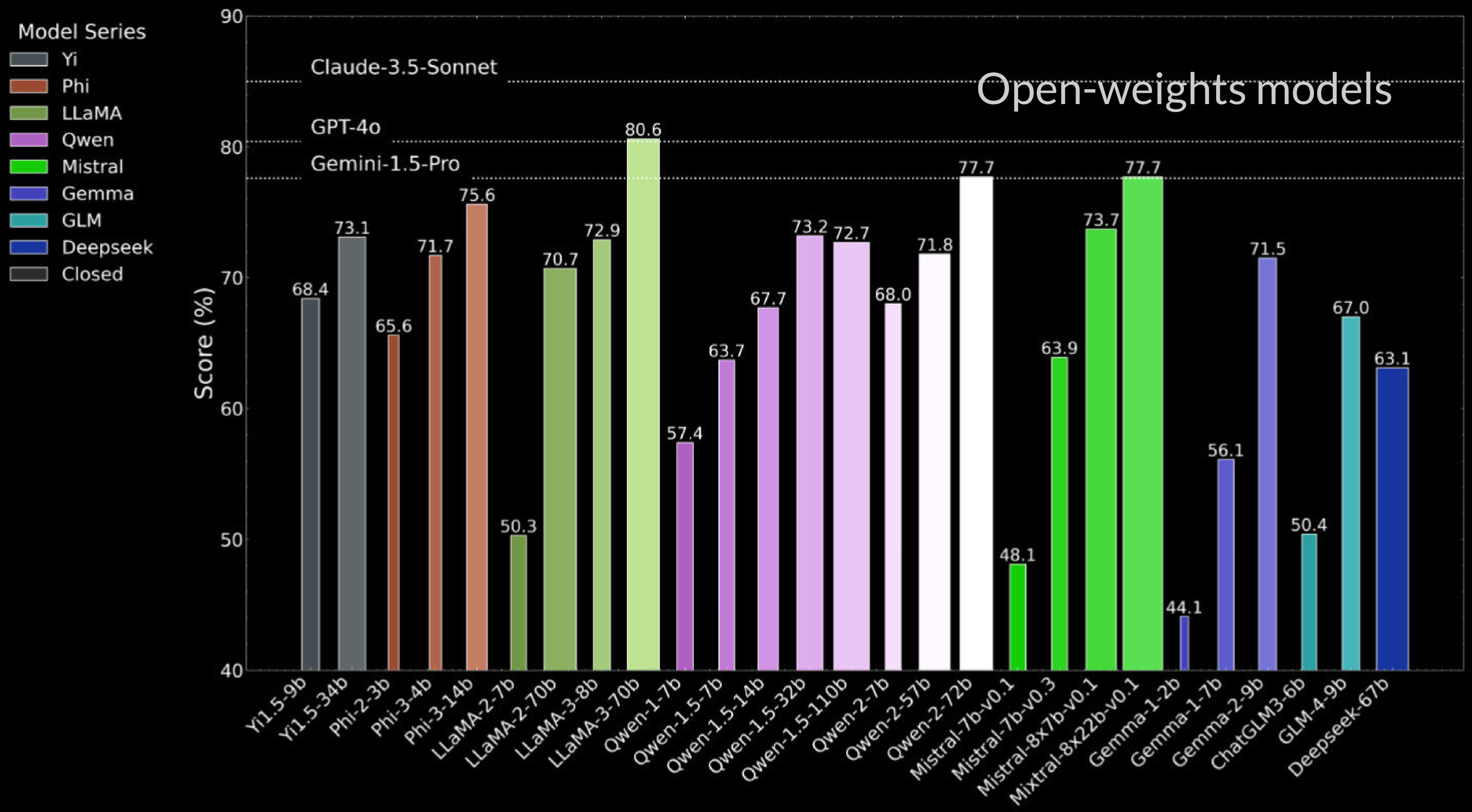
Submitted to ApJL

40 pages of all you need to know about LLM for astronomy

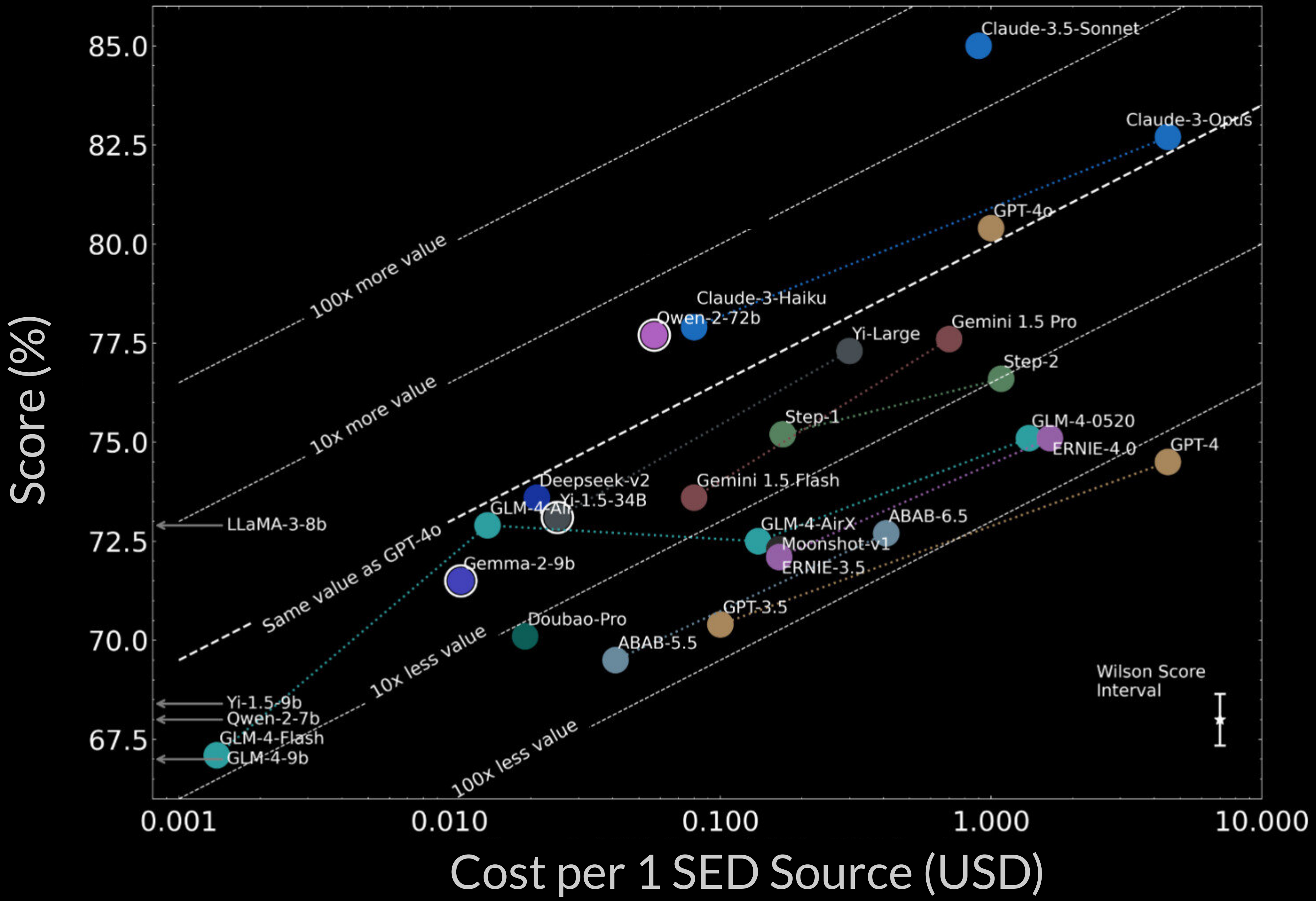
ABSTRACT

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 astronomical subfields and assesses response calibration, crucial for potential deployment in research environments. Claude-3.5-Sonnet outperforms competitors by up to 4.6 percentage points, achieving 85.0% accuracy. We observed a universal trade-off in proprietary models: a 10-fold cost increase yields a 3.5-point accuracy improvement. This suggests a 10-fold reduction in cost every 3 months to a year, offering optimistic prospects for LLM deployment in astronomy research. Open-source models have rapidly improved, with LLaMA-3-70b (80.6%) and Qwen-2-72b (77.7%) now competing with some proprietary models. We identify significant performance variations across topics, with models generally struggling more in exoplanet-related fields, recent stellar astrophysics, and observational techniques. These challenges likely stem from less abundant training data, limited historical context, and rapid recent developments in these areas. This pattern is observed across both open-weights and proprietary models, with regional dependencies evident, highlighting the impact of training data diversity on model performance in specialized scientific domains. Top-performing models demonstrate well-calibrated confidence, with correlations above 0.9 between confidence and correctness, though they tend to be slightly underconfident. The development for fast, low-cost inference of open-weights models presents new opportunities for affordable deployment in astronomy. The rapid progress observed suggests that LLM-driven research in astronomy may become feasible in the near future, potentially revolutionizing how we approach large-scale data analysis in the field.

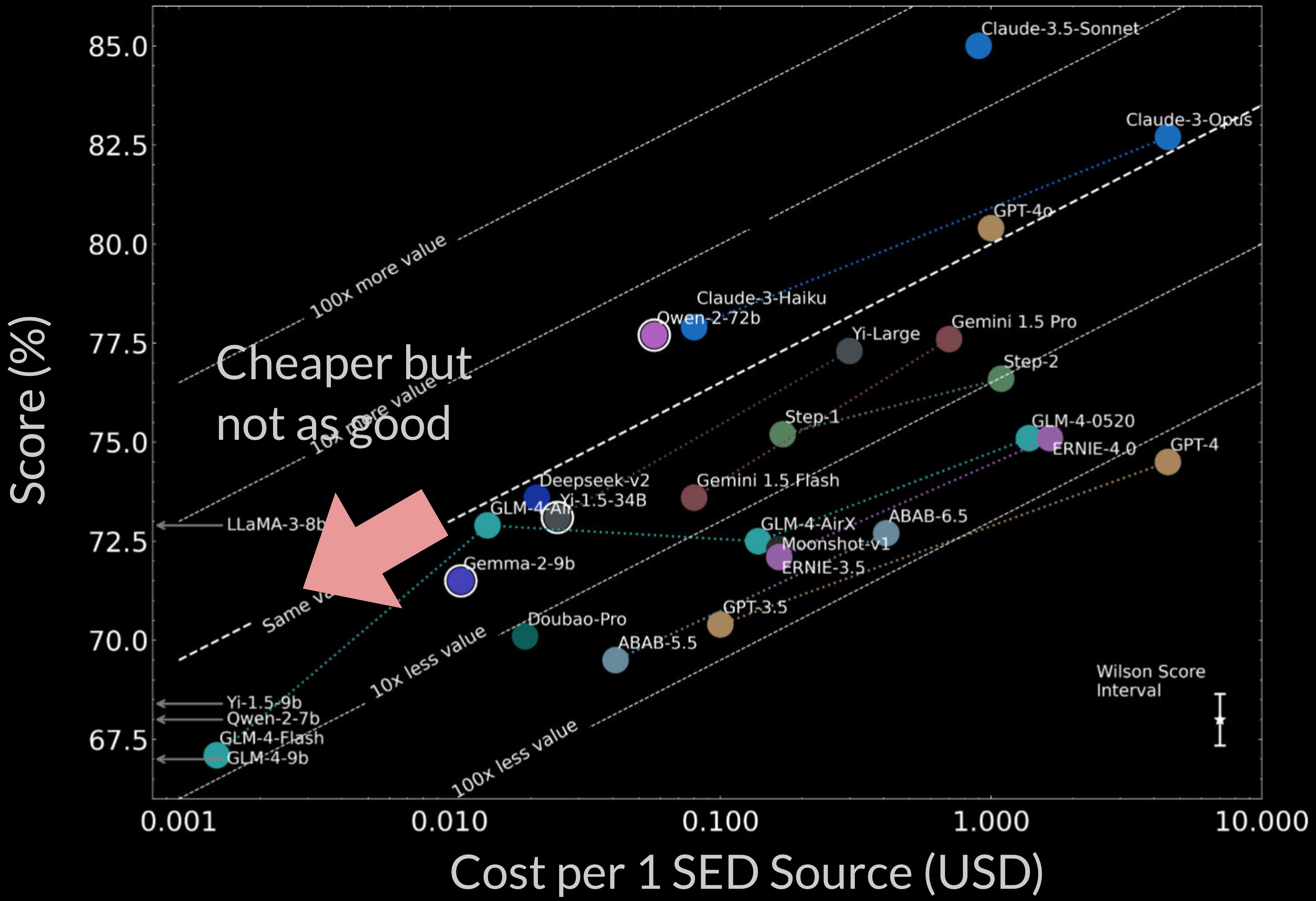

Model Series
GPT
Claude
Gemini
GLM
ERNIE
Deepseek
Step
Doubao
ABAB
Yi
Moonshot
Proprietary models
Score (%)
90
85
80
75
70
65
60
70.4
74.5
80.4
75.3
77.9
76.7
82.7
85.0
71.0
73.6
77.6
64.3
67.1
72.9
72.5
75.1
72.1
75.1
73.6
75.2
76.6
60.5
70.1
69.5
72.7
70.3
77.3
72.3
GPT-3.5
GPT-4
GPT-4o
Claude-2.0
Claude-3.0-Haiku
Claude-3.0-Sonnet
Claude-3.0-Opus
Claude-3.5-Sonnet
Gemini-1.0-Pro
Gemini-1.5-Flash
Gemini-1.5-Pro
GLM-3
GLM-4-Flash
GLM-4-Air
GLM-4-AirX
GLM-4-0520
ERNIE-3.5
ERNIE-4.0
Deepseek-v2
Step-1
Step-2
Doubao-Lite
Doubao-Pro
ABAB-5.5
ABAB-6.5
Yi-Medium
Yi-Large
Moonshot-v1

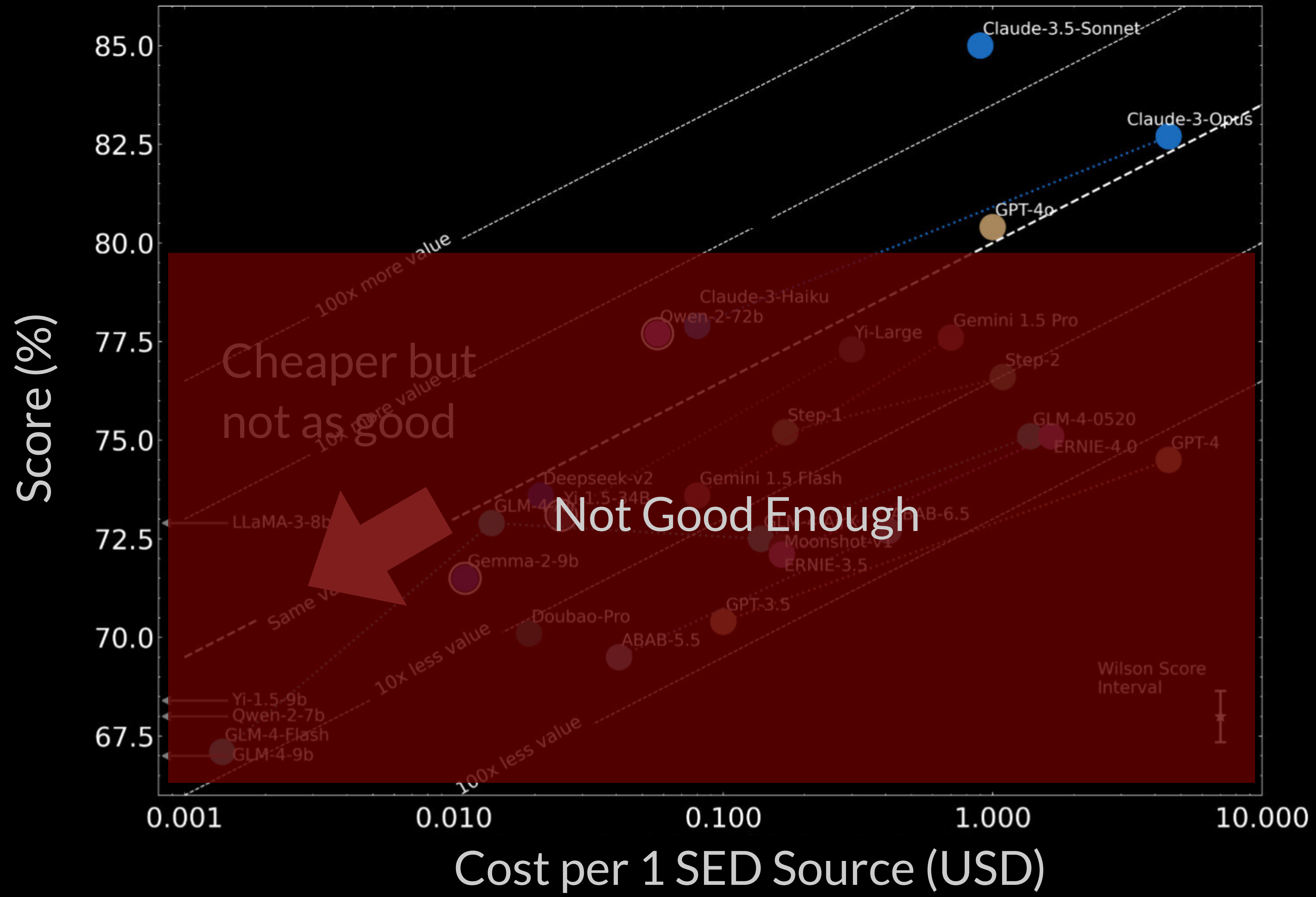


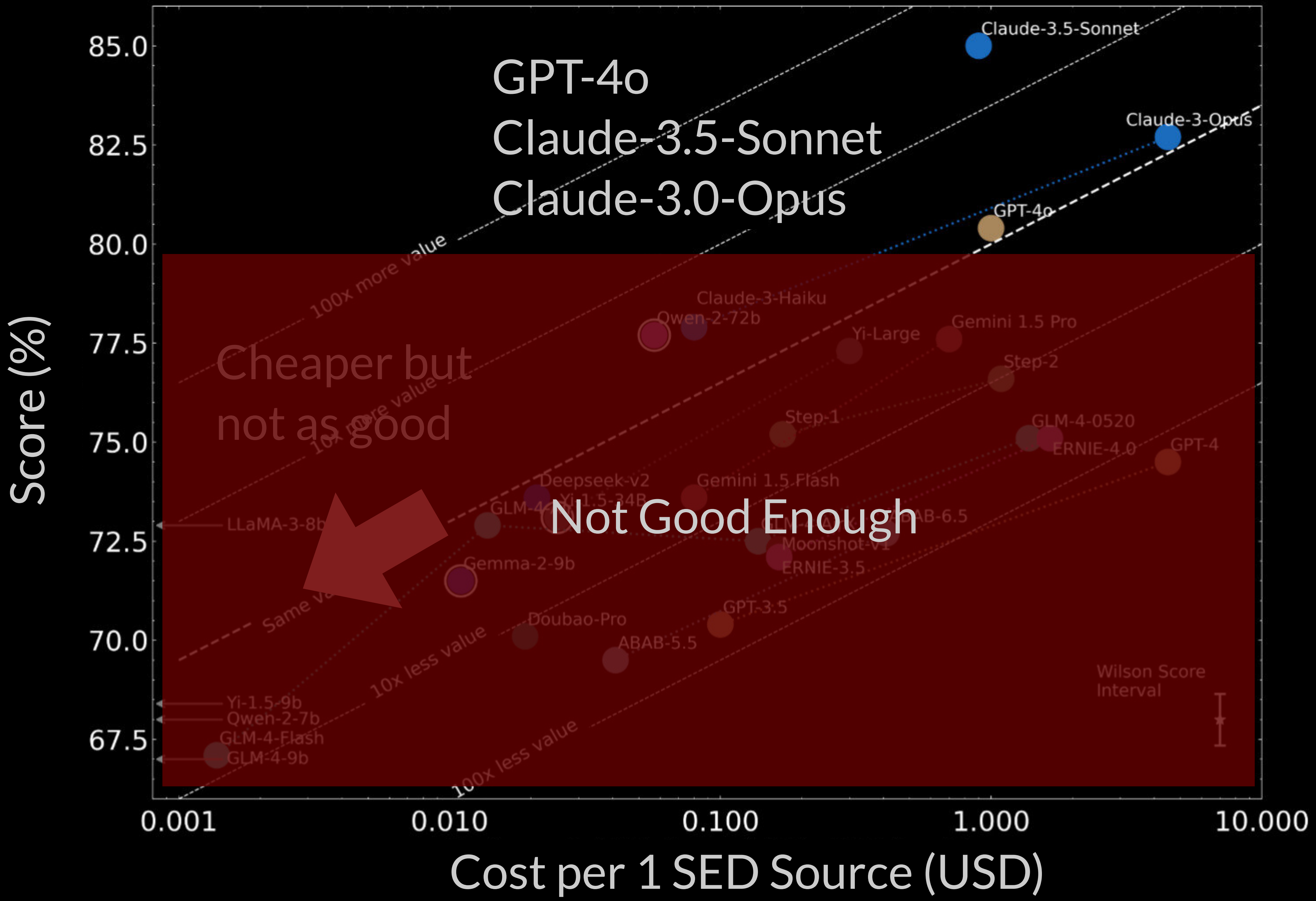
Open-weights models
Model Series
Yi
Phi
LLaMA
Qwen
Mistral
Gemma
GLM
Deepseek
Closed
Score (%)
Claude-3.5-Sonnet
GPT-4o
Gemini-1.5-Pro
Yi1.5-9b 68.4
Yi1.5-34b 73.1
Phi-2-3b 65.6
Phi-3-4b 71.7
Phi-3-14b 75.6
LLaMA-2-7b 50.3
LLaMA-2-70b 70.7
LLaMA-3-8b 72.9
LLaMA-3-70b 80.6
Qwen-1-7b 57.4
Qwen-1.5-7b 63.7
Qwen-1.5-14b 67.7
Qwen-1.5-32b 73.2
Qwen-1.5-110b 72.7
Qwen-2-7b 68.0
Qwen-2-57b 71.8
Qwen-2-72b 77.7
Mistral-7b-v0.1 48.1
Mistral-7b-v0.3 63.9
Mistral-8x7b-v0.1 73.7
Mixtral-8x22b-v0.1 77.7
Gemma-1-2b 44.1
Gemma-1-7b 56.1
Gemma-2-9b 71.5
ChatGLM3-6b 50.4
GLM-4-9b 67.0
Deepseek-67b 63.1



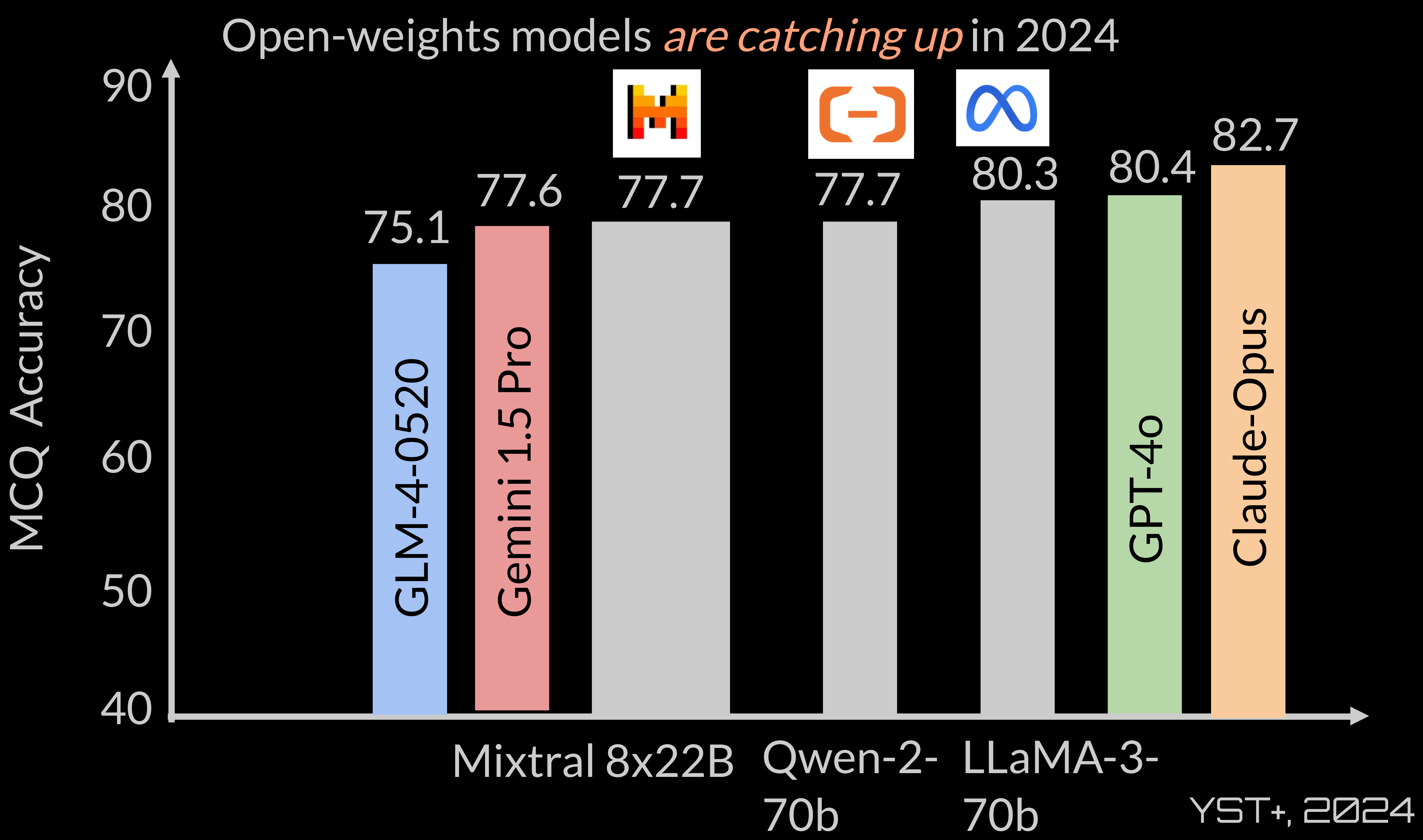
Score (%)
85.0
82.5
80.0
77.5
75.0
72.5
70.0
67.5
Cost per 1 SED Source (USD)
0.001
0.010
0.100
1.000
10.000
100x more value
10x more value
Same value as GPT-4o
10x less value
100x less value
Claude-3.5-Sonnet
Claude-3-Opus
GPT-4o
Claude-3-Haiku
Qwen-2-72b
Yi-Large
Gemini 1.5 Pro
Step-2
Step-1
GLM-4-0520
ERNIE-4.0
GPT-4
Deepseek-v2
Yi-1.5-34B
GLM-4-Air
Gemini 1.5 Flash
LLaMA-3-8b
GLM-4-AirX
ABAB-6.5
Moonshot-v1
ERNIE-3.5
Gemma-2-9b
GPT-3.5
Doubao-Pro
ABAB-5.5
Yi-1.5-9b
Qwen-2-7b
GLM-4-Flash
GLM-4-9b
Wilson Score
Interval


Cheaper but
not as good
Score (%)
Cost per 1 SED Source (USD)
85.0
82.5
80.0
77.5
75.0
72.5
70.0
67.5
0.001
0.010
0.100
1.000
10.000
100x more value
10x more value
Same value
10x less value
100x less value
Claude-3.5-Sonnet
Claude-3-Opus
GPT-4o
Claude-3-Haiku
Qwen-2-72b
Yi-Large
Gemini 1.5 Pro
Step-2
Step-1
GLM-4-0520
ERNIE-4.0
GPT-4
Deepseek-v2
Yi-1.5-34B
GLM-4-Air
LLaMA-3-8b
Gemini 1.5 Flash
GLM-4-AirX
ABAB-6.5
Moonshot-v1
ERNIE-3.5
Gemma-2-9b
GPT-3.5
Doubao-Pro
ABAB-5.5
Yi-1.5-9b
Qwen-2-7b
GLM-4-Flash
GLM-4-9b
Wilson Score
Interval


Cheaper but
not as good
Not Good Enough
Score (%)
85.0
82.5
80.0
77.5
75.0
72.5
70.0
67.5
Cost per 1 SED Source (USD)
0.001
0.010
0.100
1.000
10.000
Claude-3.5-Sonnet
Claude-3-Opus
GPT-4o
100x more value
Claude-3-Haiku
Qwen-2-72b
Yi-Large
Gemini 1.5 Pro
Step-2
Step-1
GLM-4-0520
ERNIE-4.0
GPT-4
Deepseek-v2
Gemini 1.5 Flash
LLaMA-3-8b
Moonshot-v1
ERNIE-3.5
Gemma-2-9b
Doubao-Pro
GPT-3.5
ABAB-5.5
10x less value
100x less value
Yi-1.5-9b
Qwen-2-7b
GLM-4-Flash
GLM-4-9b
Wilson Score
Interval


GPT-4o
Claude-3.5-Sonnet
Claude-3.0-Opus
Cheaper but
not as good
Not Good Enough
Score (%)
Cost per 1 SED Source (USD)
85.0
82.5
80.0
77.5
75.0
72.5
70.0
67.5
0.001
0.010
0.100
1.000
10.000
Claude-3.5-Sonnet
Claude-3-Opus
GPT-4o
100x more value
10x more value
Same value
10x less value
100x less value
Claude-3-Haiku
Qwen-2-72b
Yi-Large
Gemini 1.5 Pro
Step-2
Step-1
GLM-4-0520
ERNIE-4.0
GPT-4
Deepseek-v2
Gemini 1.5 Flash
LLaMA-3-8b
Moonshot-v1
ERNIE-3.5
Gemma-2-9b
GPT-3.5
Doubao-Pro
ABAB-5.5
Yi-1.5-9b
Qwen-2-7b
GLM-4-Flash
GLM-4-9b
Wilson Score
Interval

Open-weights large language models?


Open-weights models are catching up in 2024
MCQ Accuracy
90
80
70
60
50
40
75.1
GLM-4-0520
77.6
Gemini 1.5 Pro
77.7
Mixtral 8x22B
77.7
Qwen-2-70b
80.3
LLaMA-3-70b
80.4
GPT-4o
82.7
Claude-Opus
YST+, 2024

Open-source large language models
are *as good as* the closed-source models

Open-source large language models
are *as good as* the closed-source models

at the ~70B level

Still it is *not very scalable*

Still it is *not very scalable*

LLaMA-3 70b throughput on *four A100 GPUs*

= ~ 100 tokens / second

LLaMA-3 70b throughput on *four A100 GPUs*

= ~ 100 tokens / second

1 SED source = 15 GPU minutes

LLaMA-3 70b throughput on *four A100 GPUs*

= ~ 100 tokens / second

1 SED source = 15 GPU minutes

1B sources = 10M GPU days

LLaMA-3 70b throughput on *four A100 GPUs*

= ~ 100 tokens / second

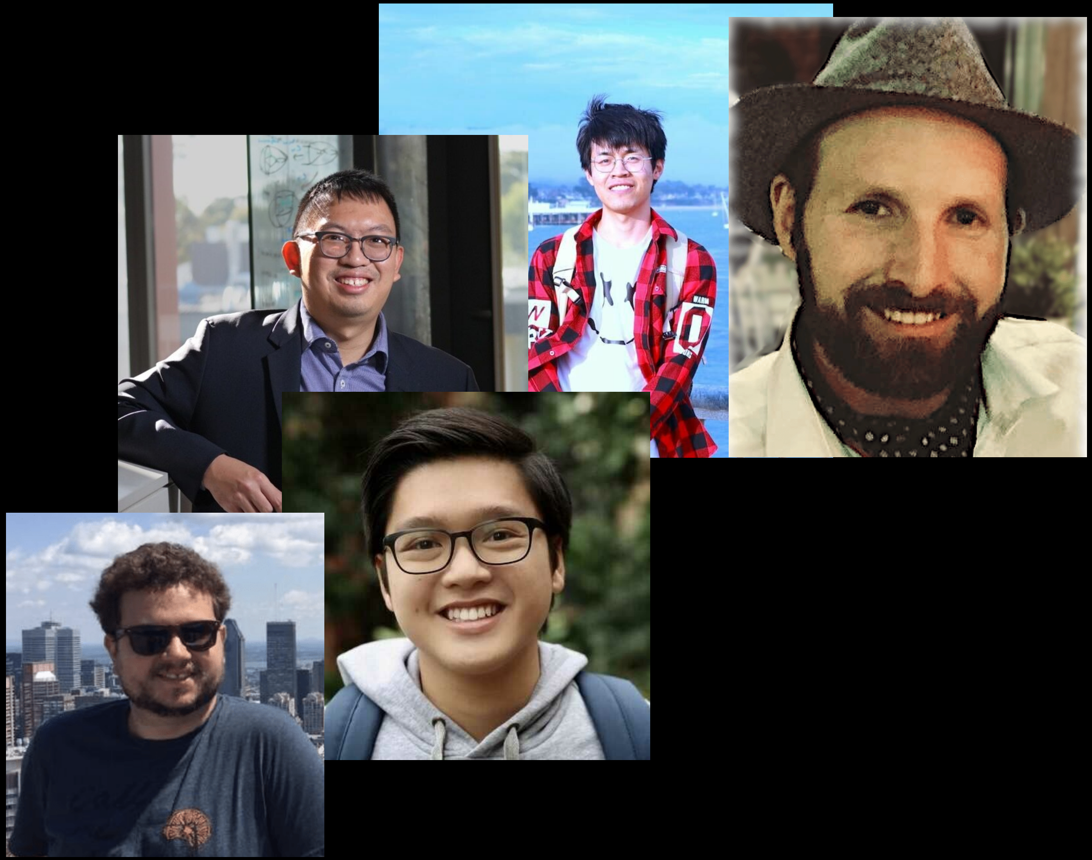
1 SED source = 15 GPU minutes

1B sources = 10M GPU days

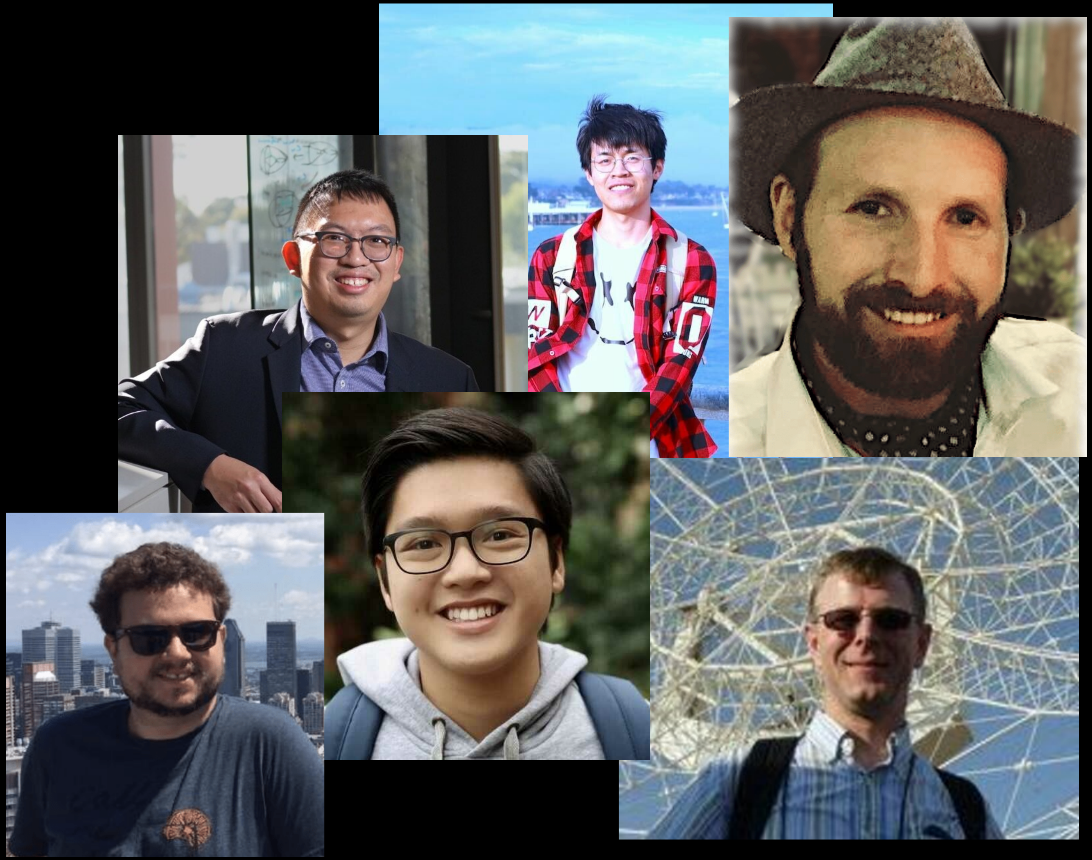
A cluster with *1000 A100 GPUs*
running for *30 years*

Can the light-weight *~7B* open-source models *perform equally* with some "fine-tuning" ?

Previously UniverseTBD -> now *AstroMLab* (astromlab.org)



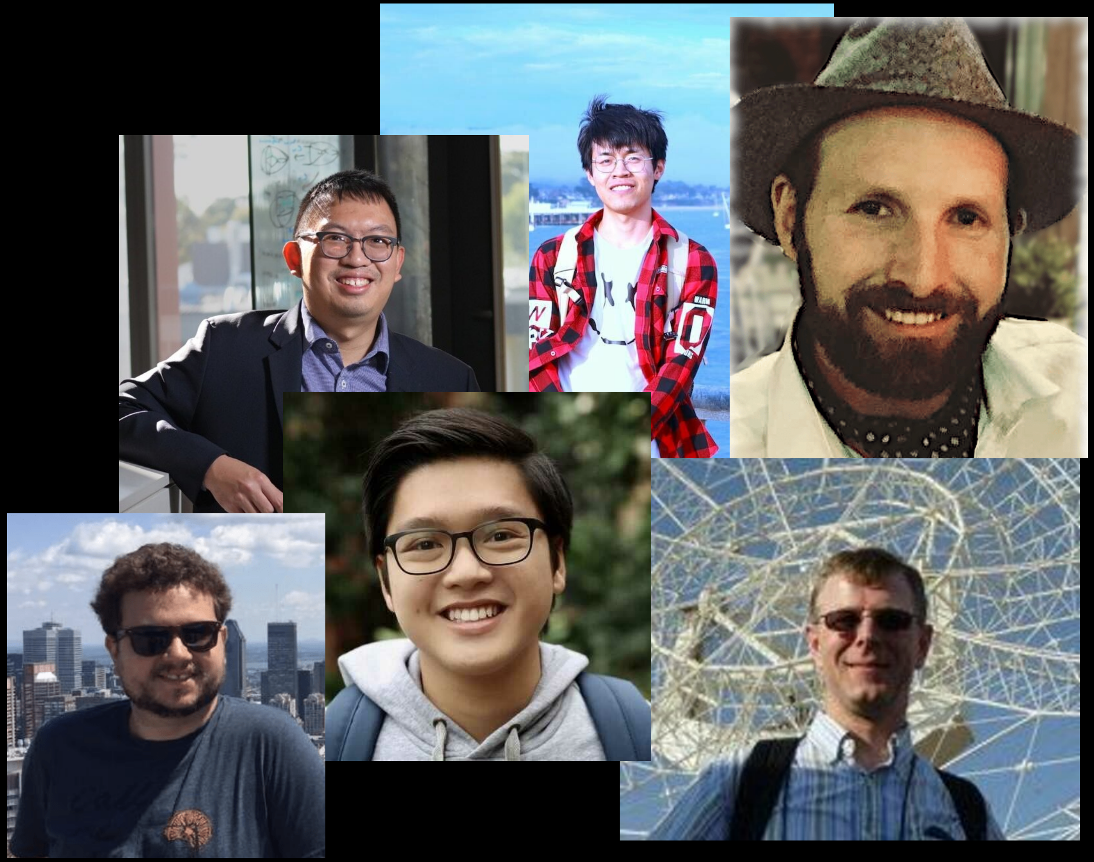
Previously UniverseTBD -> now *AstroMLab* (astromlab.org)



Harvard-Smithsonian ADS

Previously UniverseTBD -> now *AstroMLab* (astromlab.org)

Natural Language Processing experts

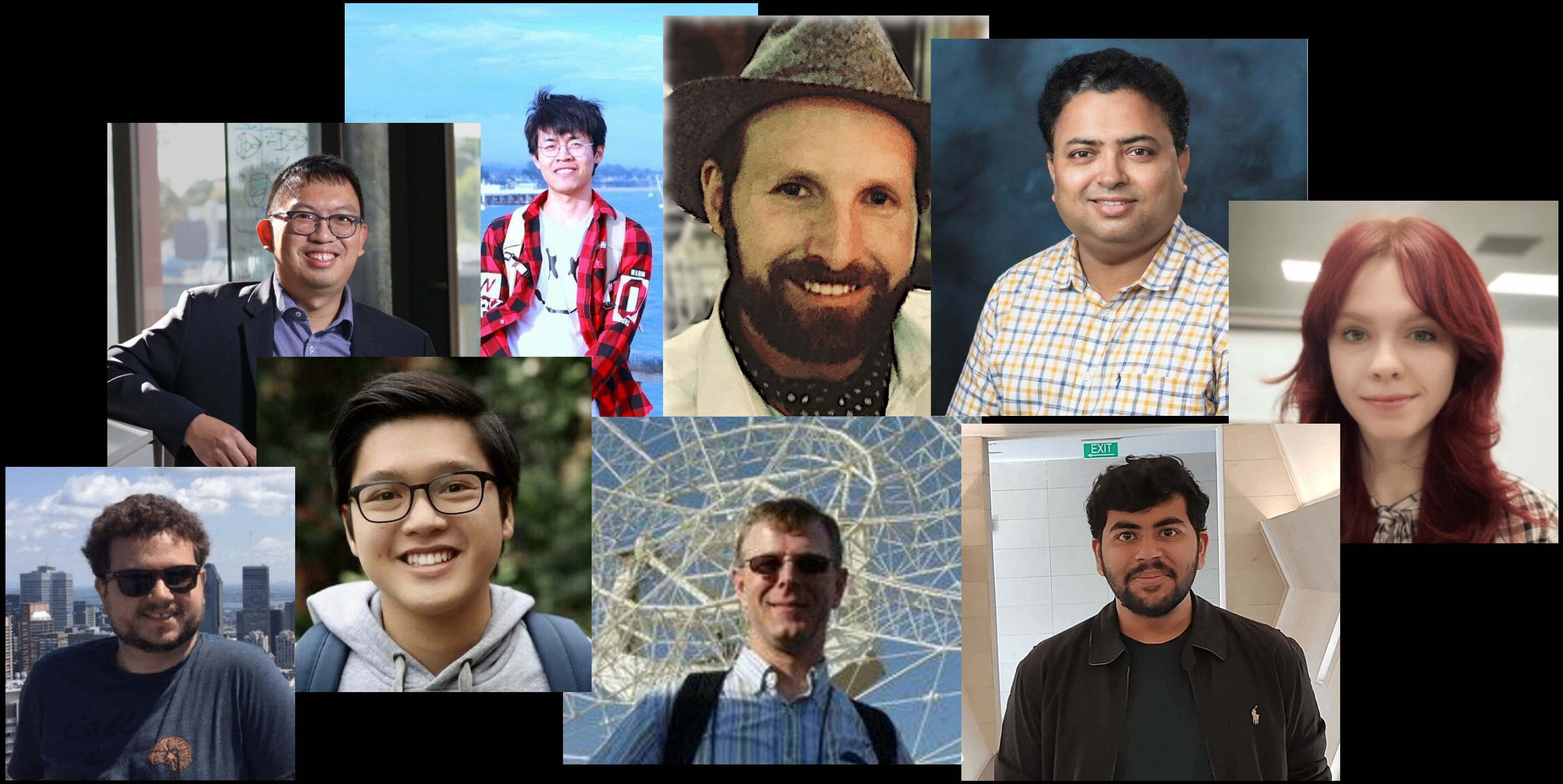


Harvard-Smithsonian ADS

Previously UniverseTBD -> now *AstroMLab* (astromlab.org)

Natural Language Processing experts

Oak Ridge National Lab



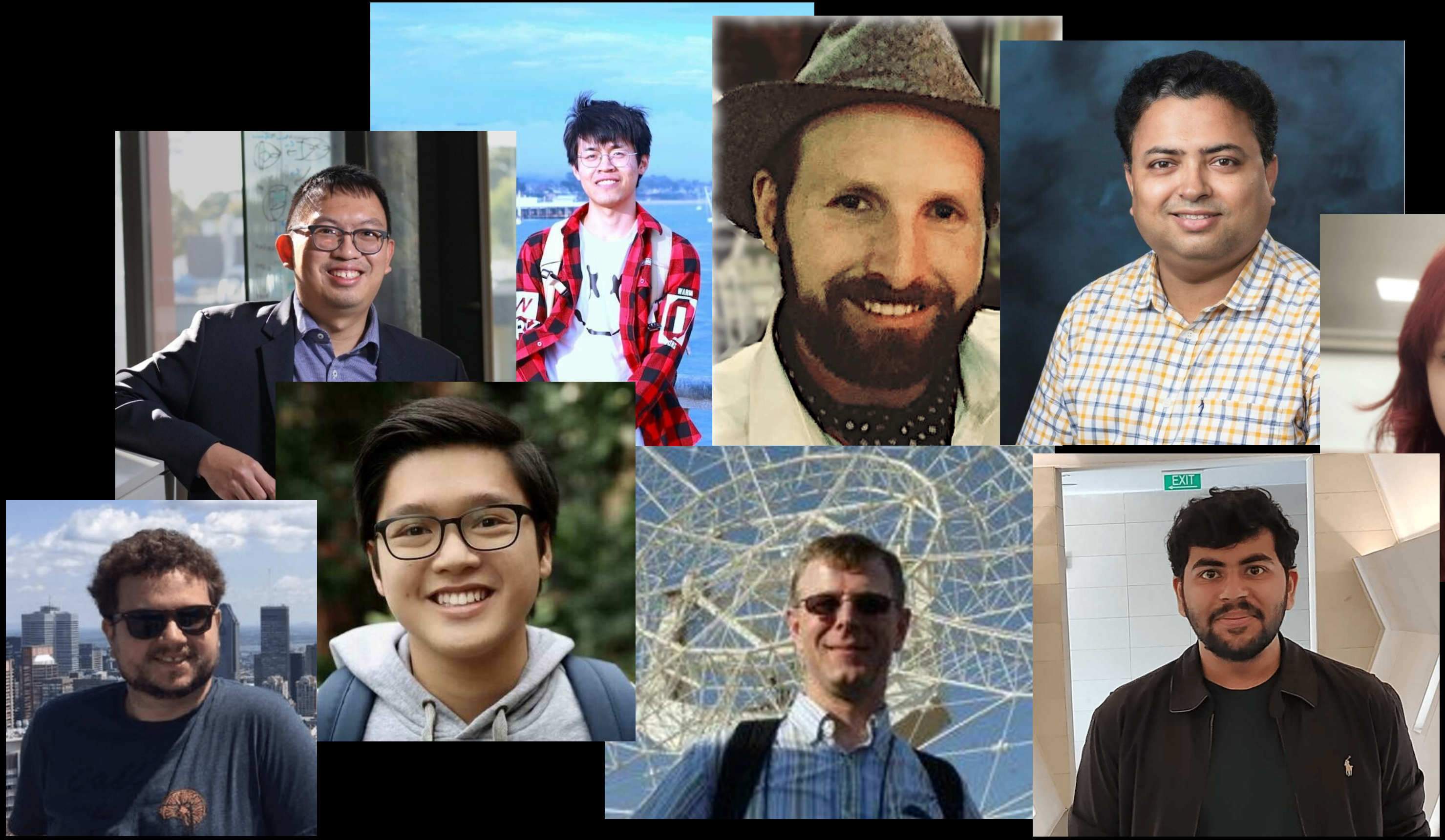
Harvard-Smithsonian ADS

Previously UniverseTBD -> now *AstroMLab* (astromlab.org)

Natural Language Processing experts

Oak Ridge National Lab

Argonne National Lab



Harvard-Smithsonian ADS

Previously UniverseTBD -> now *AstroMLab* (astromlab.org)

Natural Language Processing experts

Oak Ridge National Lab

Argonne National Lab

U. Ilinois Urbana-Champaign

Harvard-Smithsonian ADS

Continual pre-training on all the *330K astronomy papers*

AstroLLaMA: Towards Specialized Foundation Models in Astronomy

Published on Sep 12 · ⭐ Featured in Daily Papers on Sep 13

Authors: Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciucă, Charlie O'Neill, Ze-Chang Sun, Maja Jabłońska, Sandor Kruk, Ernest Perkowski, Jack Miller, Jason Li, Josh Peek, Kartheik Iyer, Tomasz Różański, Pranav Khetarpal, Sharaf Zaman, David Brodrick, Sergio J. Rodríguez Méndez, Thang Bui, Alyssa Goodman, Alberto Accomazzi, Jill Naiman, Jesse Cranney +2 authors

Continual pre-training on all the *330K astronomy papers*

OPEN ACCESS

AstroLLaMA-Chat: Scaling AstroLLaMA with Conversational and Diverse Datasets

Ernest Perkowski[15,1], Rui Pan[15,2], Tuan Dung Nguyen[3], Yuan-Sen Ting[4,5,6,7], Sandor Kruk[1], Tong Zhang[8], Charlie O'Neill[9], Maja Jablonska[4], Zechang Sun[10], Michael J. Smith[11]

+ Show full

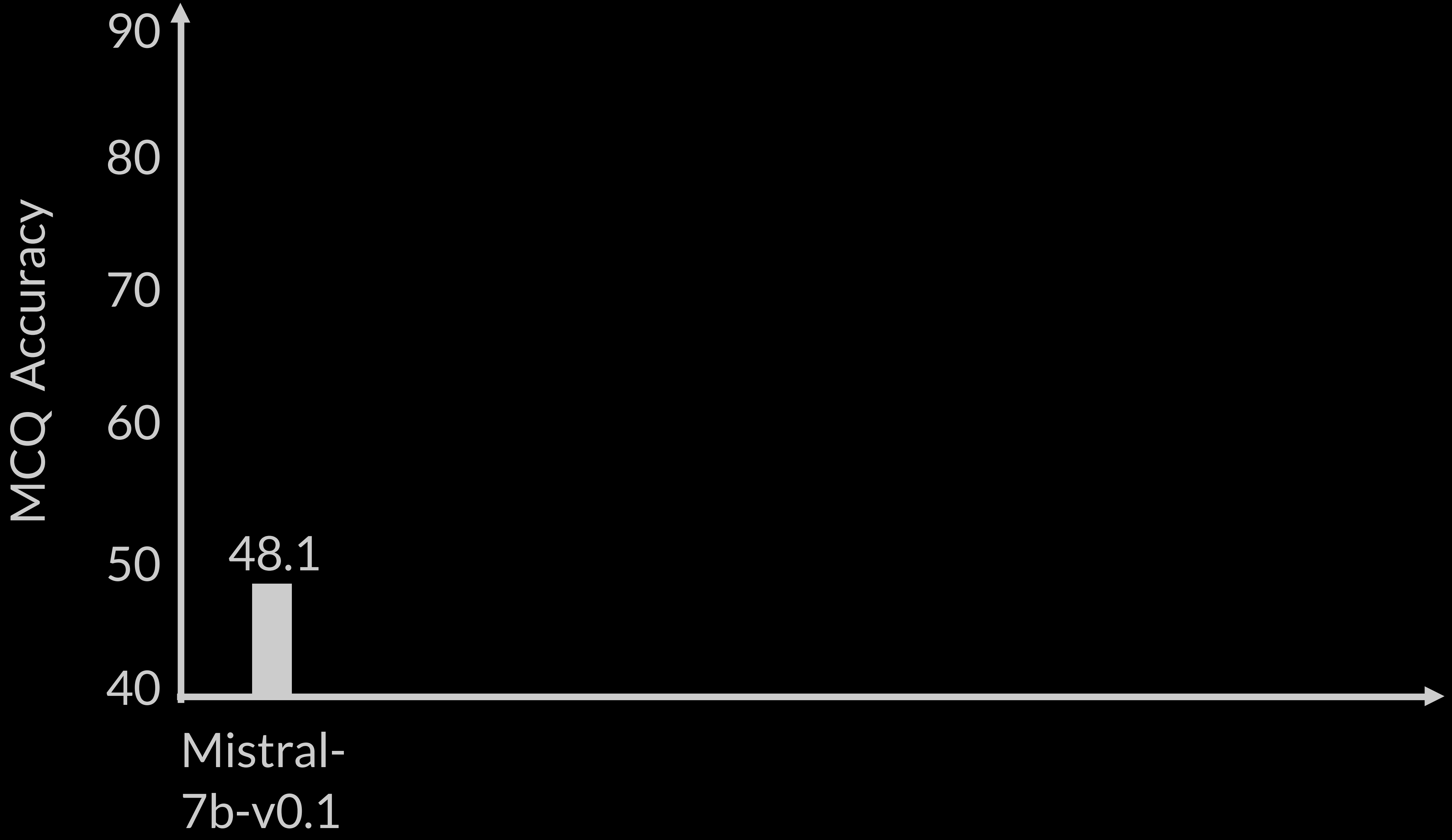


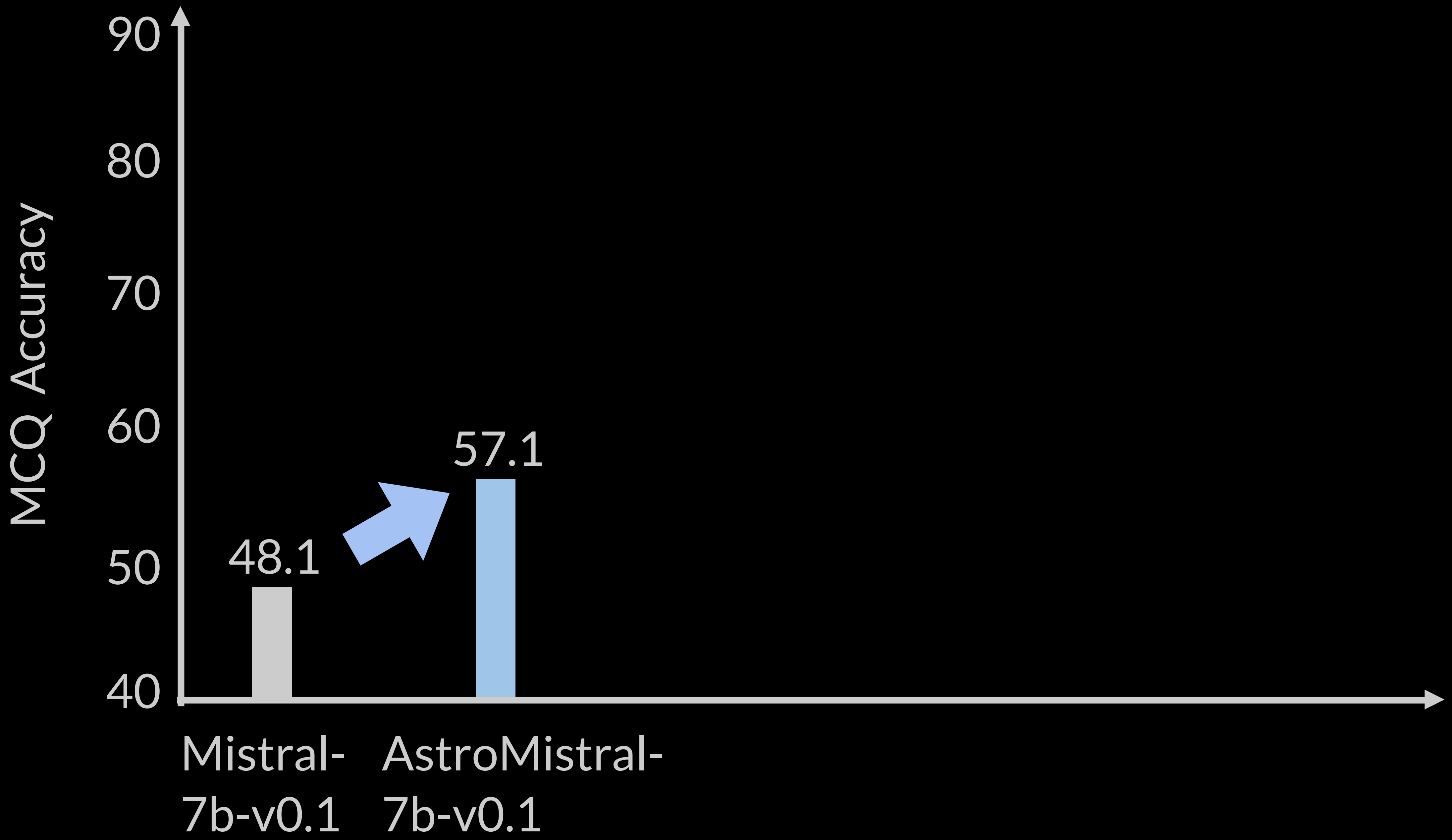
Published Ja… A… e… So

Research No… Nu

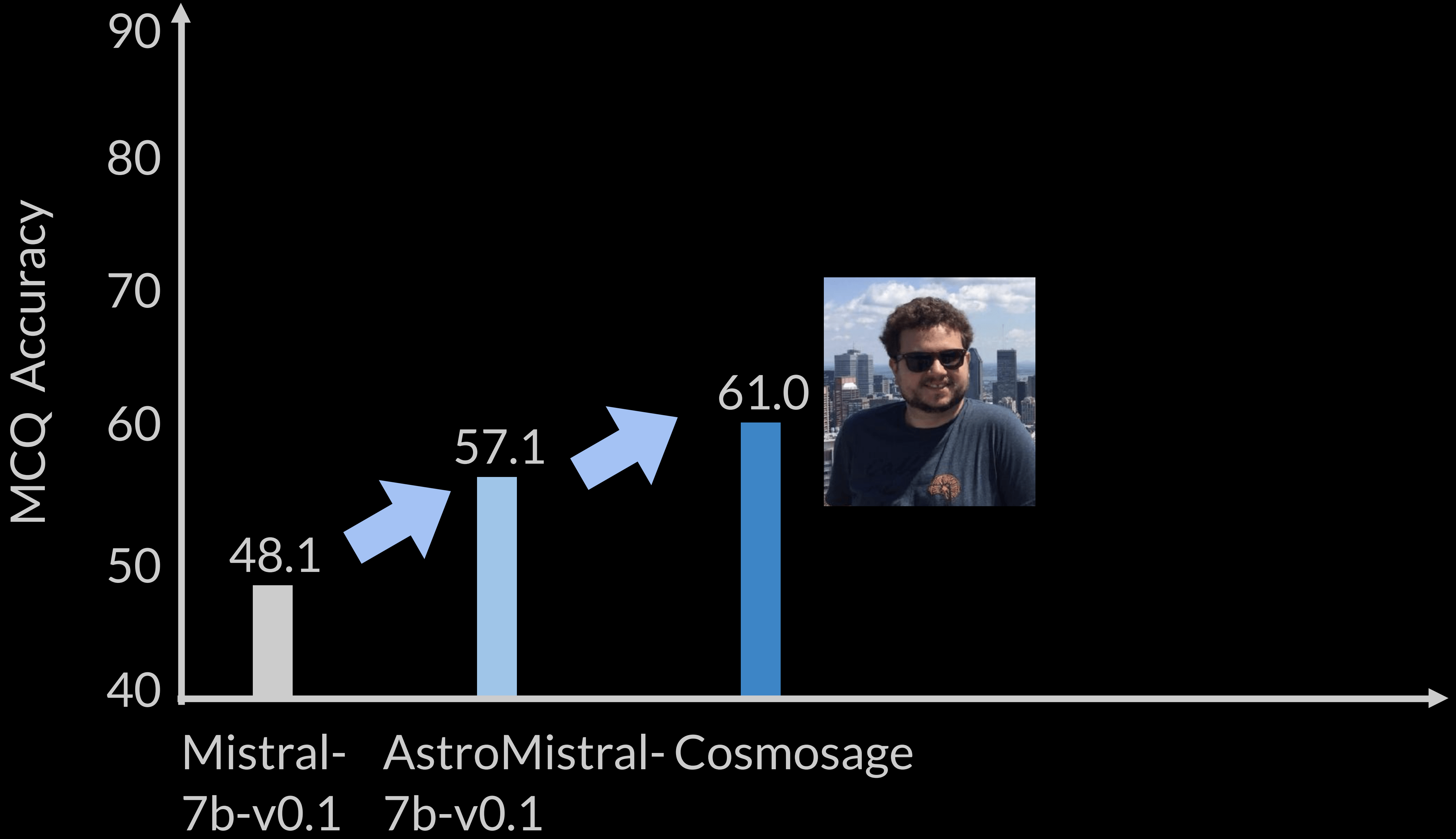
Citation Erne… es.

DOI 10.3847

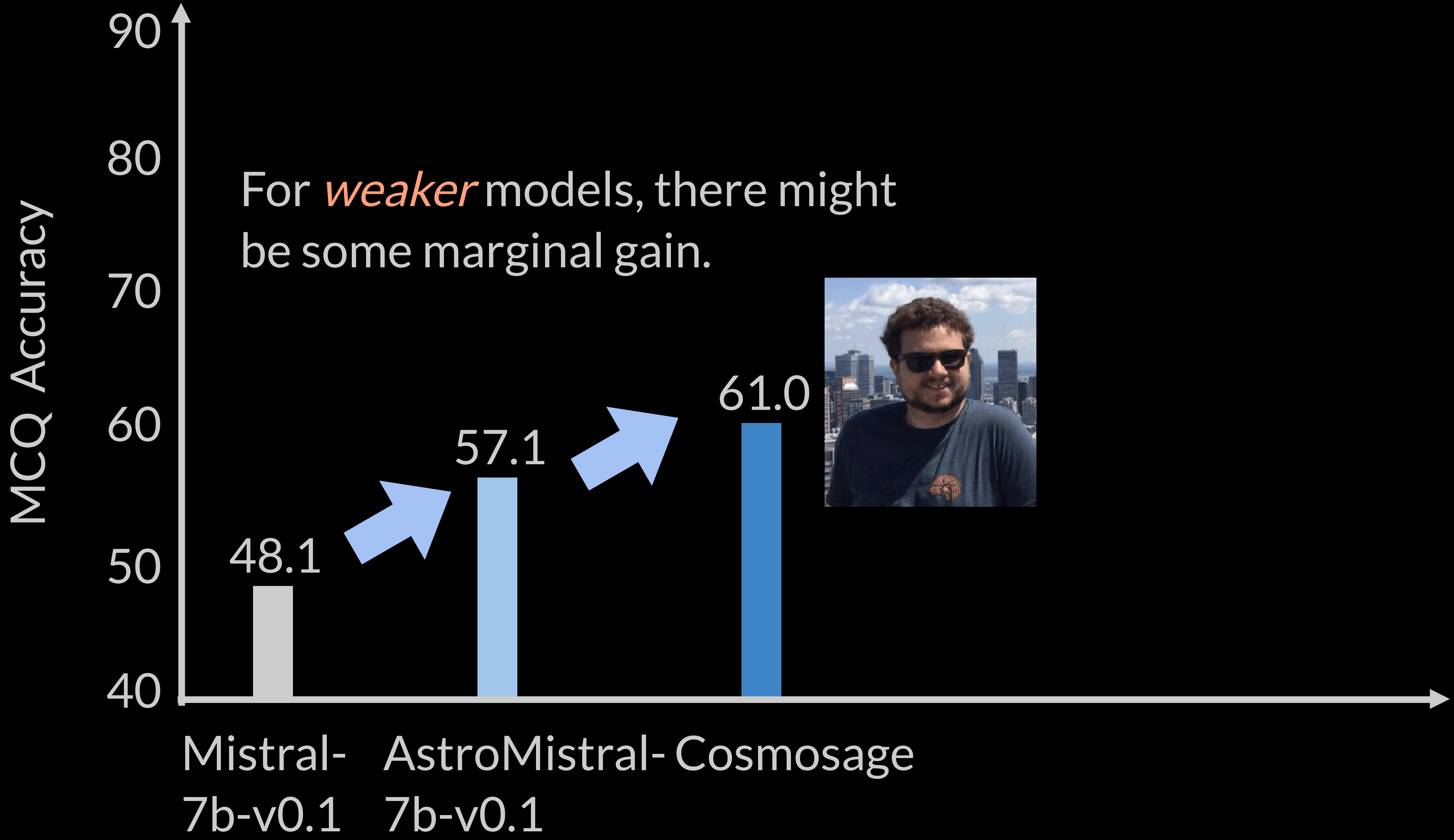

MCQ Accuracy
90
80
70
60
50
40
48.1
Mistral-
7b-v0.1



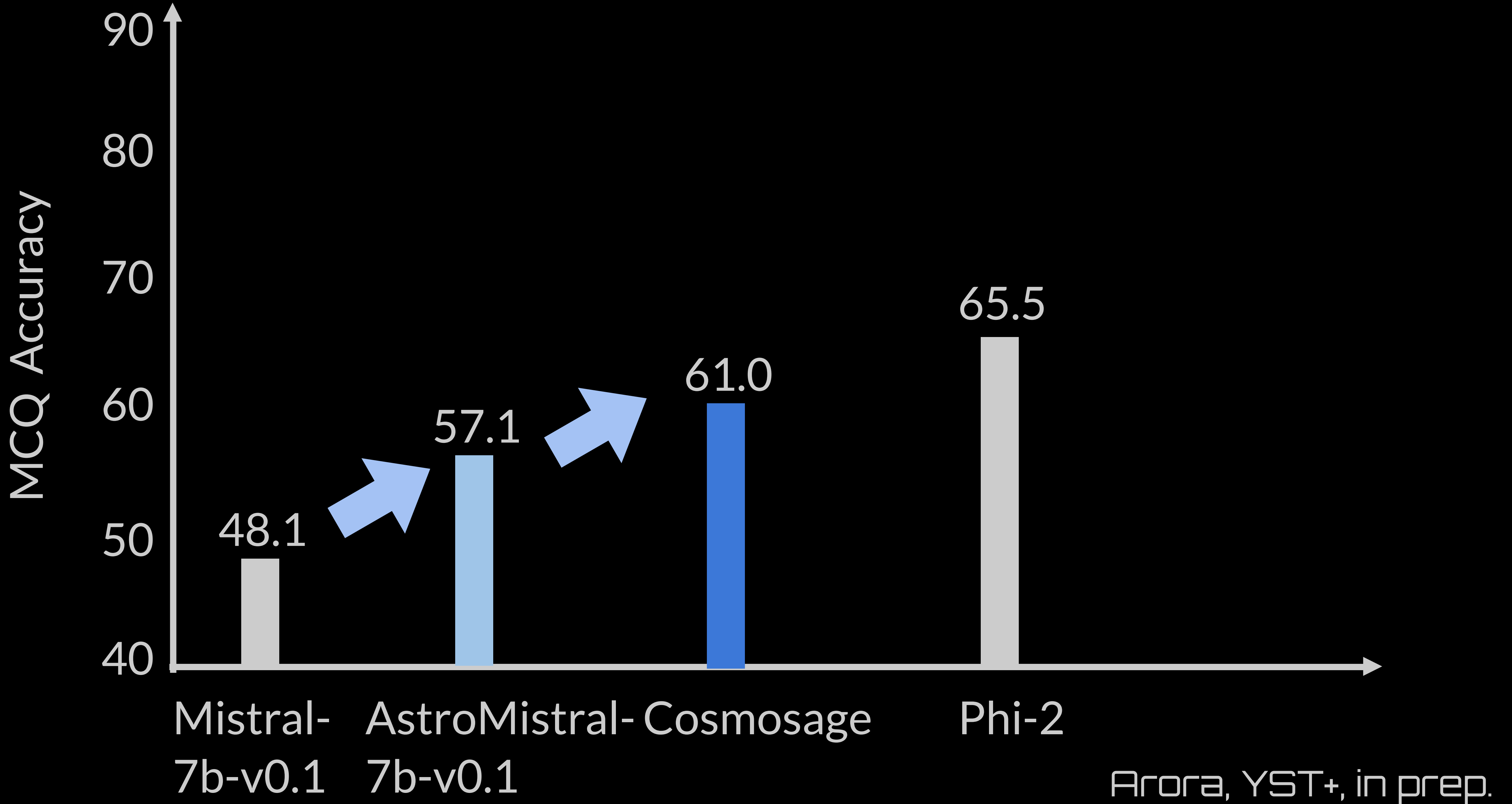
MCQ Accuracy
90
80
70
60
50
40
48.1
57.1
Mistral-7b-v0.1
AstroMistral-7b-v0.1



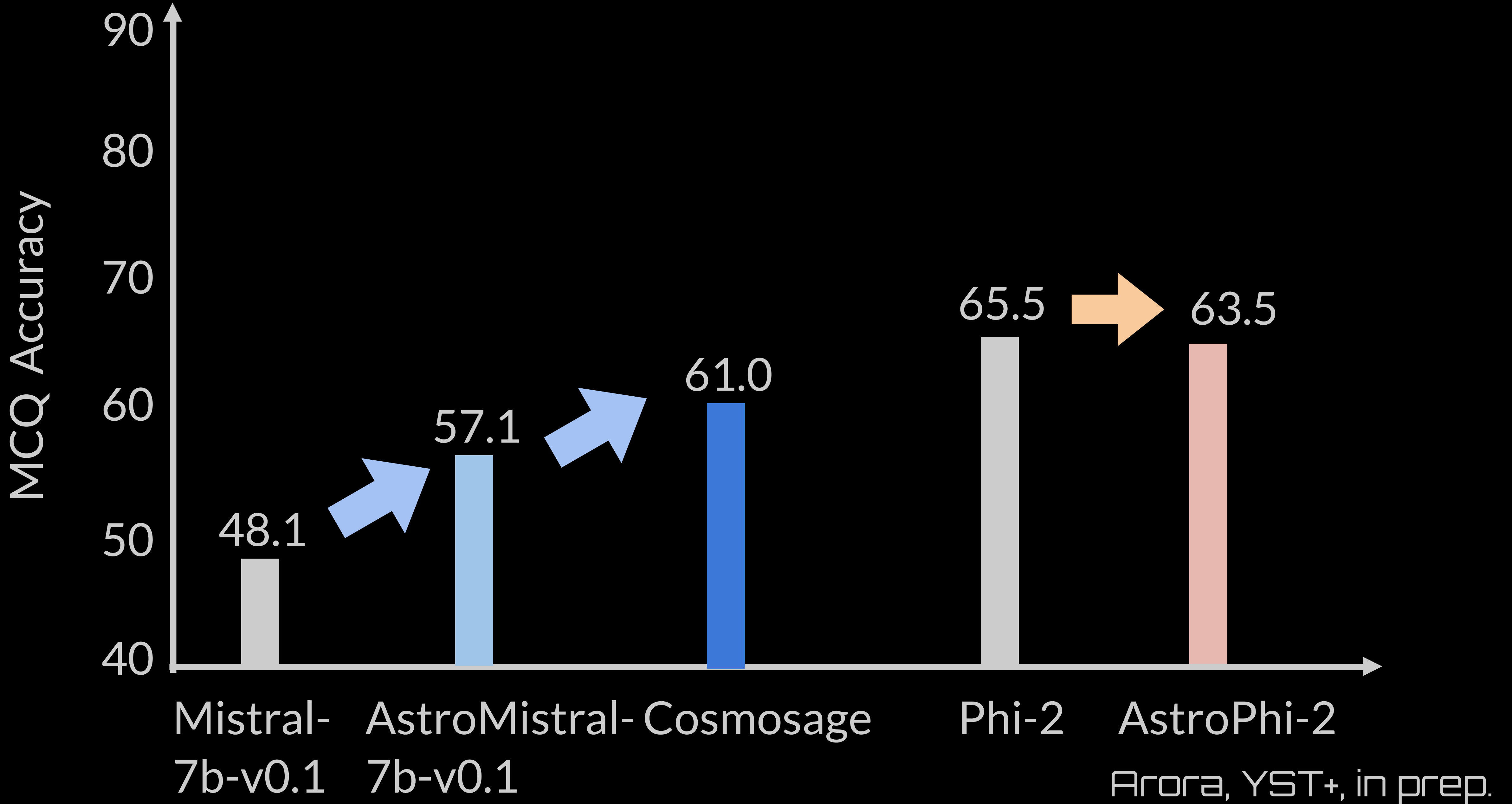
MCQ Accuracy
90
80
70
60
50
40
48.1
57.1
61.0
Mistral-7b-v0.1
AstroMistral-7b-v0.1
Cosmosage



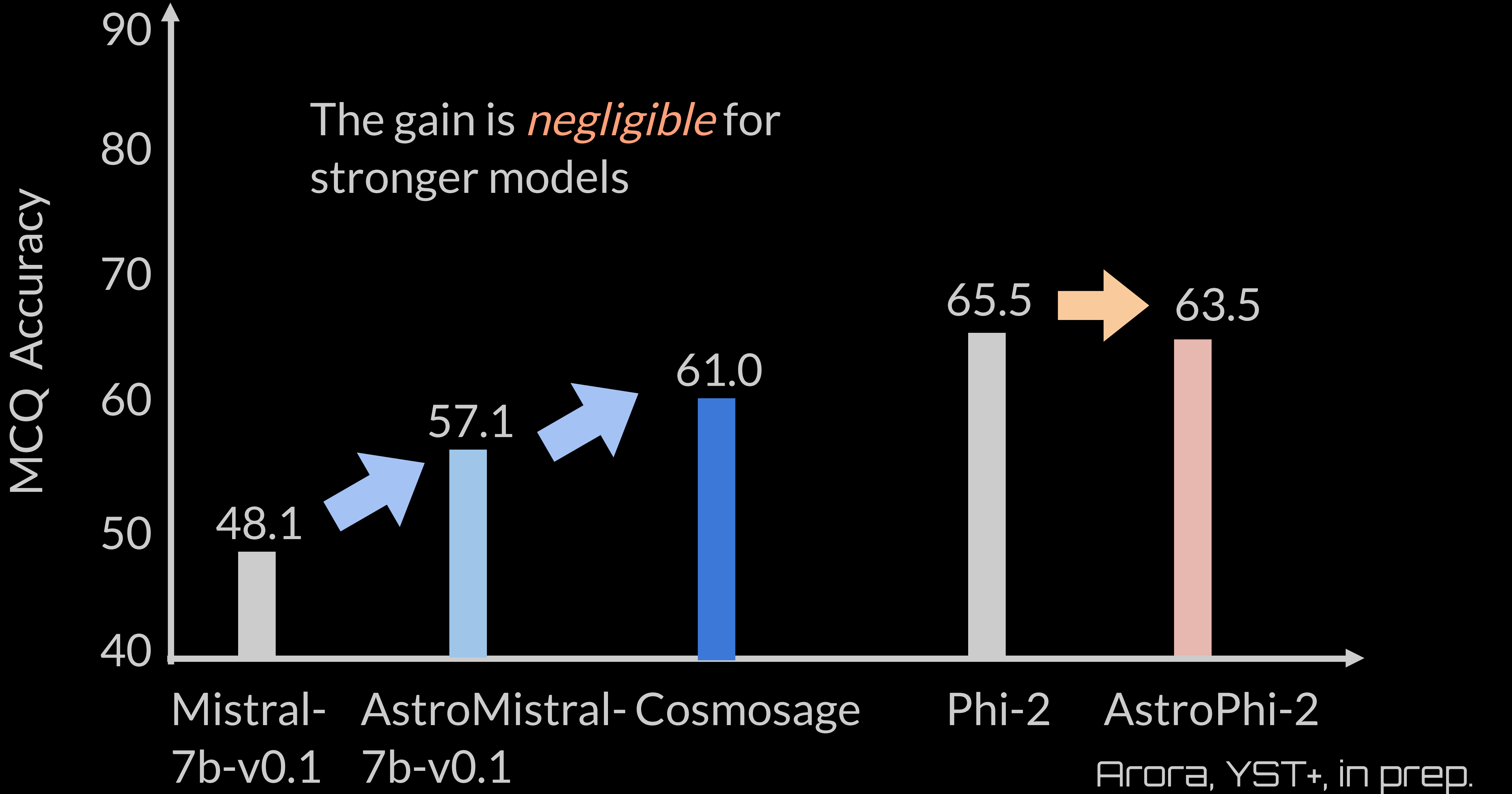
MCQ Accuracy
90
80
70
60
50
40
For *weaker* models, there might be some marginal gain.
48.1
57.1
61.0
Mistral-7b-v0.1
AstroMistral-7b-v0.1
Cosmosage



90
80
70
60
50
40
MCQ Accuracy
48.1
57.1
61.0
65.5
Mistral-7b-v0.1
AstroMistral-7b-v0.1
Cosmosage
Phi-2
Arora, YST+, in prep.



MCQ Accuracy
90
80
70
60
50
40
48.1
57.1
61.0
65.5
63.5
Mistral-7b-v0.1
AstroMistral-7b-v0.1
Cosmosage
Phi-2
AstroPhi-2
Arora, YST+, in prep.



The gain is *negligible* for stronger models
MCQ Accuracy
90
80
70
60
50
40
48.1
57.1
61.0
65.5
63.5
Mistral-7b-v0.1
AstroMistral-7b-v0.1
Cosmosage
Phi-2
AstroPhi-2
Arora, YST+, in prep.

Current LLMs remain *brittle*, and efforts to improve them through continual pretraining *have yet* to yield significant gains.

So what should we do ?

Warren Buffet :

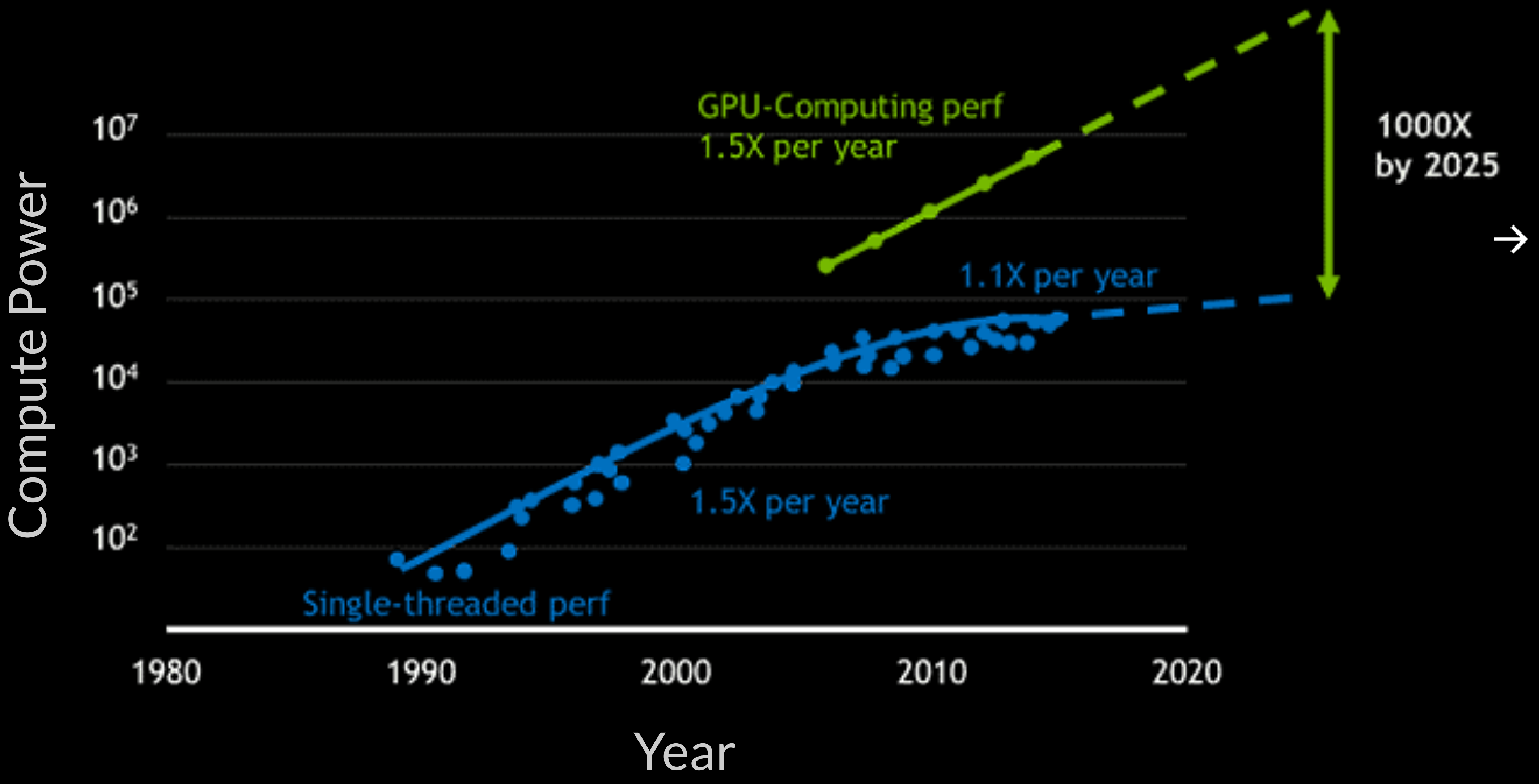
"The trick is, when there is nothing to do, *do nothing.*"

Warren Buffet :

"The trick is, when there is nothing to do, *do nothing.*"

"The stock market is designed to transfer money from the Active to the *Patient.*"

Huang's Law


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

Tech / Tech Trends

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

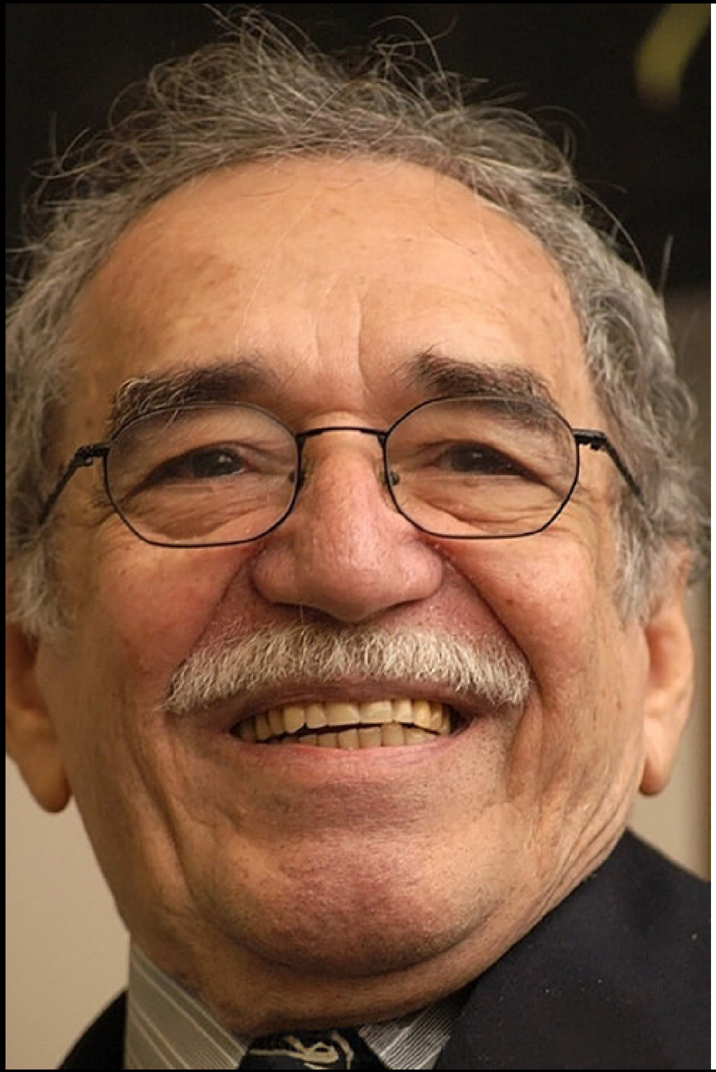
- Zhipu AI's GLM series of large language models now costs 90 per cent less than the current industry average of 1 yuan per 1 million tokens

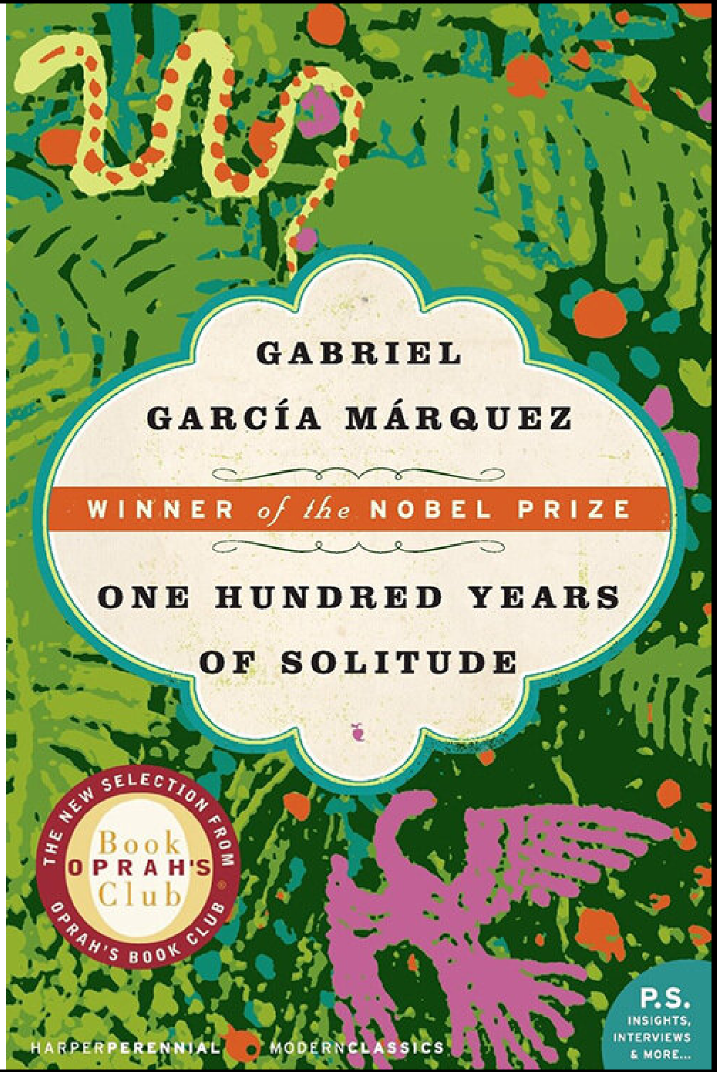
0.1 Yuan
per 1 million tokens

Ben Jiang in Beijing + FOLLOW

Published: 8:00pm, 5 Jun 2024

Why you can trust SCMP




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

= 0.03 Euro

Cornell University

arXiv > astro-ph

Astrophysics (since April 1992)

= 40 Euro

AirAsia
NOW
EVERYONE
CAN ~~FLY~~
LLM
Buy Now
Fly Now!

Summary :

Summary :

- Despite the surge in interest, the *knowledge graph* reveals a striking *lack of innovation* at the intersection of AI and astronomy.

Summary :

- Despite the surge in interest, the *knowledge graph* reveals a striking *lack of innovation* at the intersection of AI and astronomy.
- Current research prioritizes tool development over *creating meta-methods* that could accelerate the discovery of 'unknown unknowns.'

Summary :

- Despite the surge in interest, the *knowledge graph* reveals a striking *lack of innovation* at the intersection of AI and astronomy.
- Current research prioritizes tool development over *creating meta-methods* that could accelerate the discovery of 'unknown unknowns.'
- By interacting with real-world data, *LLM agents* can distill causal relationships, conducting end-to-end research autonomously.

Summary :

- Despite the surge in interest, the *knowledge graph* reveals a striking *lack of innovation* at the intersection of AI and astronomy.
- Current research prioritizes tool development over *creating meta-methods* that could accelerate the discovery of 'unknown unknowns.'
- By interacting with real-world data, *LLM agents* can distill causal relationships, conducting end-to-end research autonomously.
- The astronomical cost of *applying at scale* limits our progress, but the growth in GPU computing power will inevitably overcome this.