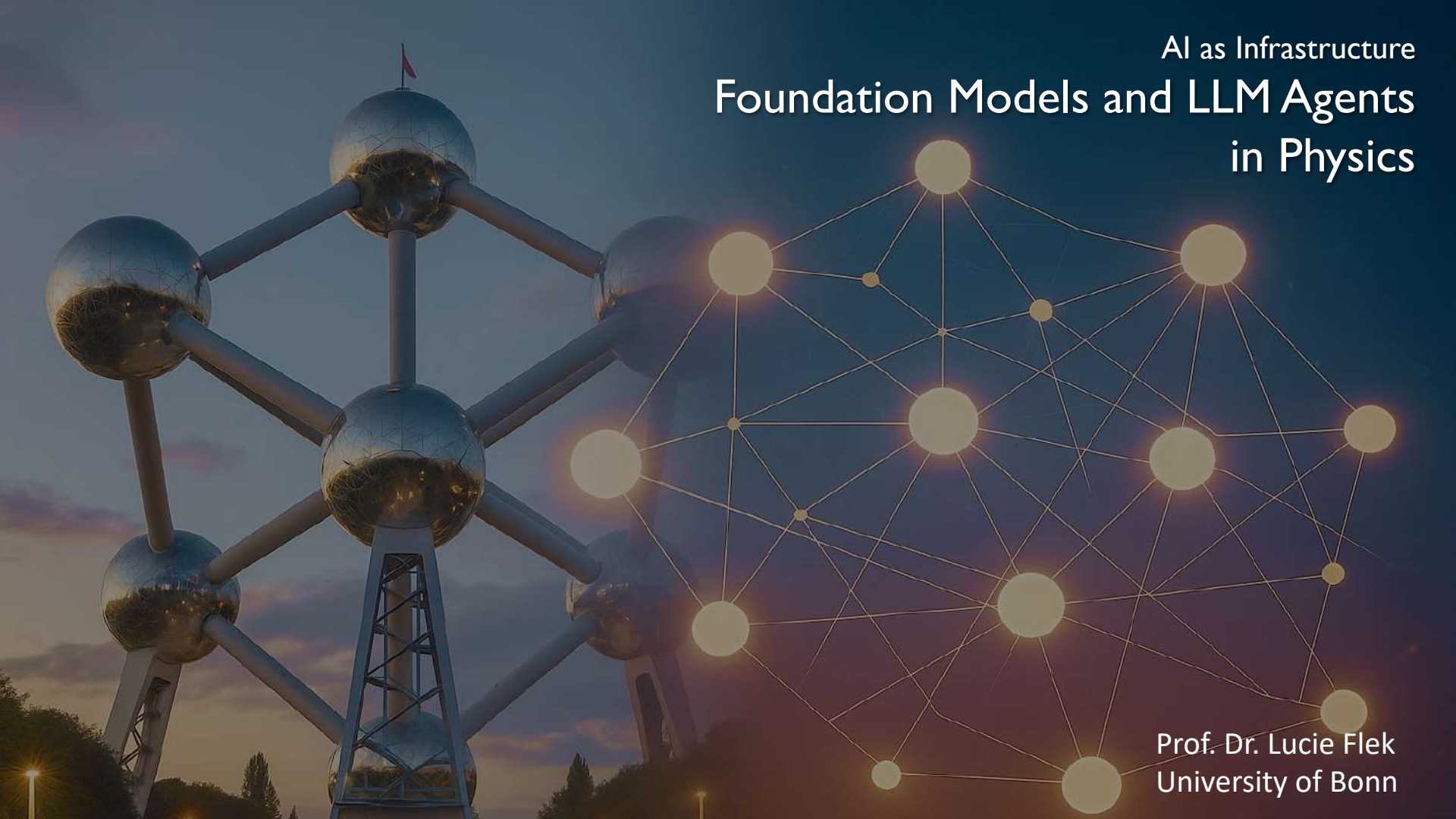


AI as Infrastructure

Foundation Models and LLM Agents in Physics

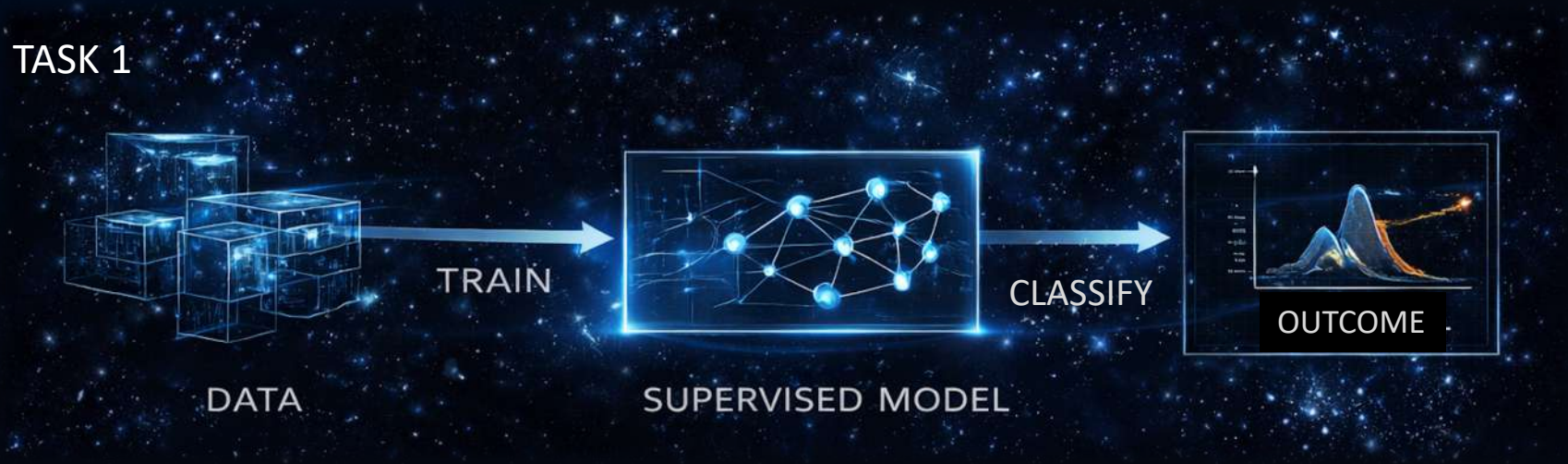
Prof. Dr. Lucie Flek
University of Bonn

Why Now?

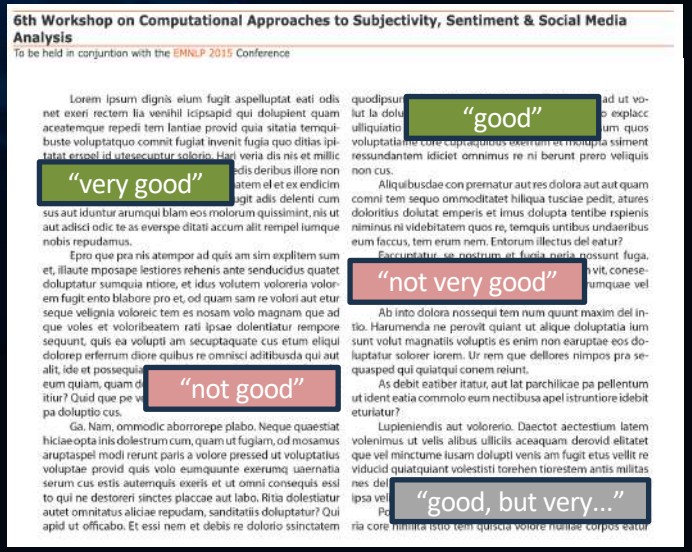

Data Scale ↑
Simulation Cost ↑
Workflow Complexity ↑
Human Iteration = Bottleneck

What ML for Physics used to be

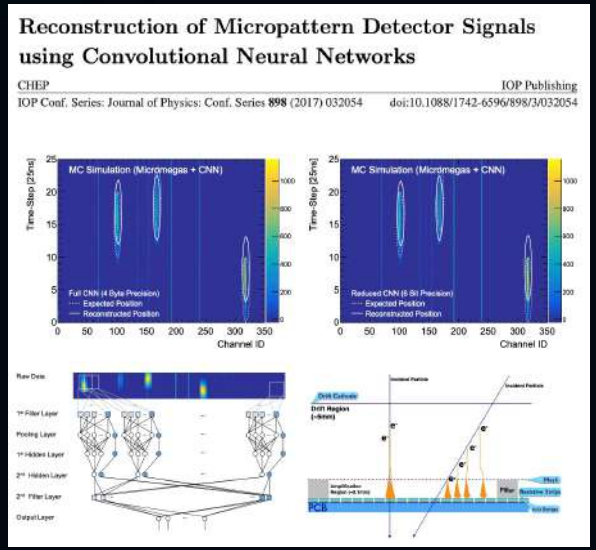
- **Task specific, supervised, one model per analysis**
- **High performance, but low reuse beyond task**



Machine Learning: Patterns = Knowledge



Flek et al., WASSA 2015

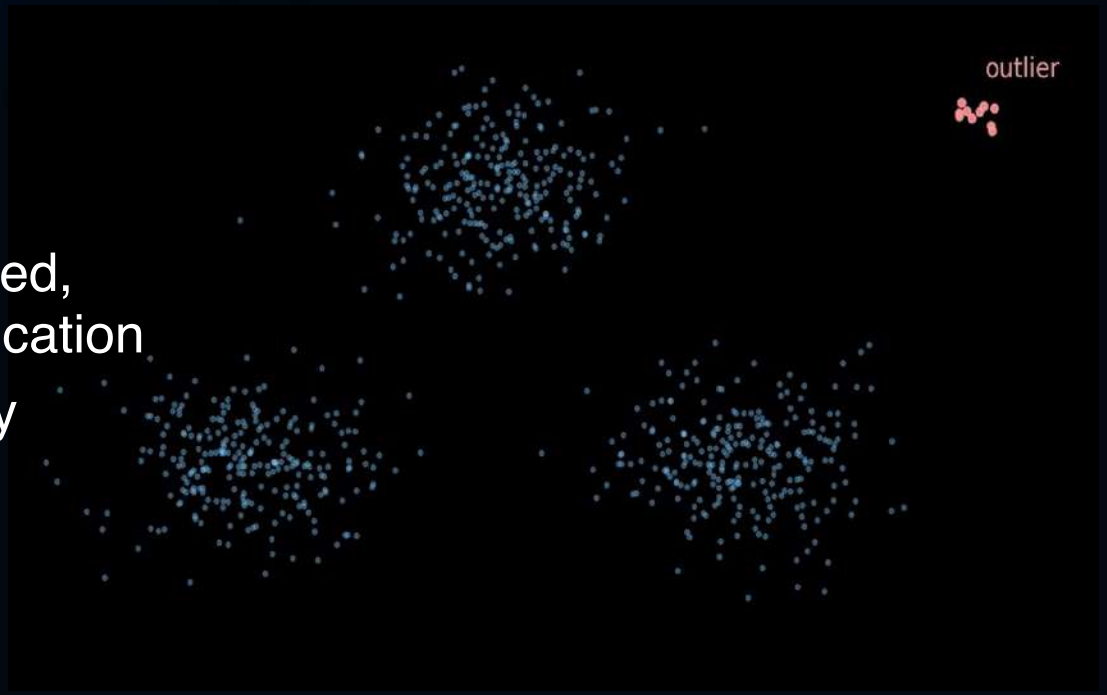


Flek et al., CHEP 2015

- Traditional statistical tools are great for **independent and identically distributed** (IID) variables
- Rarely the case in raw data; besides, data sometimes comes in infeasible to compute quantities
- Equivariant models such as CNNs excel in pattern recognition

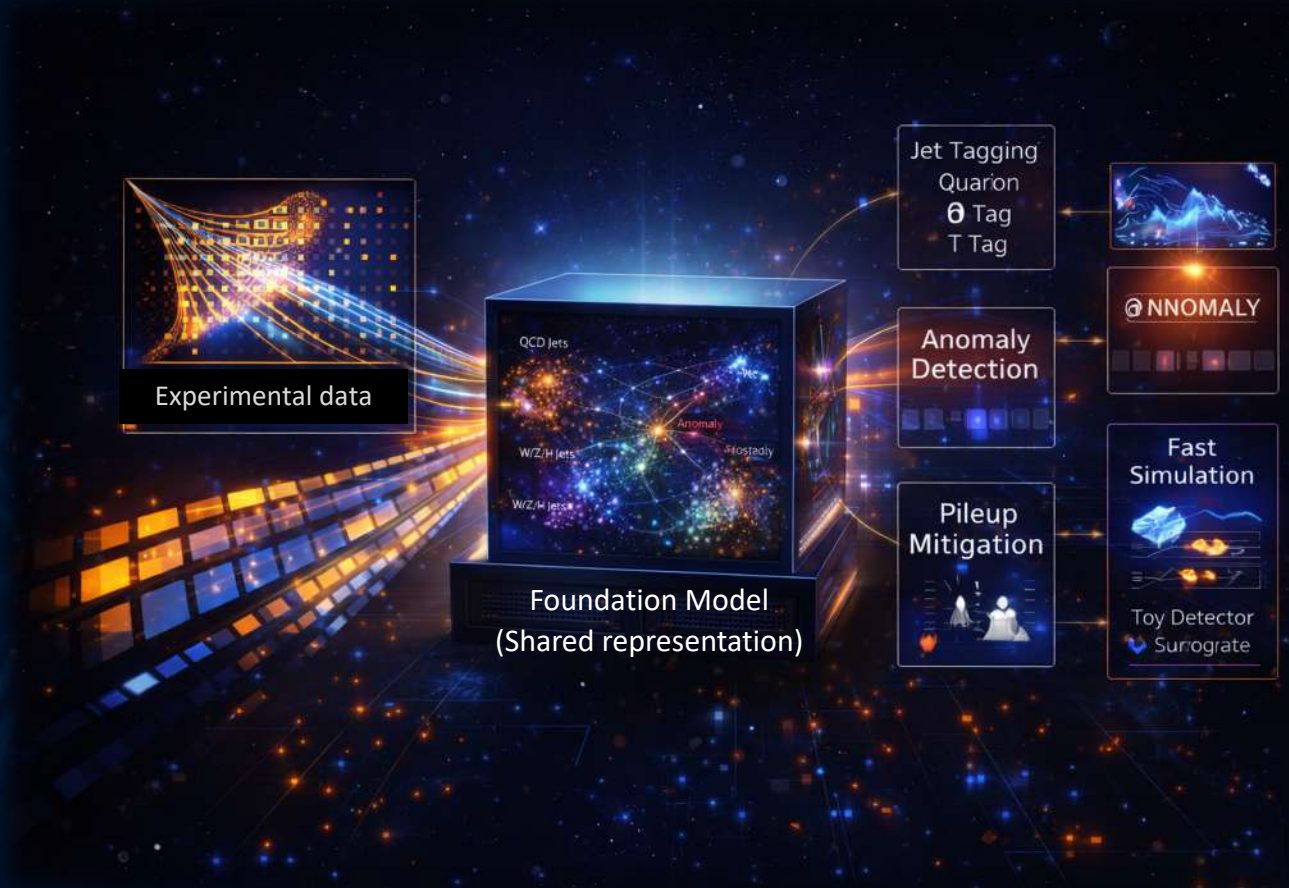
Why task-based ML is not enough

- Labels are scarce and biased, discovery is not just classification
- Domain shift and reusability
- Workfllow complexity

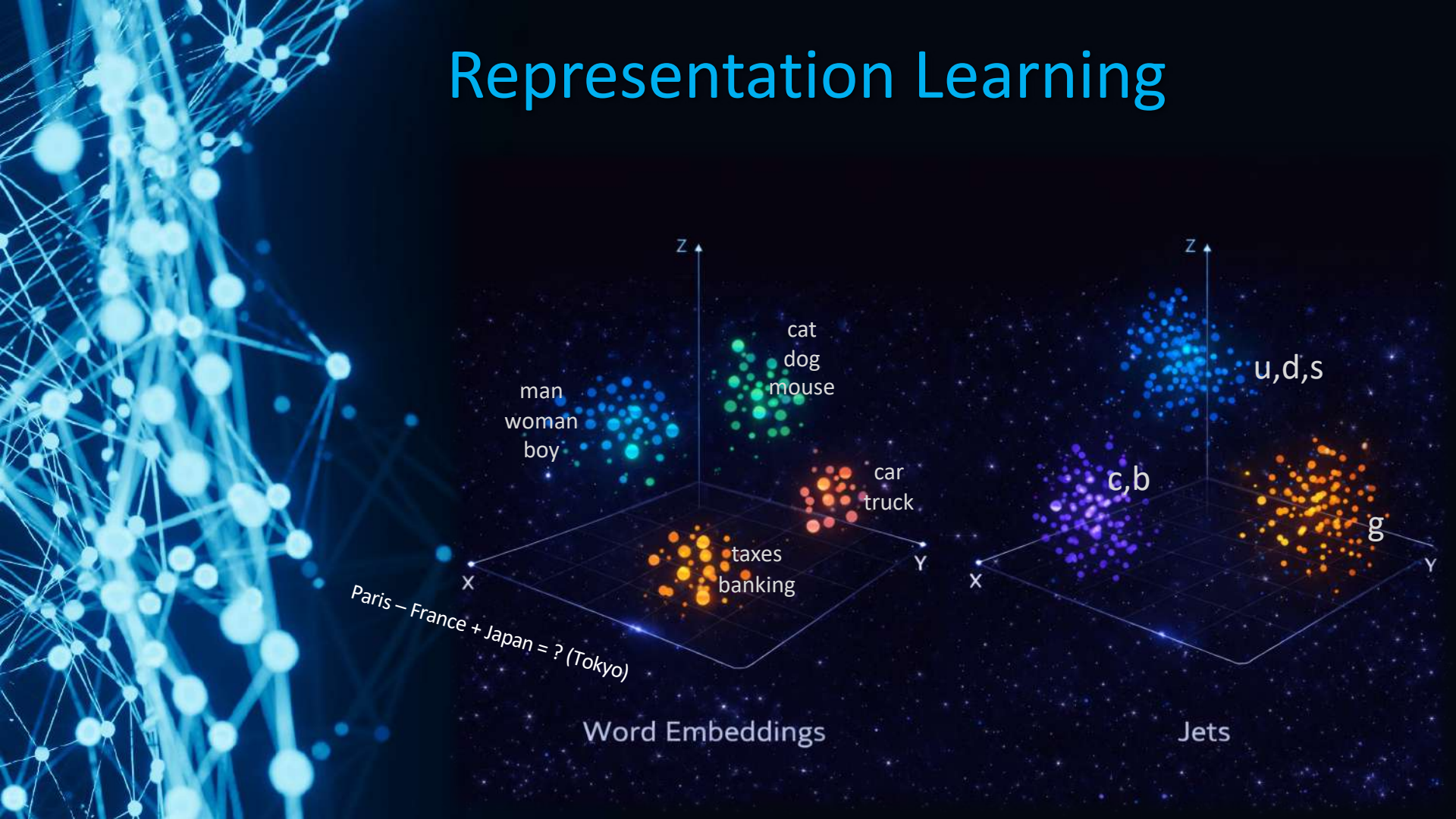


What is a Foundation Model

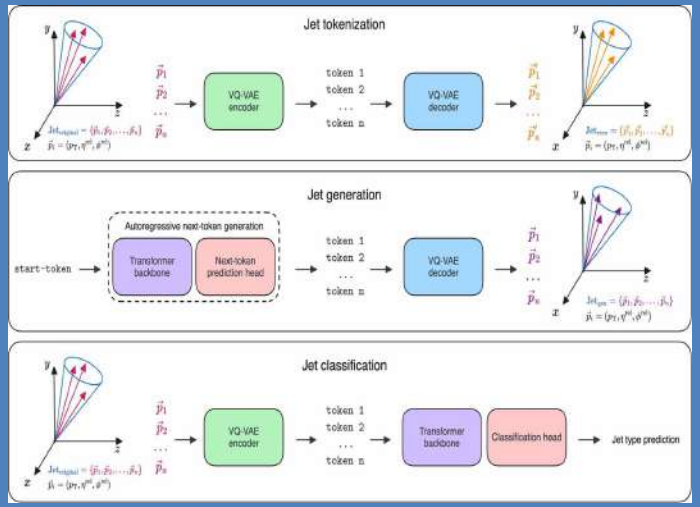
Foundation Models in HEP



Representation Learning

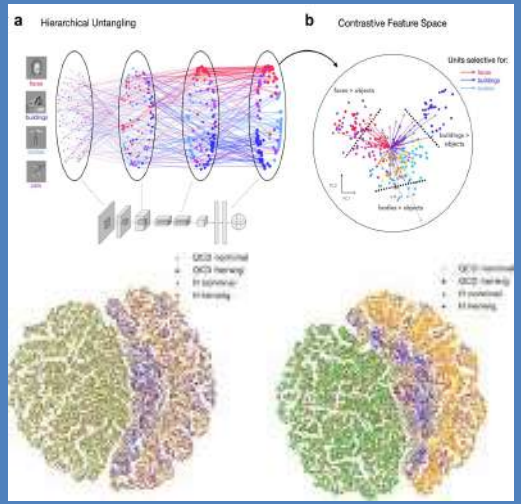


How does it work



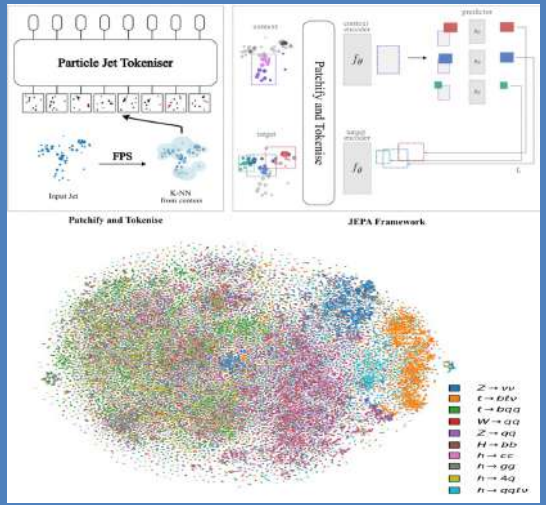
OmniJet-Alpha (2024)

The first cross-task foundation model for particle physics



RS3L (2024)

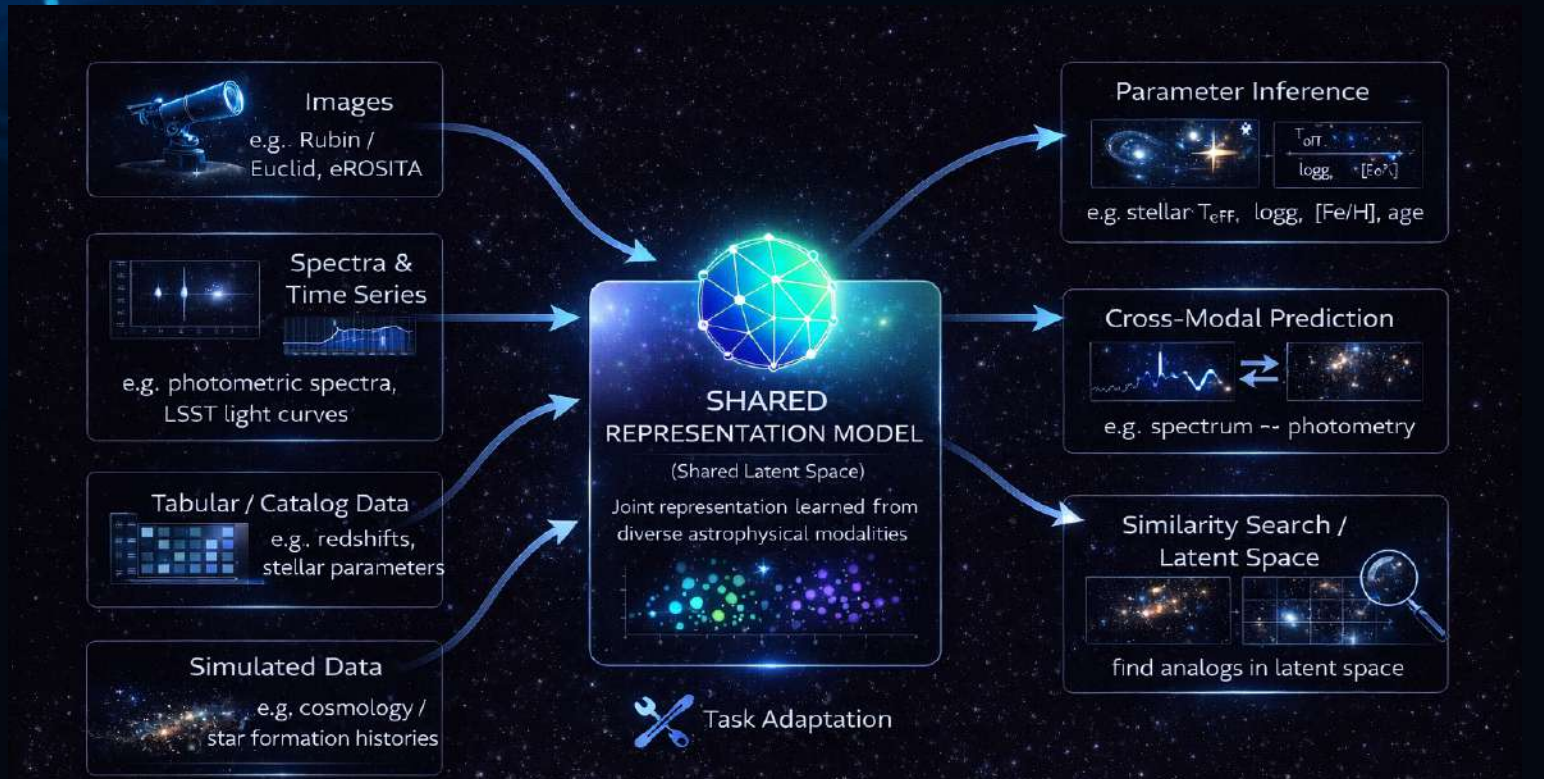
Contrastive Pretraining via Resimulation



HEP-JEPA (2025)

Joint Embedding Predictive Architecture for collider physics

Foundation Models in Astrophysics



Foundation Models in Astrophysics

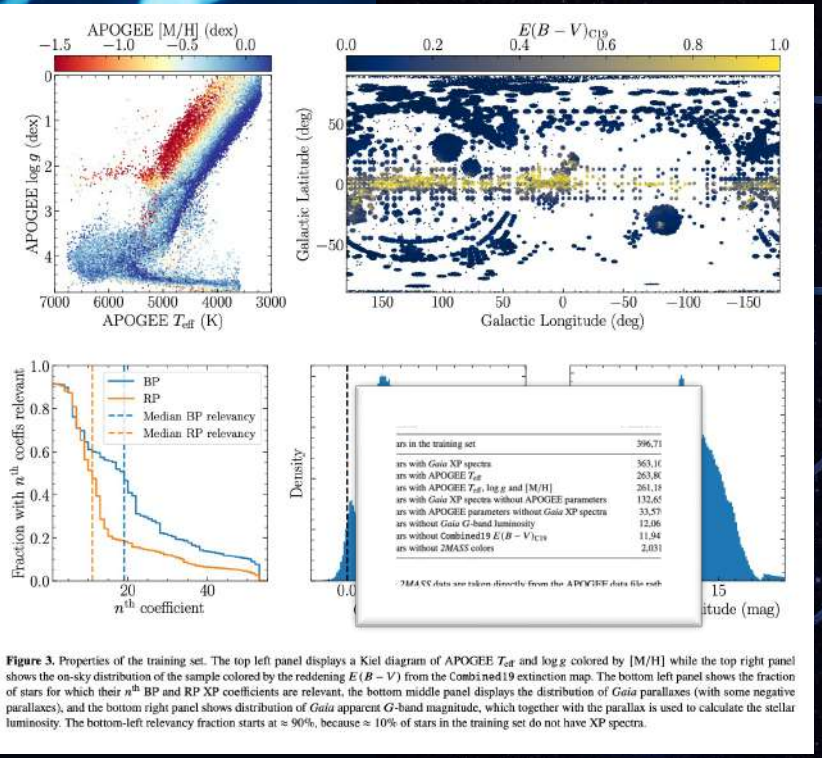
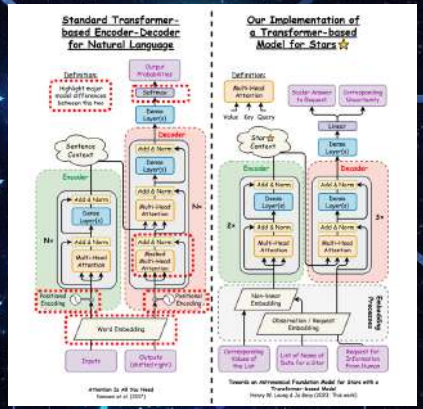
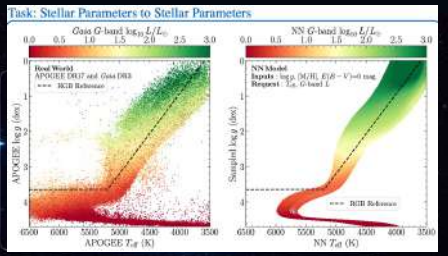
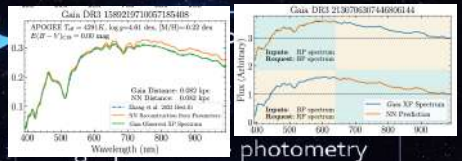
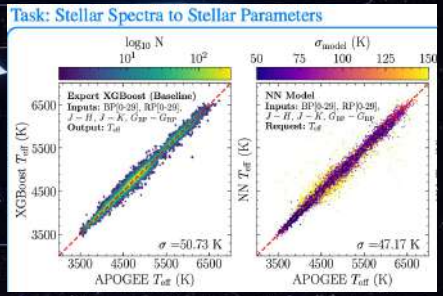


Figure 3. Properties of the training set. The top left panel displays a Kiel diagram of APOGEE $T_{\rm eff}$ and log g colored by [M/H] while the top right panel shows the on-sky distribution of the sample colored by the reddening $E(B-V)$ from the Combined19 extinction map. The bottom left panel shows the fraction of stars for which their $n^{\rm th}$ BP and RP XP coefficients are relevant, the bottom middle panel displays the distribution of *Gaia* parallaxes (with some negative parallaxes), and the bottom right panel shows distribution of *Gaia* apparent G-band magnitude, which together with the parallax is used to calculate the stellar luminosity. The bottom-left relevancy fraction starts at ≈ 90%, because ≈ 10% of stars in the training set do not have XP spectra.







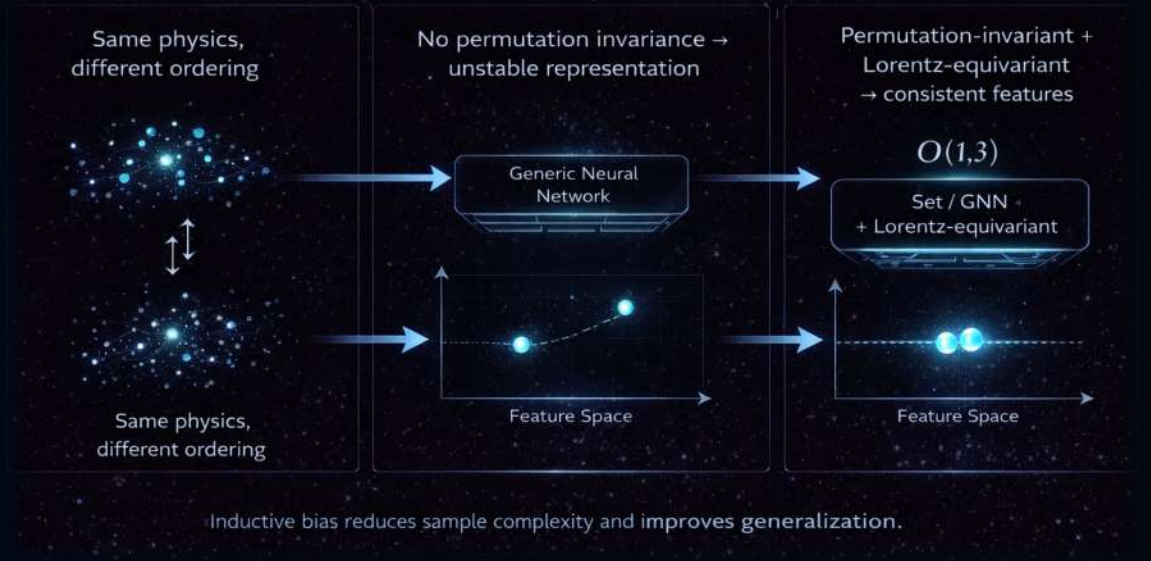


Task Adaptation

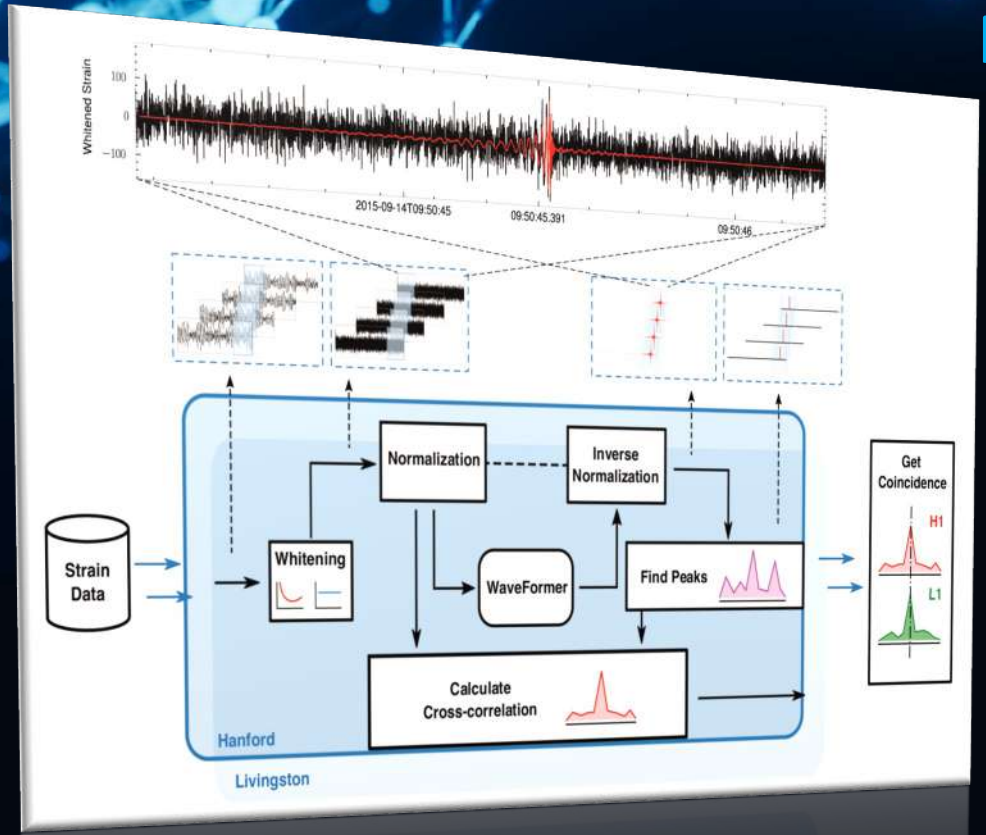
Astronomical Foundation Model for Stars (H. Leung, 2023)

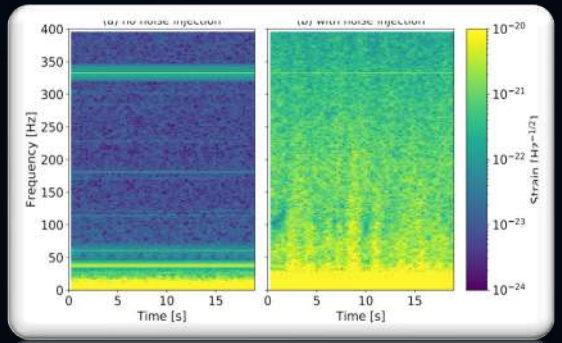
Physics-Aware Deep Learning

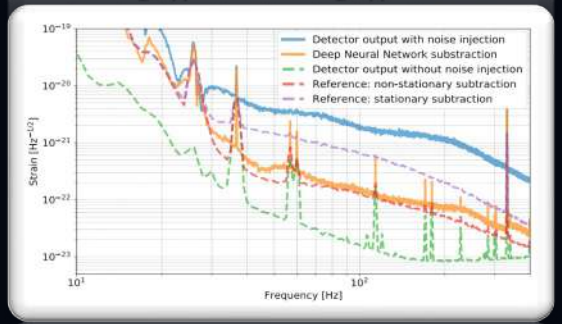
- Lorentz Invariance, detector geometry…
- symmetry-aware models in cosmology
- physically constrained time-series models in GW



Noise removal in GW



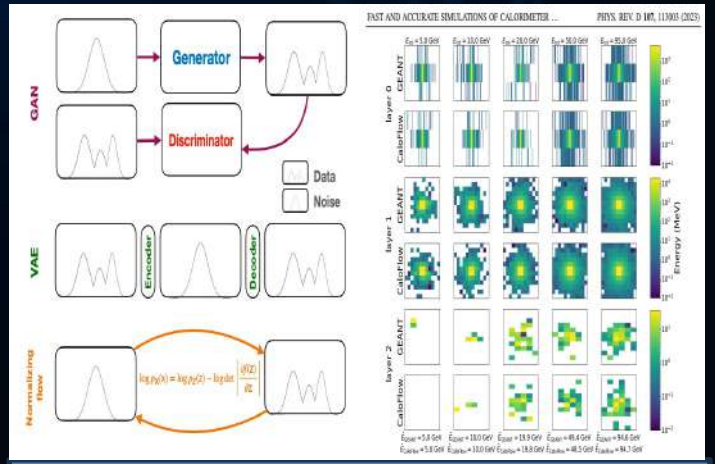




WaveFormer, DeepClean, Physics-informed ML for non-linear and non-stationary noise (slowxfast CNNs…)

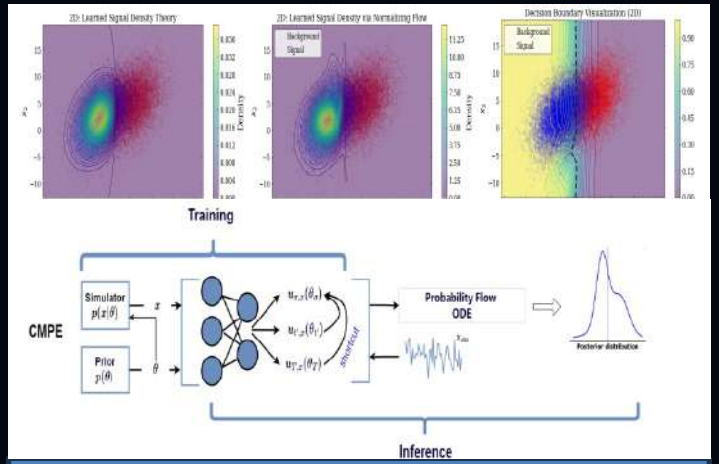
Simulation-Based Inference

- Orders of magnitude speed-up



Surrogates for Simulations

CaloFlow etc., but also e.g. GW (BBH NR) surrogates…
Typically GANs, VAEs, NFs



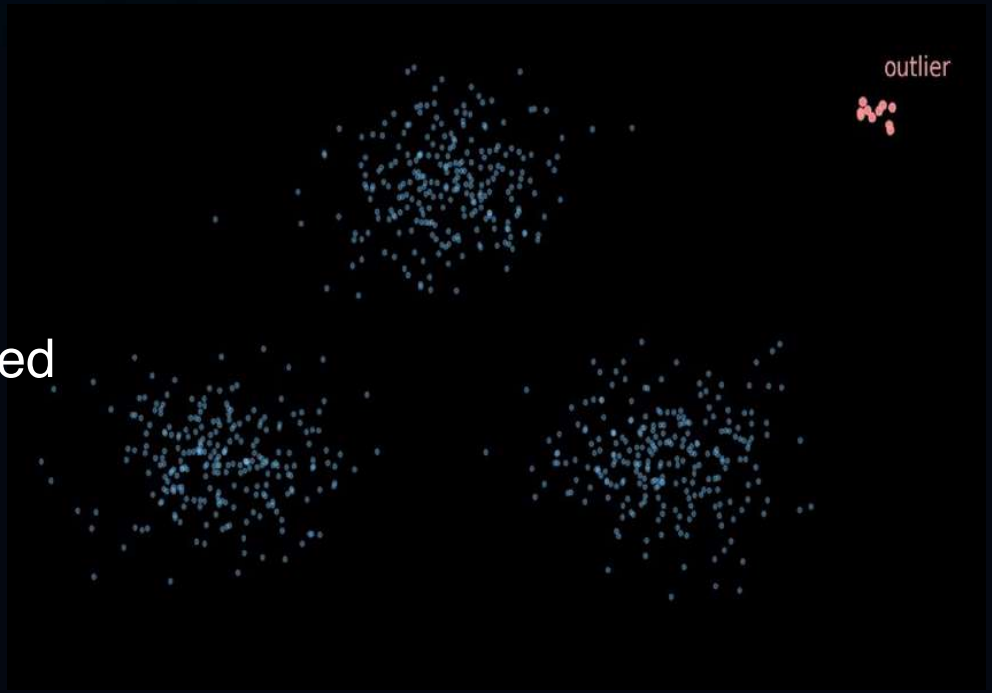
Simulation-Based inference

Simulation-based inference with consistency models for neural posterior est. (CMPE), Contrastive Normalizing Flows (CNFs)...

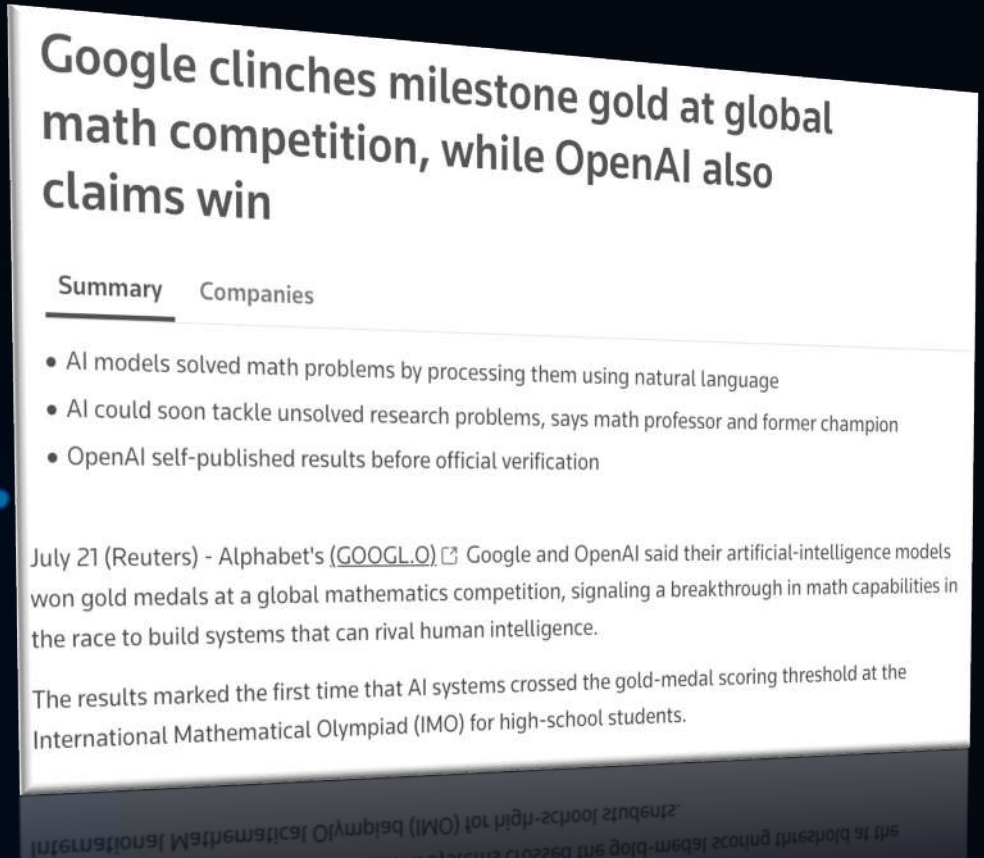
- Surrogate: $\theta \rightarrow x$ (fast detector response)
- SBI: $x \rightarrow p(\theta|x)$ (posterior, calibrated)

Discovery-oriented ML

- How to find rare/unexpected events without signal labels?
- Input: events / embeddings
- Output: anomaly score + ranked candidates (+ neighbors)



What about large language models?



Google clinches milestone gold at global math competition, while OpenAI also claims win

Summary | Companies

- AI models solved math problems by processing them using natural language
- AI could soon tackle unsolved research problems, says math professor and former champion
- OpenAI self-published results before official verification

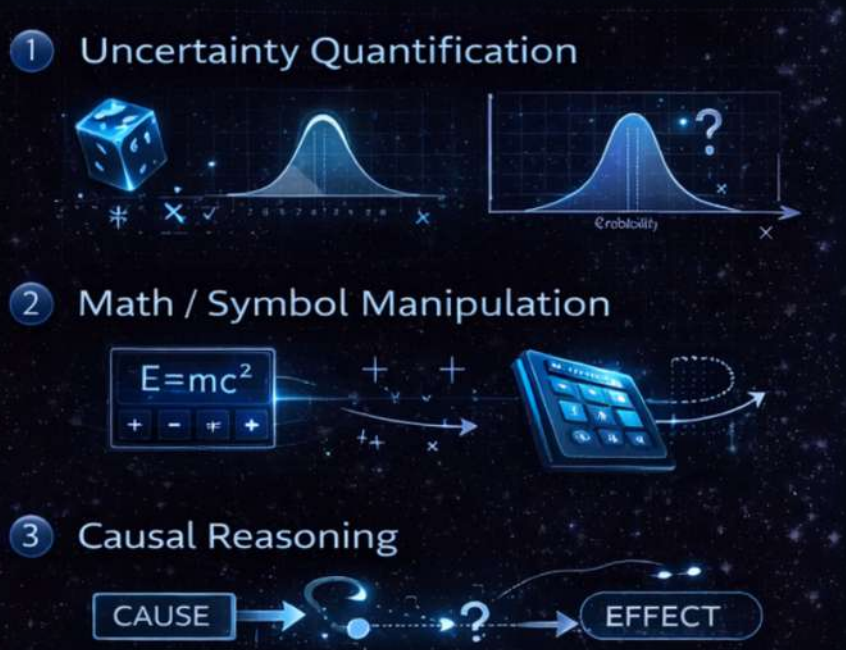
July 21 (Reuters) - Alphabet's (GOOGL.O) Google and OpenAI said their artificial-intelligence models won gold medals at a global mathematics competition, signaling a breakthrough in math capabilities in the race to build systems that can rival human intelligence.

The results marked the first time that AI systems crossed the gold-medal scoring threshold at the International Mathematical Olympiad (IMO) for high-school students.

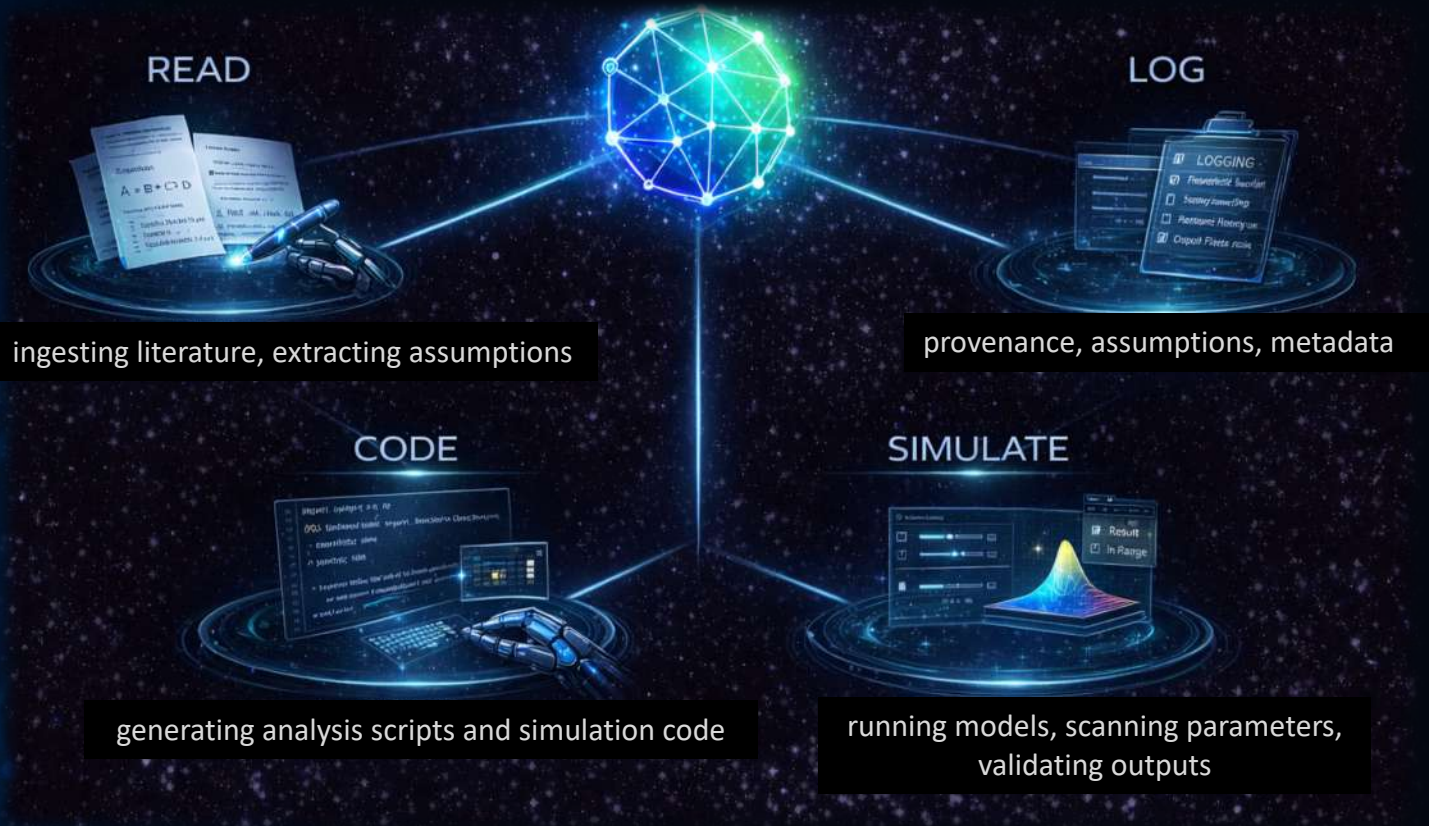
What about large language models?

LLMs are NOT physics inference engines

- **They are really bad at:**



What LLMs are good at



ingesting literature, extracting assumptions

provenance, assumptions, metadata

generating analysis scripts and simulation code

running models, scanning parameters, validating outputs

From LLMs to AI Agents

- Agent = LLM + tools + state + checks
- LLM chooses steps — tools produce numbers

LLM Agents: eROSITA

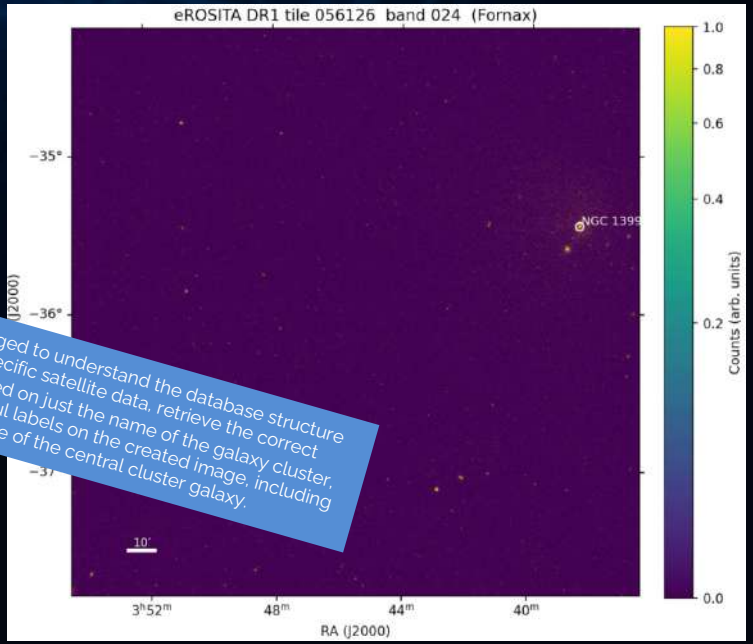
"Download one tile of eROSITA data that includes the Fornax galaxy cluster and plot an image"

Thought for 6m 3s

Thought for 38s

Thought for 18s

Short version: yes—Fornax sits in eROSITA-DE DR1's footprint (Galactic $l \approx 236.72^\circ, b \approx -53.63^\circ$), so there is a DR1 skytile covering it. DR1 data are organized into $3.6^\circ \times 3.6^\circ$ overlapping tiles named by `RRRDDD` (≈ RA° and (Dec°+90) rounded to the survey grid) with per-band images for 021–027/024, etc.



LLM managed to understand the database structure of these specific satellite data, retrieve the correct dataset based on just the name of the galaxy cluster, and put useful labels on the created image, including even the name of the central cluster galaxy.

LLM managed to download several tiles to create a larger image with the galaxy cluster in the center but **failed to notice there are overlap regions, resulting in grid-like artefacts in the merging process.**

LLM Agents in physics: GW

"Download the H1 and L1 detector strain data for GW150914 over a 12 sec window centered on the merger using TimeSeries fetch, then plot the q-transform spectroscopyplot with a normalized energy bar, whiten the data, do a lowpass filter at 250 hz and high pass at 30 hz and plot it, with both strain data on the same graph."

SCIENTIFIC INTERPRETER AGENT

Query + AI

CODER AGENT

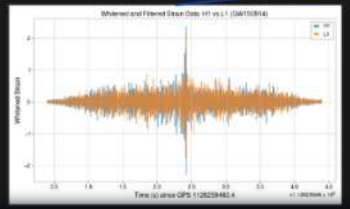
Query + DB + AI

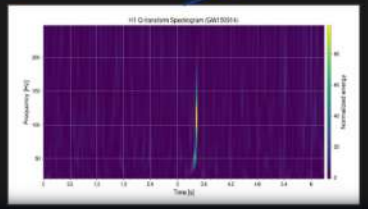
EXECUTOR AGENT

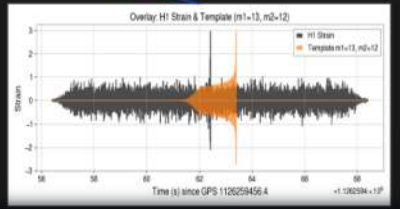
Task script execution

DEBUGGER AGENT

DB + AI





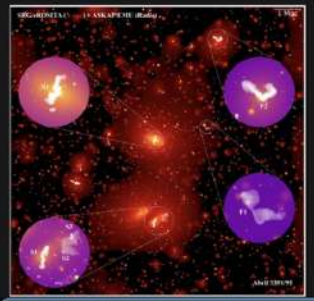


Credit:

LLM Agents: cosmology

simulation-based inference emulating dark matter halo mass functions or non-linear matter power spectra in varied cosmologies learning from N-body dark matter simulations.

Develop specific AI-agent for each major step.



SCIENTIFIC INTERPRETER AGENT
Interprets the query

ORCHESTRATOR AGENT
Routes to LLM Agents

N-BODY SIMULATION AGENT
Designs cosmology grid
Manages n-body runs

COMPRESSION AGENT
Does PCA projection
Measures P_nl(k, z), n(M, z)

EMULATOR CONSTRUCTION AGENT
Learns conditional probabilities in the grid

SB INFERENCE AGENT
Uses emulators as forward models, generates mock summaries, trains flow/NPE/NRE

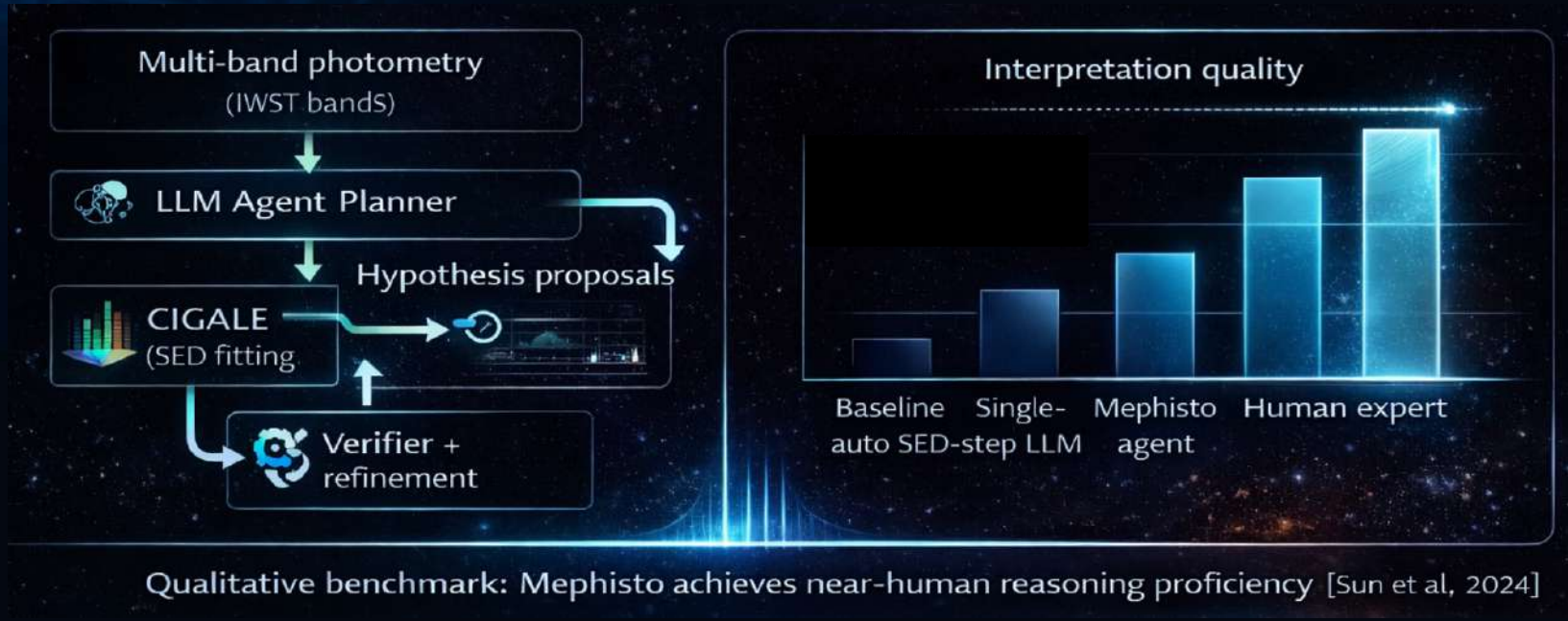
USER - SCIENTIST
Interprets posteriors and refines design

Provenance: dataset version • config hash • code commit • random seed • environment

Rule: if a verification check fails → stop and ask for human input

LLM Agents: Mephisto

Interpreting Multi-band Galaxy Observations
with Large Language Model-Based Agents" (arXiv:2409.14807)



Limitations

Verifiable science agents

Agents should route uncertainty — NOT invent it

Schema Checks

- Expected columns present
- Units properly declared
- Headers match requirements

Physics Checks

- Coordinate sanity
- Dimensional analysis
- Time window validation

Tool Evidence

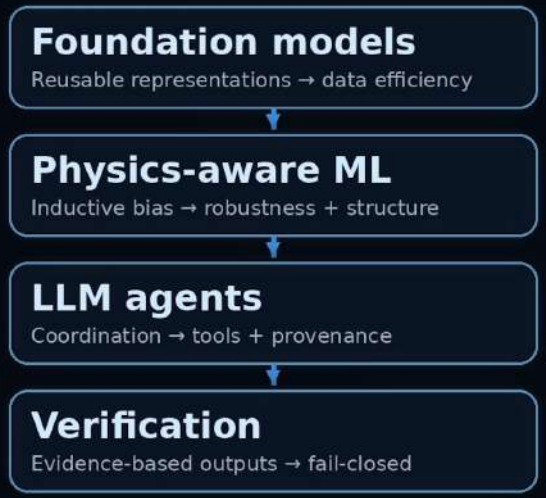
- Validates parameter claims
- Attaches output artifacts
- Executes code/tools

AI-Assisted Discovery Flow

Denario-style modular agents streamline discovery steps while keeping humans in the loop.

https://astropilot-ai.github.io/DenarioPaperPage/

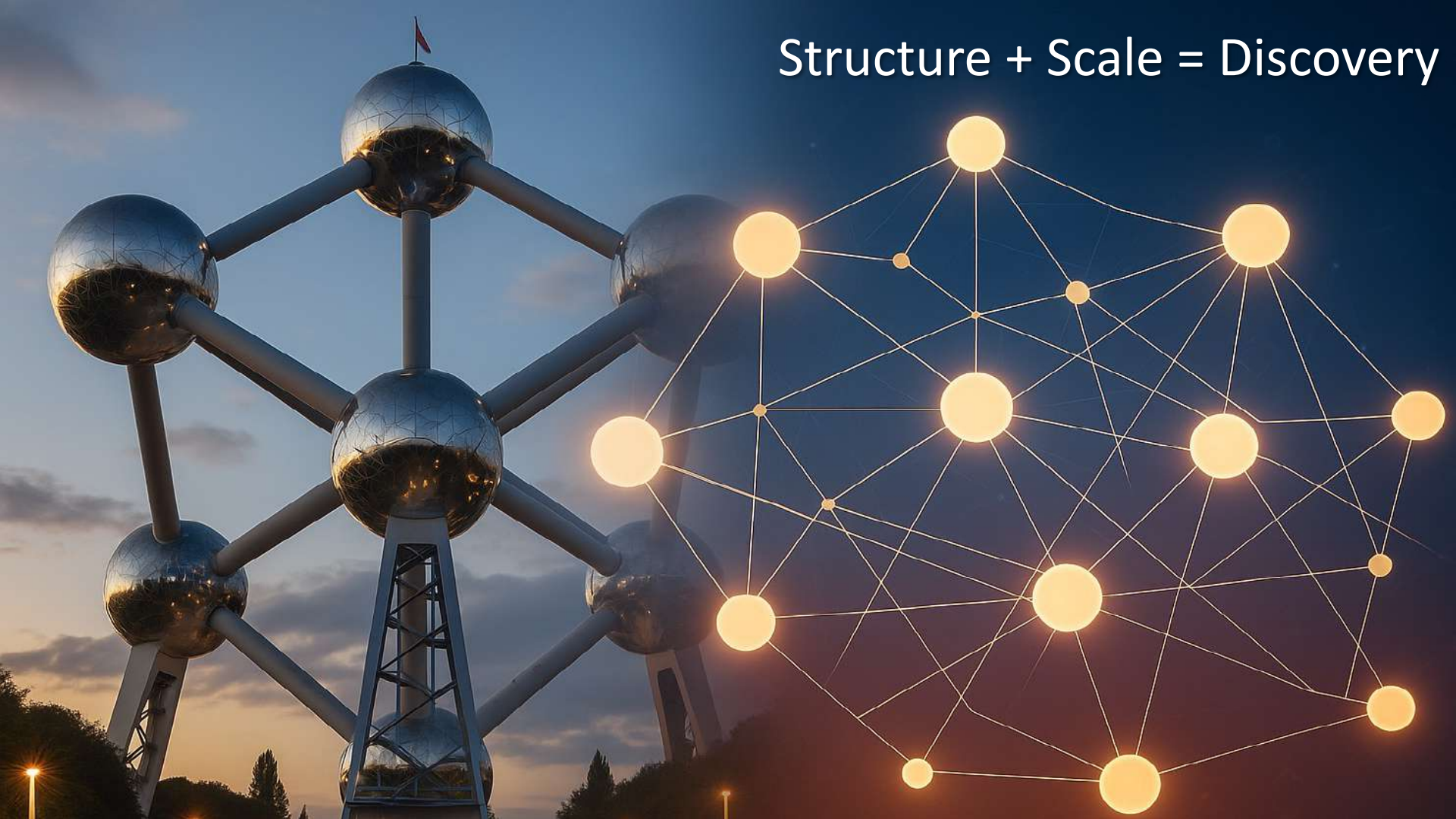
Take-aways



Agents don’t replace inference — they coordinate inference. Science stays physics-first, but becomes hopefully faster, more reproducible and easier to verify.

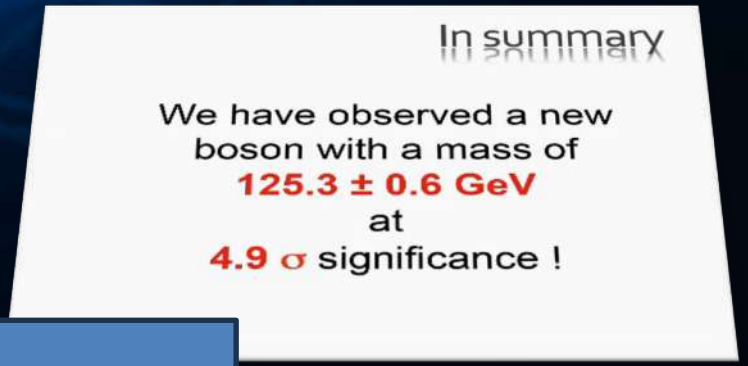
Thank you for your

!


Structure + Scale = Discovery

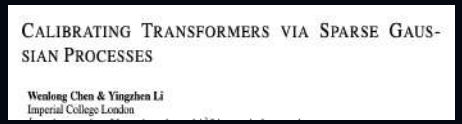

Backup Slides

Uncertainty = Good Science



CMS slides on Higgs, CERN, Dec 2012

It took AI engineers a decade to adopt a physics mindset…


CALIBRATING TRANSFORMERS VIA SPARSE GAUSSIAN PROCESSES

Wenlong Chen & Yingzhen Li
Imperial College London

ArXiV CS.AI, 10 Sept2025 https://arxiv.org/pdf/2303.02444

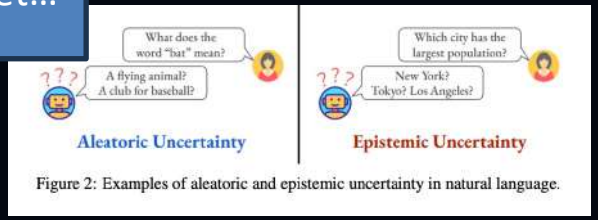
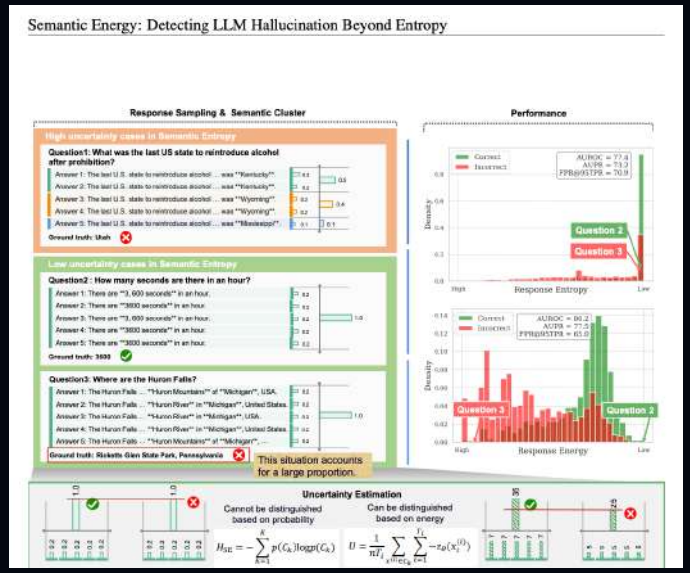


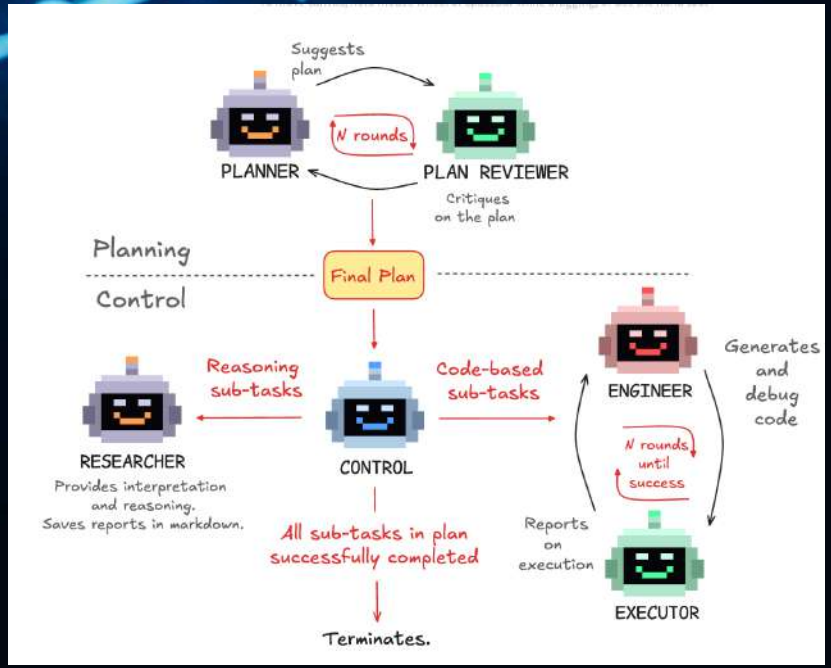
Figure 2: Examples of aleatoric and epistemic uncertainty in natural language.

Kang et al. ,ArXiV CS.AI, 14 Oct 2025 https://arxiv.org/pdf/2510.12040



ArXiV CS.AI, 27 Aug 2025 https://arxiv.org/pdf/2508.14496

AI Agents

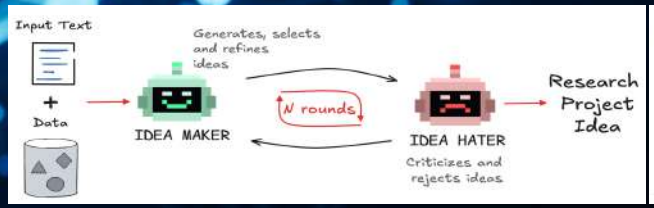


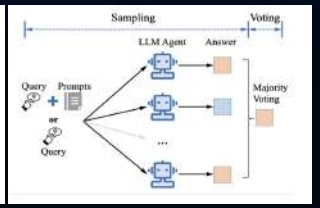
https://github.com/CMBAgents/

What agents already do:

- **Reading:** ingesting literature, extracting equations & assumptions.
- **Coding:** generating analysis scripts and simulation code.
- **Simulating:** running models, scanning parameters, validating outputs.
- **Comparing:** confronting predictions with data.
- **Logging:** tracking provenance, assumptions, metadata.

Hypothesis Generation





Input text

We have data from a recent gravitational wave event: GW231123. The data is located in five files:

- /mnt/ceph/users/fvillaescusa/AstroPilot/GW/Iteration1/data/GW231123_NRSur7dq4.csv
- /mnt/ceph/users/fvillaescusa/AstroPilot/GW/Iteration1/data/GW231123_IMRPhenomXO4a.csv
- /mnt/ceph/users/fvillaescusa/AstroPilot/GW/Iteration1/data/GW231123_SEOBNRv5PHM.csv
- /mnt/ceph/users/fvillaescusa/AstroPilot/GW/Iteration1/data/GW231123_IMRPhenomXPHM.csv
- /mnt/ceph/users/fvillaescusa/AstroPilot/GW/Iteration1/data/GW231123_IMRPhenomTPHM.csv

These files contain samples from the posterior distribution corresponding to five different gravitational-wave waveform models which are detailed below. The meaning of each column can be found in the file header, but is also provided below for reference.

We are interested in understanding which ways the different models agree or differ in their prediction for GW231123. There are complex degeneracies in the high-dimensional posterior space which make this problem challenging. Analyze the datasets in detail and tell us what you have learned from them. Mention also any interesting astrophysical insights that you learn from this analysis, and what can be robustly concluded about statistical properties of the high-mass black hole merger GW231123. Make sure there is no repetition in plots in the paper. For reference, the event was reported in https://arxiv.org/pdf/2507.08219

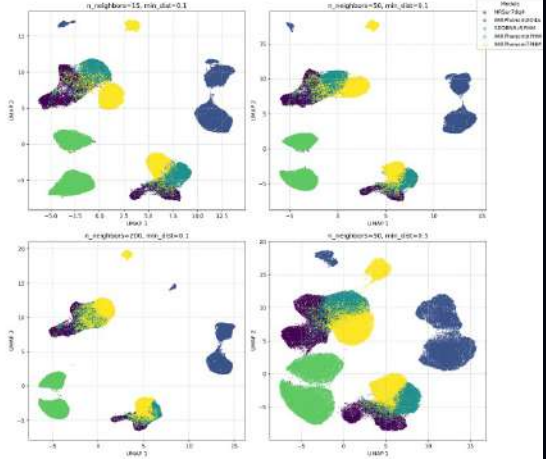


Figure 3. UMAP 2D embedding of the full posterior distributions for GW231123, colored by waveform model. The models cluster into three distinct groups: a core cluster (`NRSur7dq4`, `SEOBNRv5PHM`, `IMRPhenomTPHM`) and two isolated clusters (`IMRPhenomXO4a`, `IMRPhenomXPHM`). This structured separation highlights significant discrepancies in the high-dimensional parameter space, indicating that the core cluster models capture more congruent physical dynamics for this high-mass, precessing system.

https://arxiv.org/pdf/2510.26887

Scientific LLMs

Paper ID: 2023ARA&A..61..131F

Question: The properties of the circumgalactic medium (CGM) primarily depend on the competition between:
(A) Star formation rate and supernova feedback.
(B) Gas cooling and stellar winds.
(C) Gravity-driven infall and gas cooling.
(D) Magnetic fields and thermal conduction.

Correct Answer: C

Explanation: The article explicitly states that the defining characteristic of the CGM is the balance between gravity pulling gas inwards and cooling processes that allow gas to lose pressure and condense. This balance dictates whether the CGM is predominantly hot (slow cooling) or cold (rapid cooling).

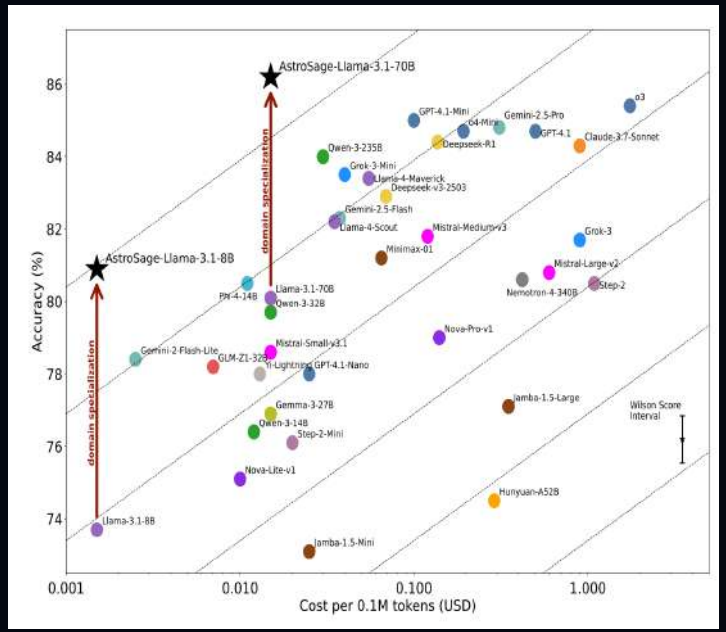
Paper ID: 2023ARA&A..61..473C

Question: What is the primary goal of calibrating subgrid feedback models in cosmological simulations?
(A) To ensure that simulations accurately reproduce the observed properties of the interstellar medium.
(B) To create a diverse range of galaxy morphologies in the simulations.
(C) To achieve convergence in simulation results across different resolutions and box sizes.
(D) To steer simulations towards producing a broadly realistic galaxy population that is consistent with key observational constraints.

Correct Answer: D

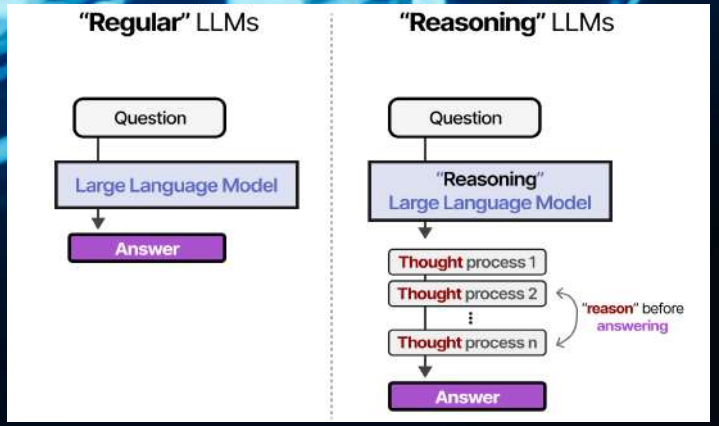
Explanation: The calibration of subgrid feedback models is primarily done to ensure that simulations produce a galaxy population that broadly aligns with key observational constraints. This is crucial because the microphysics governing feedback processes occur on scales much smaller than the resolution of cosmological simulations. By calibrating these models, simulations can better reproduce properties like the galaxy stellar mass function and the relationship between galaxy stellar mass and central supermassive black hole mass. This is discussed in section 2.4 of the article.

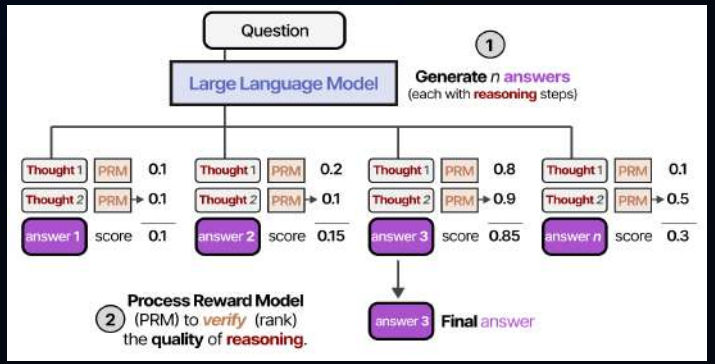
Model	Score (%)
OpenAI/GPT Series	
GPT-3.5	70.4
GPT-4	74.5
GPT-4o	80.4
Anthropic/Claude Series	
Claude-2.0	75.3
Claude-3.0-Haiku	77.9
Claude-3.0-Sonnet	76.7
Claude-3.0-Opus	82.7
Claude-3.5-Sonnet	**85.0**
Google/Gemini Series	
Gemini-1.0-Pro	71.0
Gemini-1.5-Flash	73.6
Gemini-1.5-Pro	77.6
Zhipu(智谱)/GLM Series	
GLM-3-Turbo	64.3
GLM-4-Flash	67.1
GLM-4-Air	72.9
GLM-4-AirX	72.5
GLM-4-0520	75.1
Baidu/ERNIE(文心一言) Se	
ERNIE-3.5	72.1
ERNIE-4.0	75.1
Deepseek(深度求索) Series	
Deepseek-v2	73.6



AstroSage, AstroBench https://arxiv.org/pdf/2407.11194, https://www.kaggle.com/competitions/tracs-wasp-2025/

Reasoning Errors in LLMs





Math Derivation Errors

- LLMs introduce algebraic steps that don't follow, ignores subscript
- Drop terms or misapply integration rules.
- Produce plausible-looking but invalid solutions.

HEP and Astro Reasoning Hallucinations

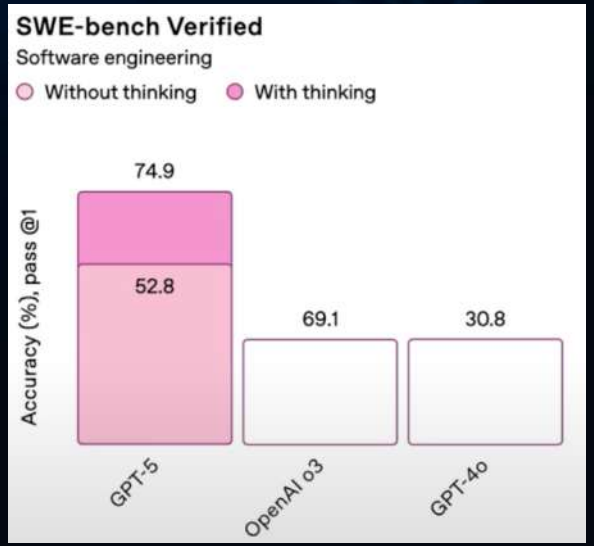
- LLMs invent detector constraints and wrong kinematic formulas.
- Produce non-existent ROOT functions, incorrect code, explanations physically false.
- mixing up redshift evolution and cosmic time,
- applying Newtonian approximations where relativistic ones are neede
- misinterpreting "magnitude" changes as linear rather than logarithmic,
- treating correlation as causation in exoplanet datasets

Error-Based Failures

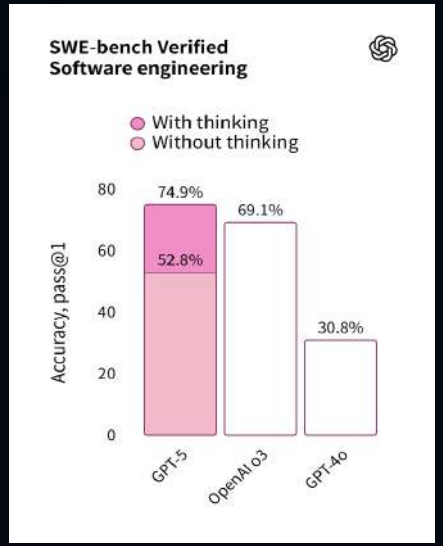
- treat statistical + systematic errors as additive when they should be quadrature-combined,
- propagate uncertainties incorrectly,
- or interpret confidence intervals as standard deviations.

Where LLMs fail

GPT5 release report v1:



GPT5 release report v2:



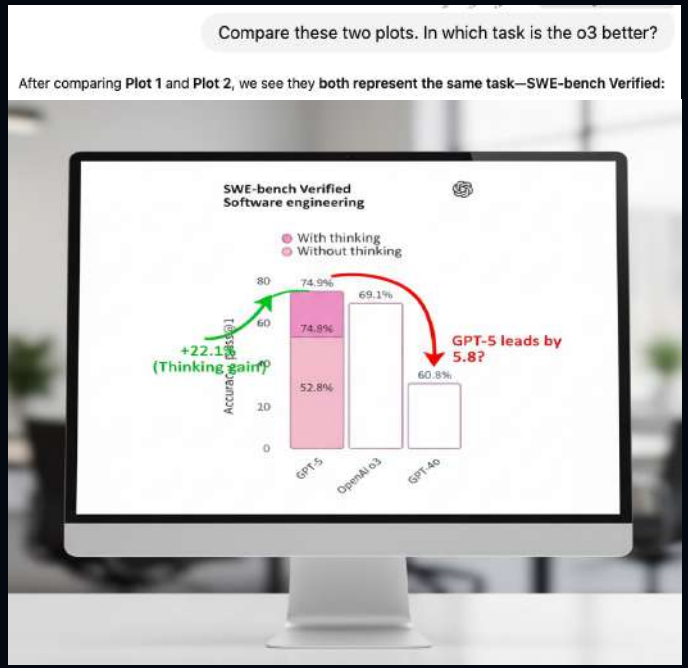


Table Understanding and (Multimodal) LLMs: A Cross-Domain Case Study on Scientific vs. Non-Scientific Data

Ekaterina Borisova[1,2], **Fabio Barth**[1], **Nils Feldhus**[1,2,3], **Raia Abu Ahmad**[1,2], **Malte Ostendorff**[4], **Pedro Ortiz Suarez**[5], **Georg Rehm**[1,6], **Sebastian Möller**[1,2]

[1]Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), [2]Technische Universität Berlin, [3]BIFOLD, [4]Deutsche Telekom, [5]Common Crawl Foundation, [6]Humboldt-Universität zu Berlin

Can LLMs Generate Tabular Summaries of Science Papers? Rethinking the Evaluation Protocol

Weiqi Wang[♣♠], **Jiefu Ou**[♣], **Yangqiu Song**[♠], **Benjamin Van Durme**[♣], **Daniel Khashabi**[♣]
[♣]Center for Speech and Language Processing, Johns Hopkins University, Baltimore, MD, USA
[♠]Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China

	Years Ended December 31,						
	2014	2013	2012	2014 vs. 2013		2013 vs. 2012	
				$ Change	% Change	$ Change	% Change
Routing	$ 2,223.9	$ 2,318.0	$ 2,037.6	$ (94.1)	(4)%	$ 280.4	14%
Switching	721.2	638.0	554.8	83.2	13 %	83.2	15%
Security	463.6	563.9	669.7	(100.3)	(18)%	(105.8)	(16)%
Total Product	3,408.7	3,519.9	3,262.1	(111.2)	(3)%	257.8	8%
Percentage of net revenues	*73.7 %*	*75.4 %*	*74.7 %*				
Total Service	1,218.4	1,149.2	1,103.3	69.2	6 %	45.9	4%
Percentage of net revenues	*26.3 %*	*24.6 %*	*25.3 %*				
Total net revenues	$ 4,627.1	$ 4,669.1	$ 4,365.4	$ (42.0)	(1)%	$ 303.7	7%

Figure 22: Table image corresponding to the ComTQA (FinTabNet) example in Figure 5.