AI at the superhuman level

Włodzisław Duch

Katedra Informatyki Stosowanej, INT WFAiIS UMK
Laboratorium Neurokognitywne, ICNT UMK

Google: Wlodzislaw Duch

AI 360°: Innowacje, Przemysł, Społeczeństwo, Nauka, Bielsko-Biała, 23.10.2025

Towards autonomous AI scientists

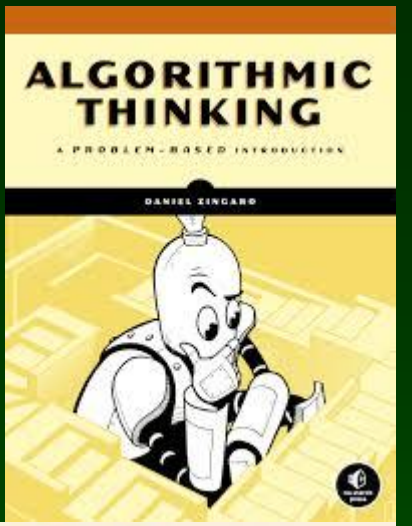
1. Few thoughts on AI: computing, intelligence, cognition.
2. Superintelligence and exponential growth.
3. State-of-the-art (SOTA).
4. Towards autonomous agentic science.
5. Agentic science future, or can science without strong support of AI agents be relevant?

ChatGPT << AI, 100-300 papers in the arxiv.cs.ai each day! Too fast to plan.
We see only the tip of the iceberg … but technological tsunami is coming.

Dilemma: Should I stay, or should I go now?
If I go, there will be trouble, and if I stay it will be double.
So come on and let me know …

Duch W and Diercksen GHF (1994) *Neural networks as tools to solve problems in physics and chemistry*.
Computer Physics Communication **82**: 91-103

Defining intelligence

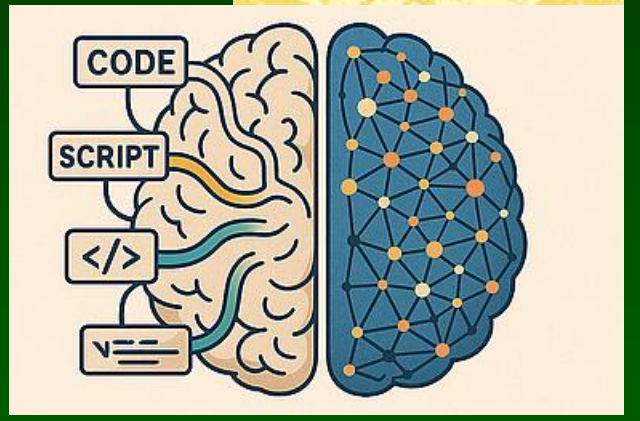


For many problems we have **effective** algorithms, so we can write programs to solve them and there is no need for AI.

AI solves problems for which there are **no effective algorithms**.

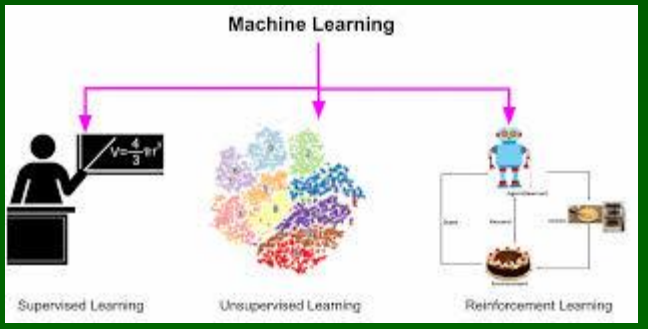
Hardware could be biological, electronic, photonic, spintronic, atomic …

Intelligence: ability to solve problems where no effective algorithms are known. AI does it using some hardware, humans use wetware.



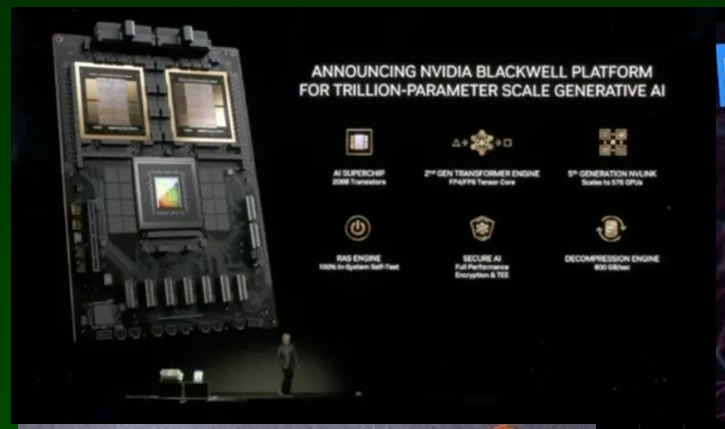
What cannot be successfully programmed but can be learned from experience may be replicated in a **supervised** way, discover in **unsupervised** way, or strategies may be developed using **reinforcement learning**.

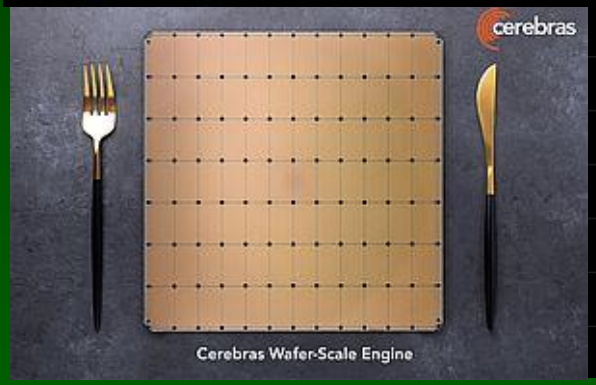
AI cannot be perfect, but it can be better than humans, learn more and work faster with texts, images, videos, sensor signals, control systems.



Humans are megalomaniacs and believe in magic in their wetware …

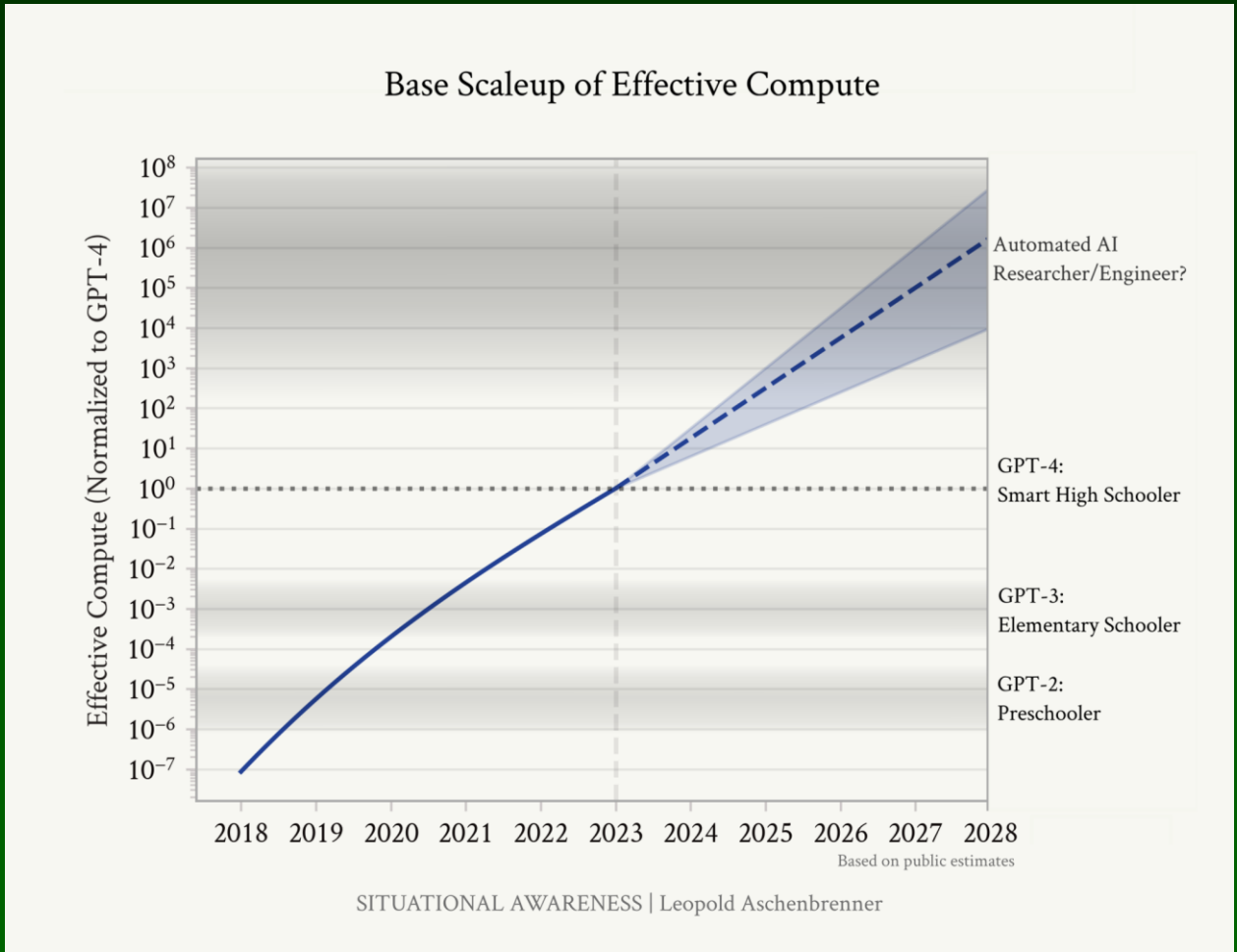
Hardware, another substrate for thinking

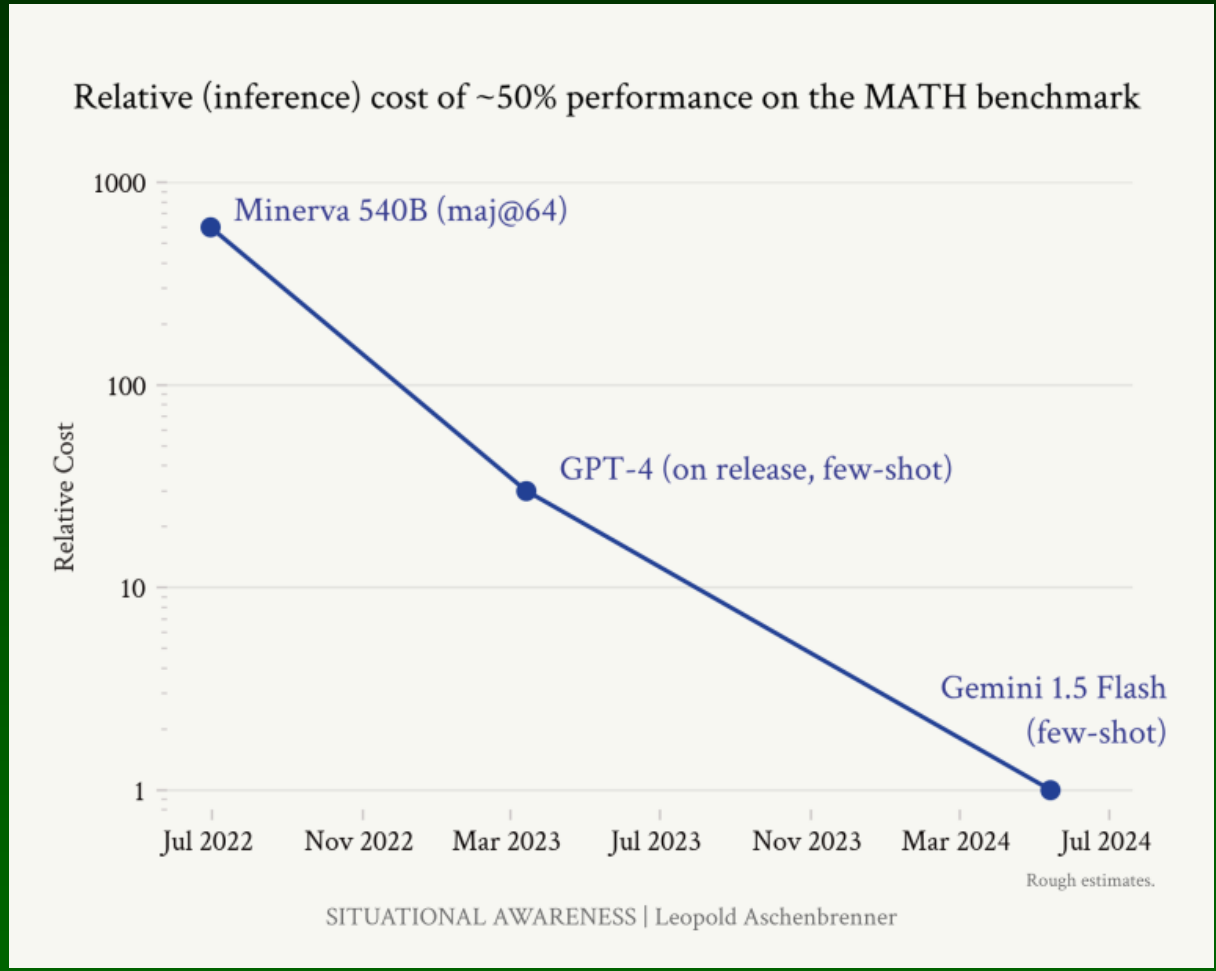




No constraints of biological brains. New energy efficient computing paradigms: photonics: Taichi chiplet 160 TOPS/W, Q.ant photonic chip, Lightmatter 3D Photonic 500-1000 TOPS/W. Groq Linear Processing Unit LPU, 400 TOPS/W. Cortical Labs organoids CL1 800 000 cells chip. Cerebras, CIMs, reversible computing, neuromorphic computing!
LLM inference cost fall between 9-900x a year, depending on the task (Epoch.ai).

Scaling costs and computations





Robotics costs also drops dramatically. World in 5 years time will be very different.
My phone can do locally live translate in 13 languages, requiring 1W of power! My brain needs 20 W.

Transformers and spreading activation

Predictive AI: search + heuristics.

Generative AI: spreading activation networks, binding relevant information.

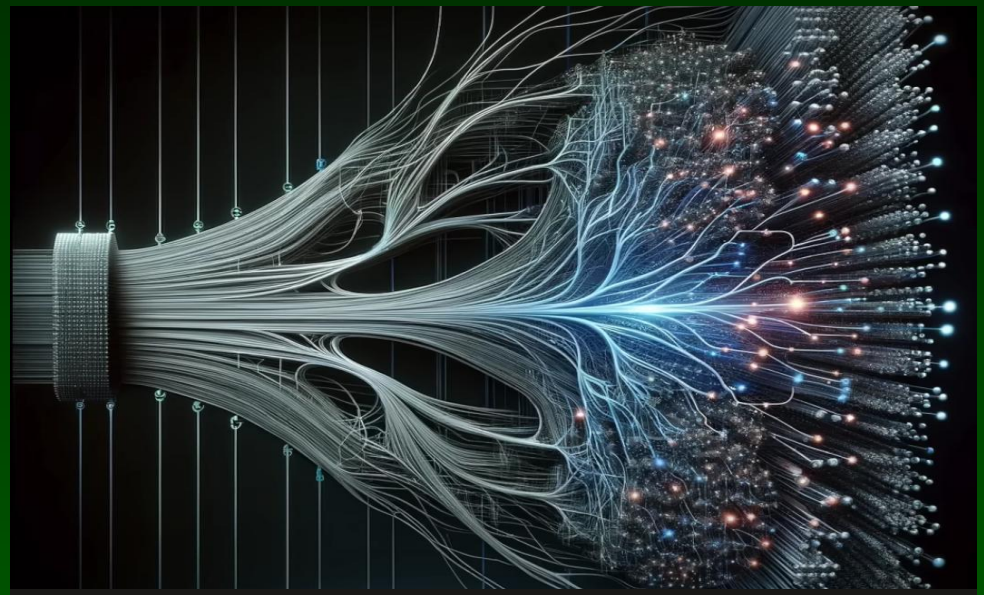
GPT = Generative Pre-trained Transformer

Convert input data into tokens, and embed them in numerical vectors space that preserves similarity of meaning (semantics) in different contexts.

Analyze sequences or fragments, pay attention to links among all tokens that add to their interpretation – select subgraphs of tokens.

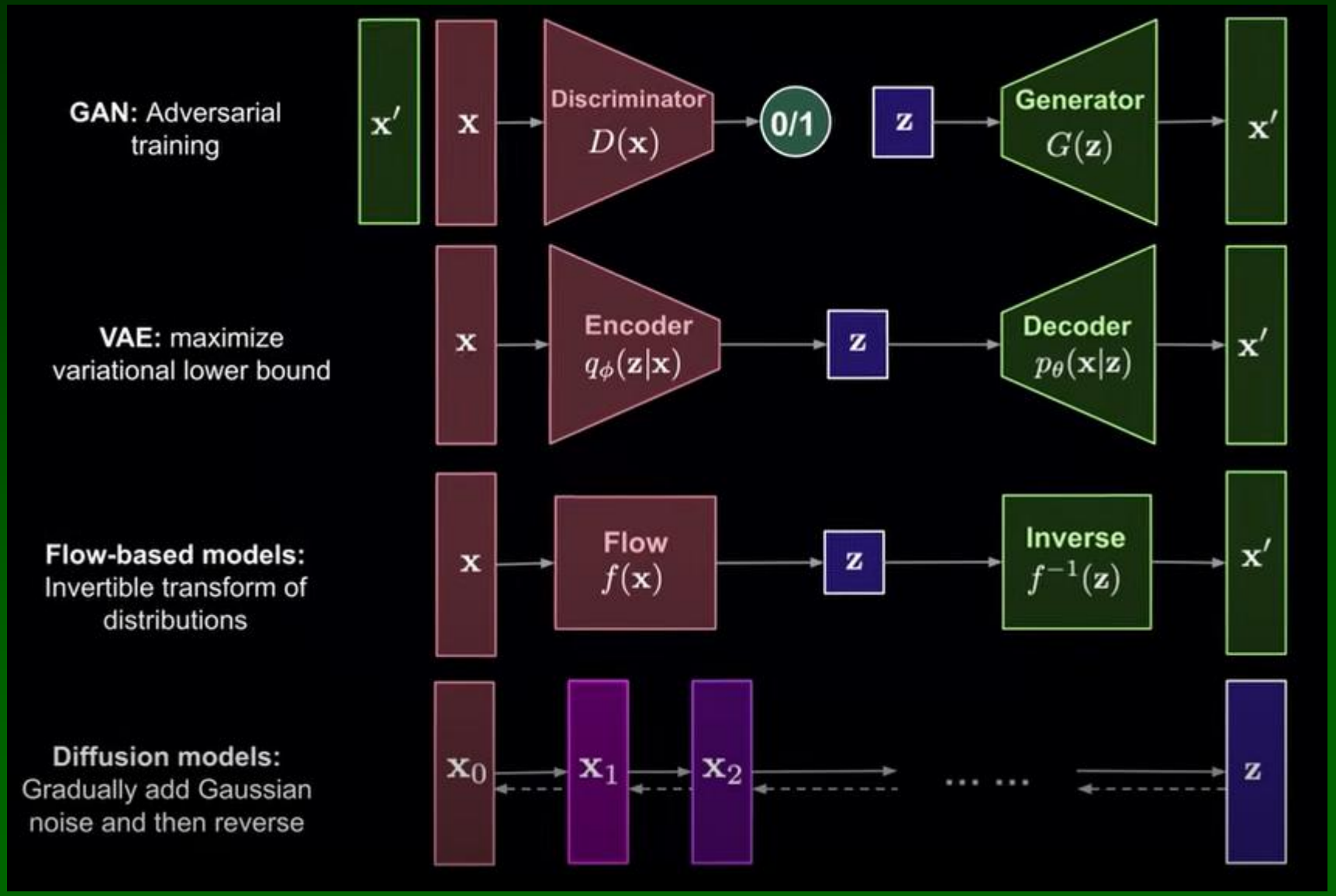
LLM visualization https://bbycroft.net/llm

Over 51 550 variants of open-source LLMs, including > 3000 for fine-tuning (LLM Explorer, 24/10/25).



Processing clinical text with domain-specific spreading activation methods. US Patent US8930178B2 (1/2015)
Duch W, Matykiewicz P, Pestian J, Towards Understanding of Natural Language: Neurocognitive Inspirations. 2007

GenAI: Latent probabilistic models



Compressed internal representations (pdfs):

GAN (2014), VAE (2022), Flows (2019), diffusion models (2015).

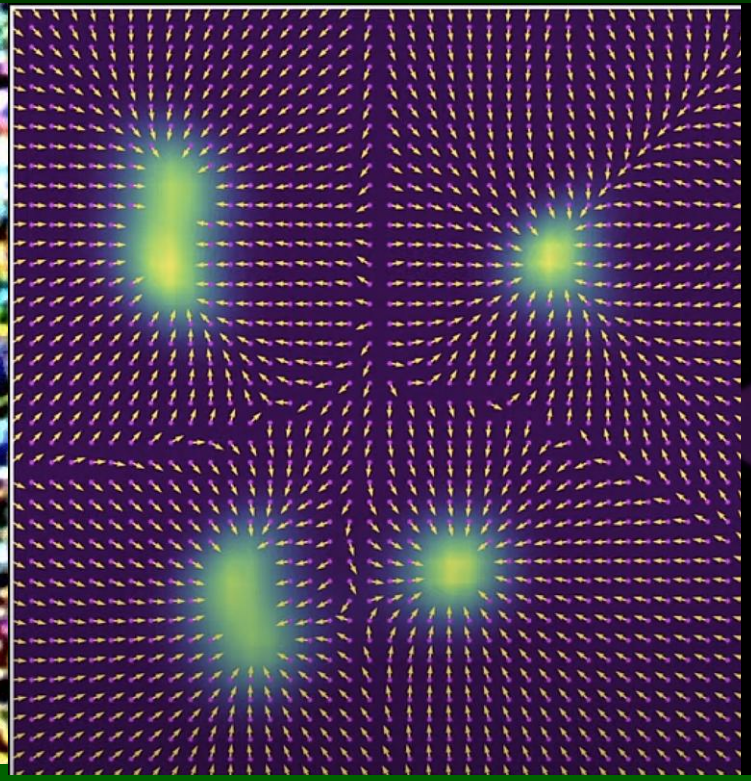
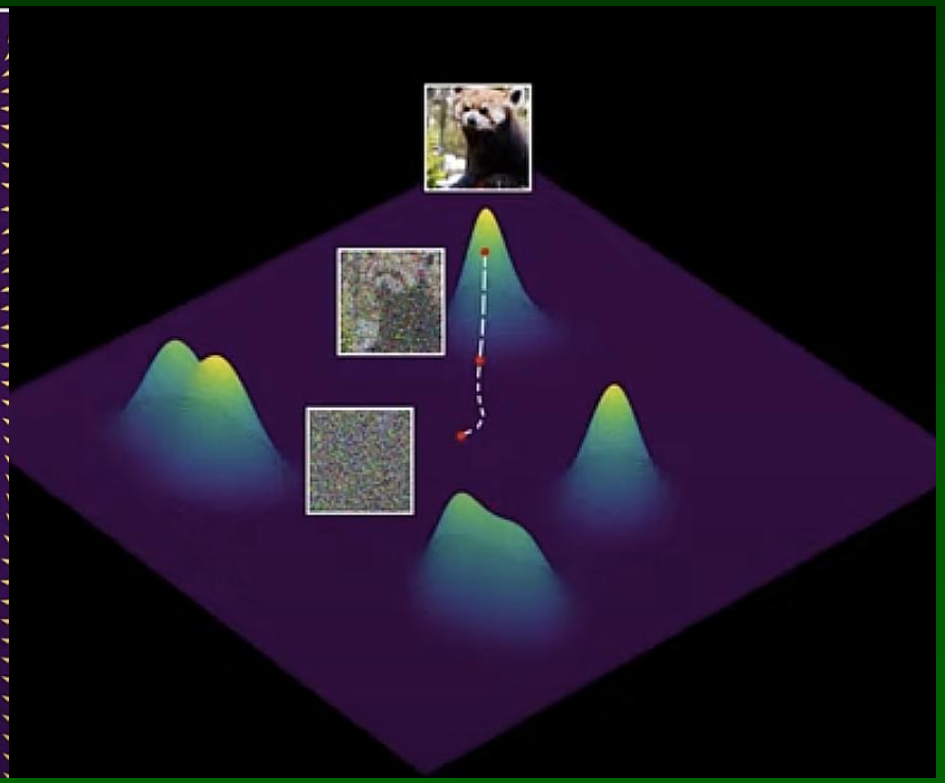
EBM, Energy-based models.

Many models based on the idea of information compression create vectors in the latent space capturing meaning in different contexts.

Algorithmic (Kolmogorov) information with low complexity is the key.

Diffusion models

Sometimes we try to complete a sentence, but sometimes it comes all at once.
Create gradient field towards attractor basins where images/concepts linked to prompts are stored.

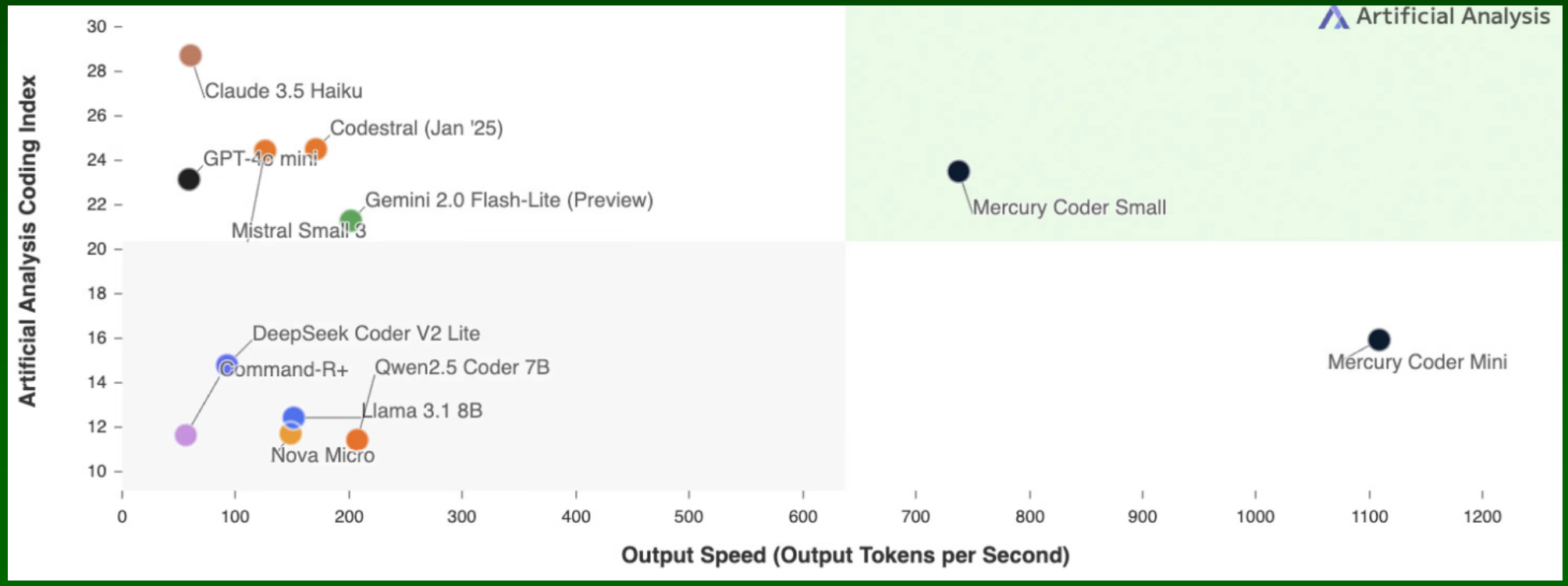




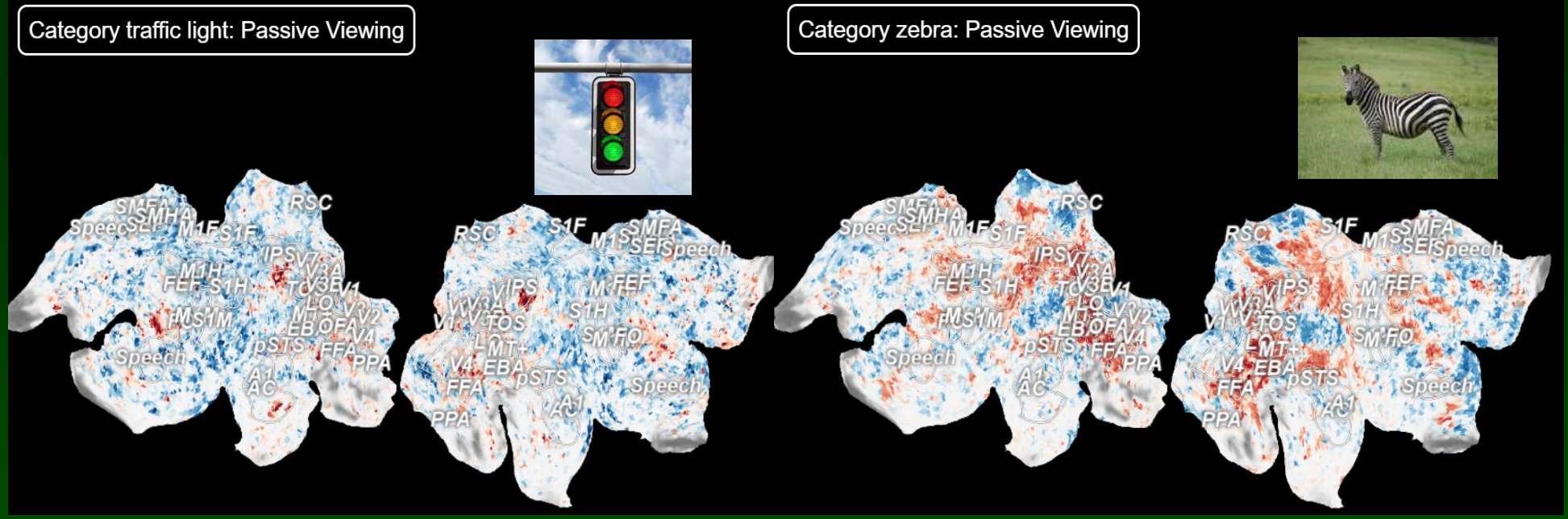
See videos at the @depth-first YT channel. Diffusion has roots in statistical physics.

Diffusion language models

Transformers work in sequential way. But LLMs may be based on diffusion models. Or IMMs?

- Nie S, … & Li, C. (2/25). *Large Language Diffusion Models*. arXiv:2502.09992
  LLaDA, a diffusion model demonstrates strong scalability, in-context learning.
- Inception.ai Mercury family of diffusion large language models (dLLMs) runs >1000 tokens/sec.
  Opens the way for costly deep reasoning that was taking a long time.
  Mercury Coder Mini is super-fast and has rank 1 on Leaderboard Copilot list.

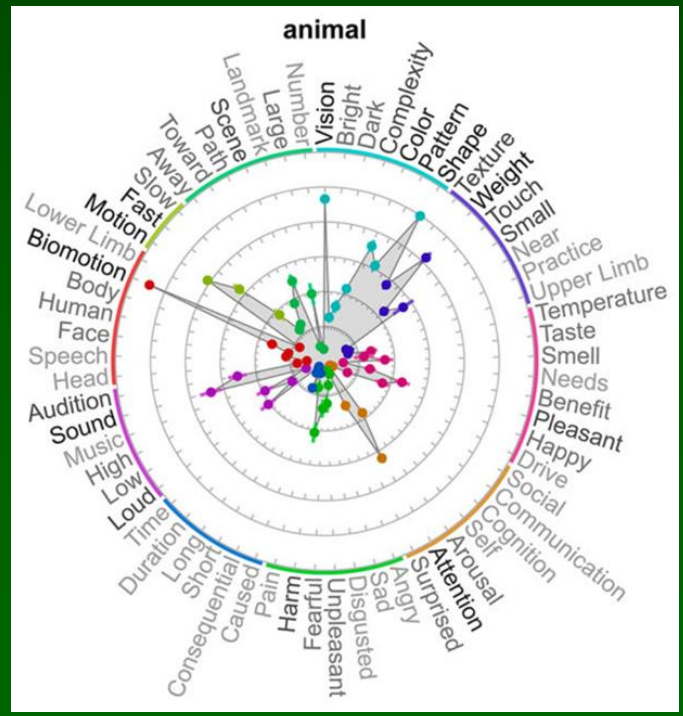




Watching traffic lights: V4: color perception. IPS: sensory-motor link, Frontal area (FEF, FO): planning behavior, action.

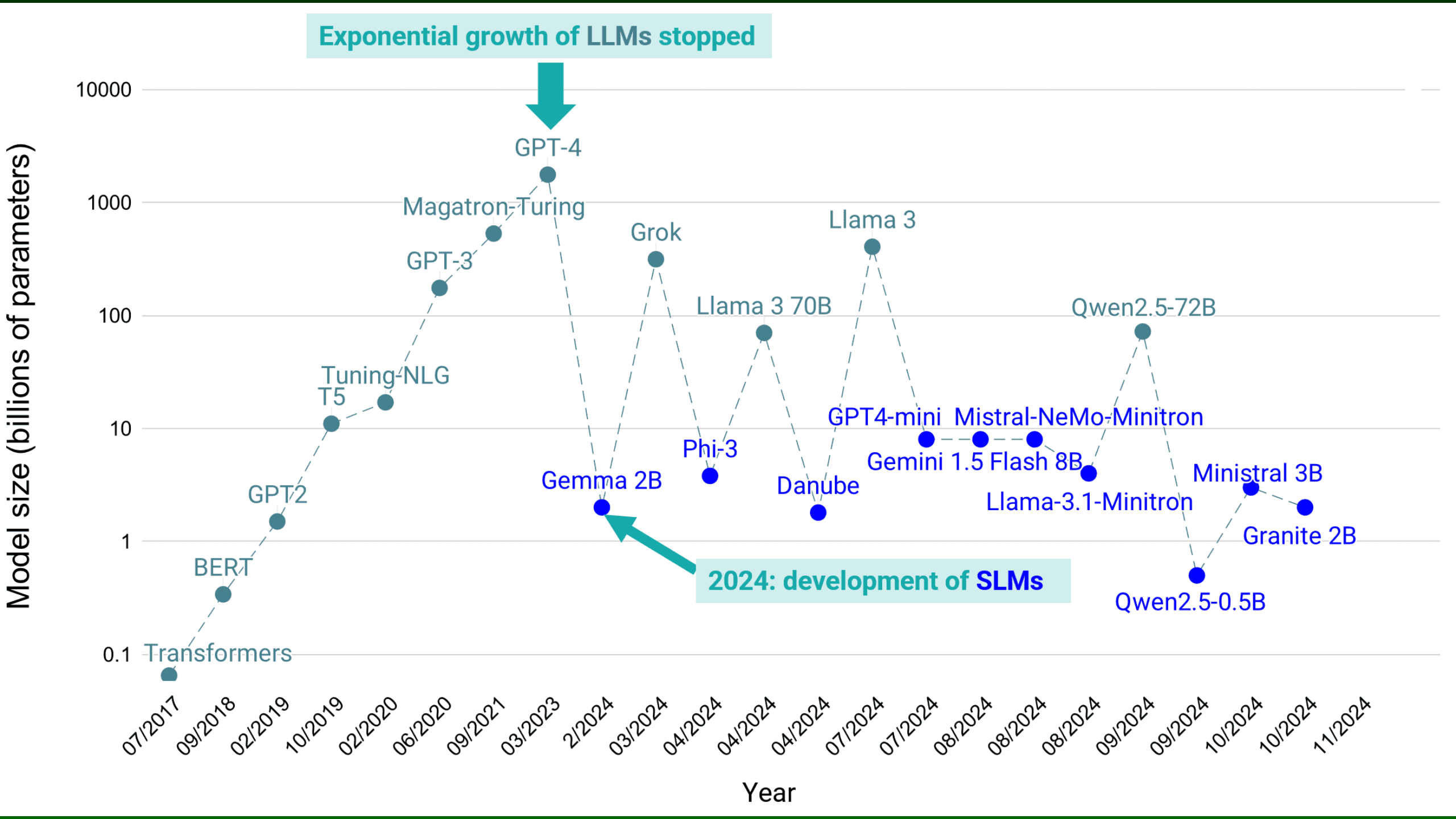
Idea: represent concepts as patterns corresponding to brain activation, or vectors in the feature space.

Can we simulate such maps of activity?

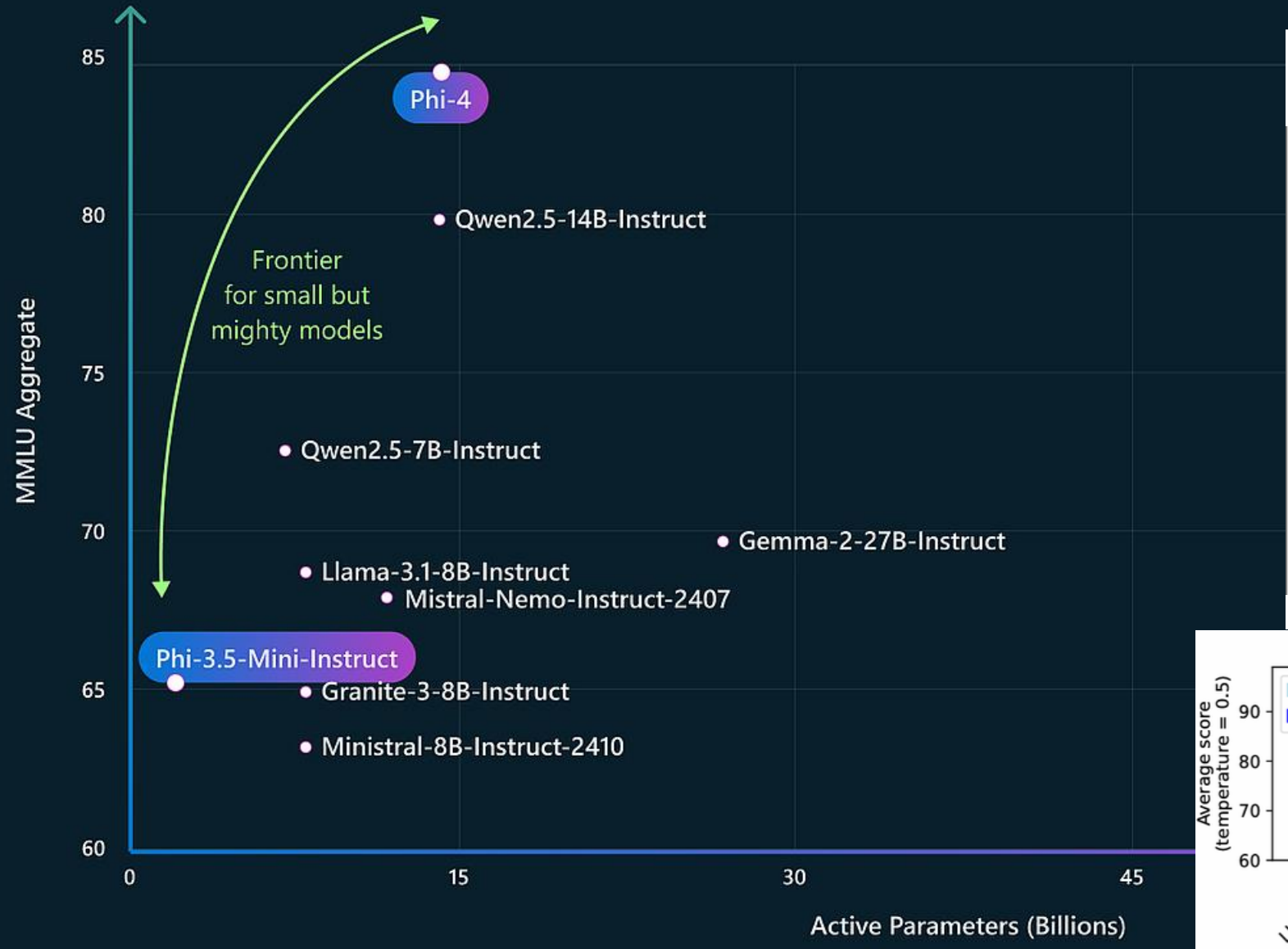


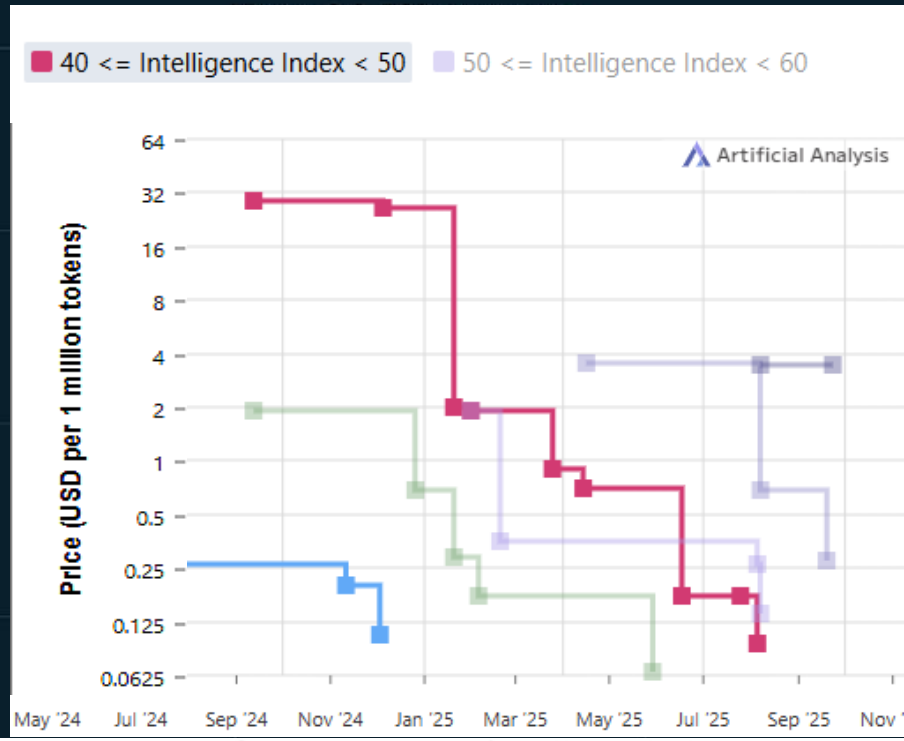
J.R. Binder et al, Toward a Brain-Based Componential Semantic Representation, 2016.

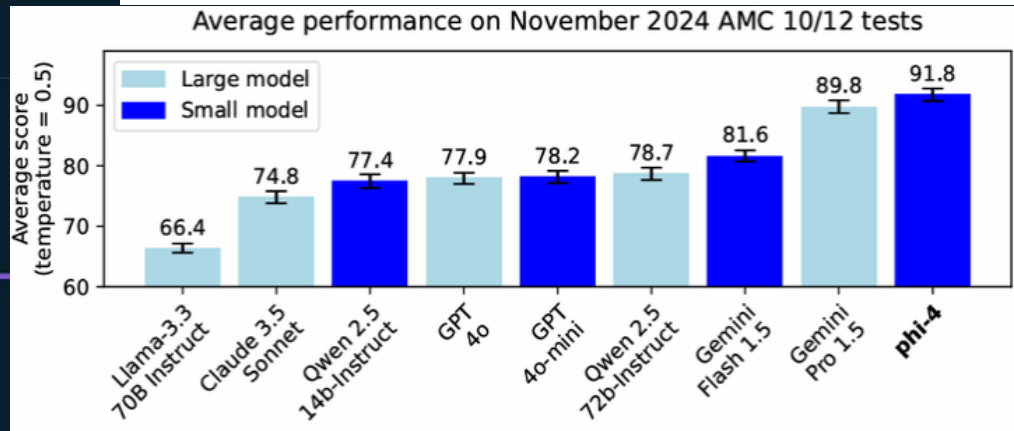
Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.


Exponential growth of LLMs stopped
Model size (billions of parameters)
10000
1000
100
10
1
0.1
Transformers
BERT
GPT2
Tuning-NLG
T5
GPT-3
Magatron-Turing
GPT-4
Gemma 2B
Grok
Phi-3
Llama 3 70B
Danube
Llama 3
GPT4-mini
Gemini 1.5 Flash 8B
Mistral-NeMo-Minitron
Llama-3.1-Minitron
Qwen2.5-72B
Qwen2.5-0.5B
Ministral 3B
Granite 2B
2024: development of SLMs
07/2017
09/2018
02/2019
10/2019
02/2020
06/2020
09/2021
03/2023
2/2024
03/2024
04/2024
04/2024
04/2024
07/2024
07/2024
08/2024
08/2024
08/2024
09/2024
09/2024
10/2024
10/2024
11/2024
Year

Phi-4 offers high quality results at a small size







Prompt optimization

GPT-4 with Medprompt achieves state-of-the-art performance on a wide range of medical challenge problems (Microsoft Research).

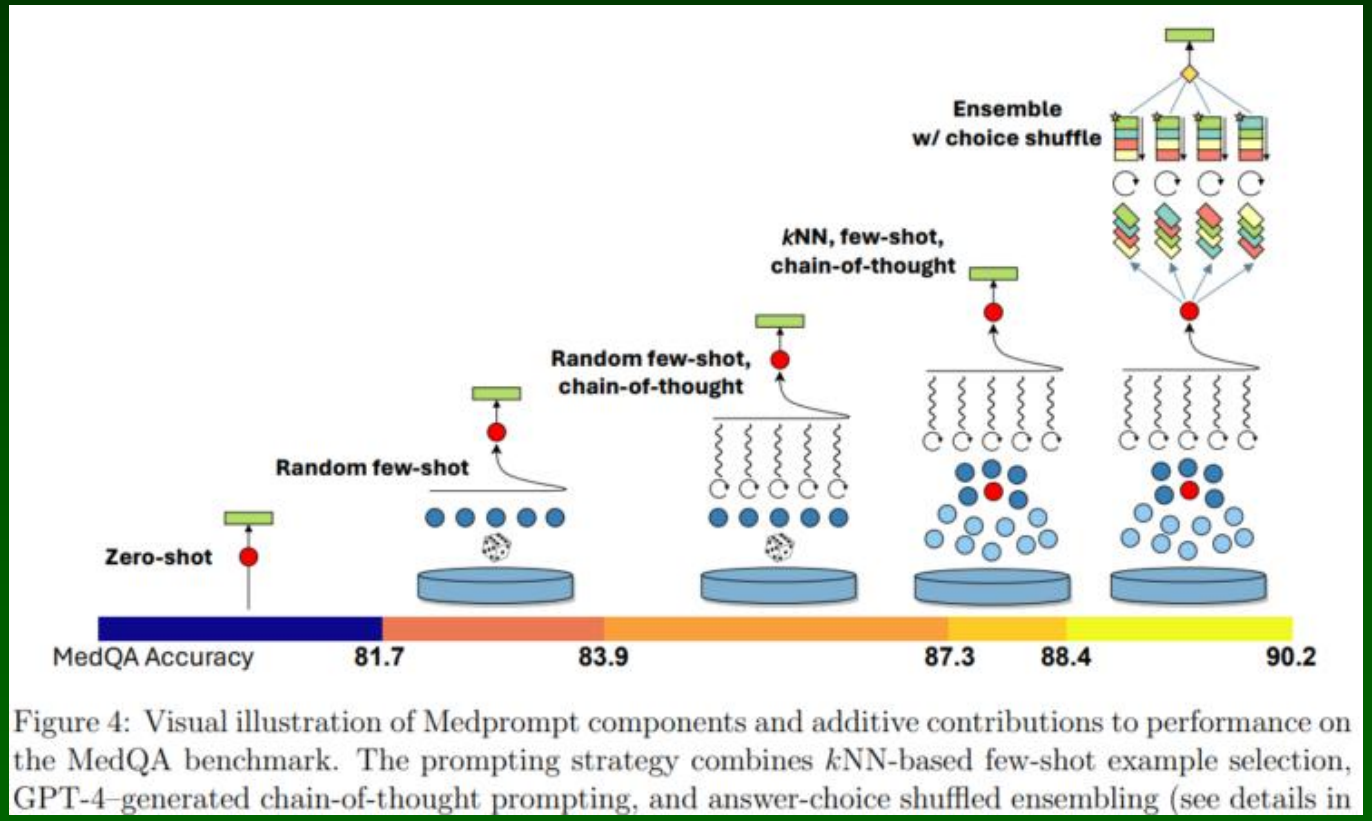
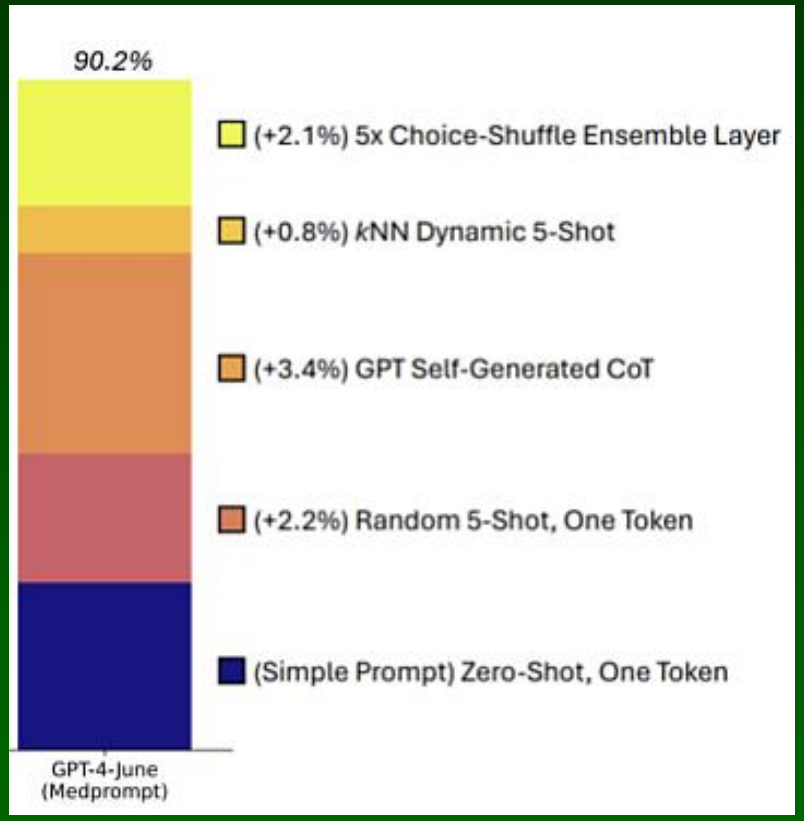


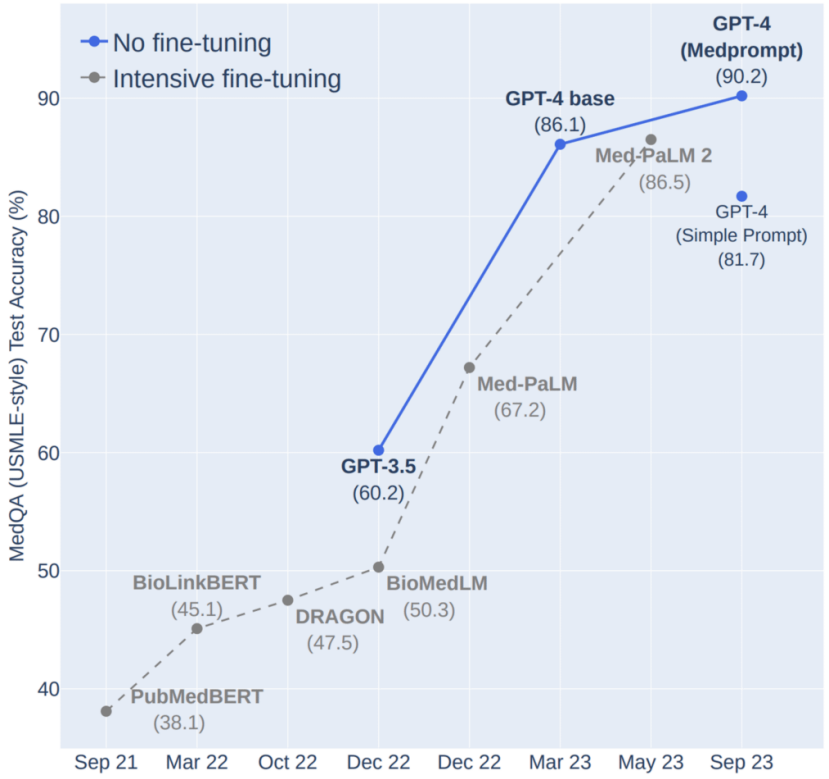
Figure 4: Visual illustration of Medprompt components and additive contributions to performance on the MedQA benchmark. The prompting strategy combines *k*NN-based few-shot example selection, GPT-4–generated chain-of-thought prompting, and answer-choice shuffled ensembling (see details in

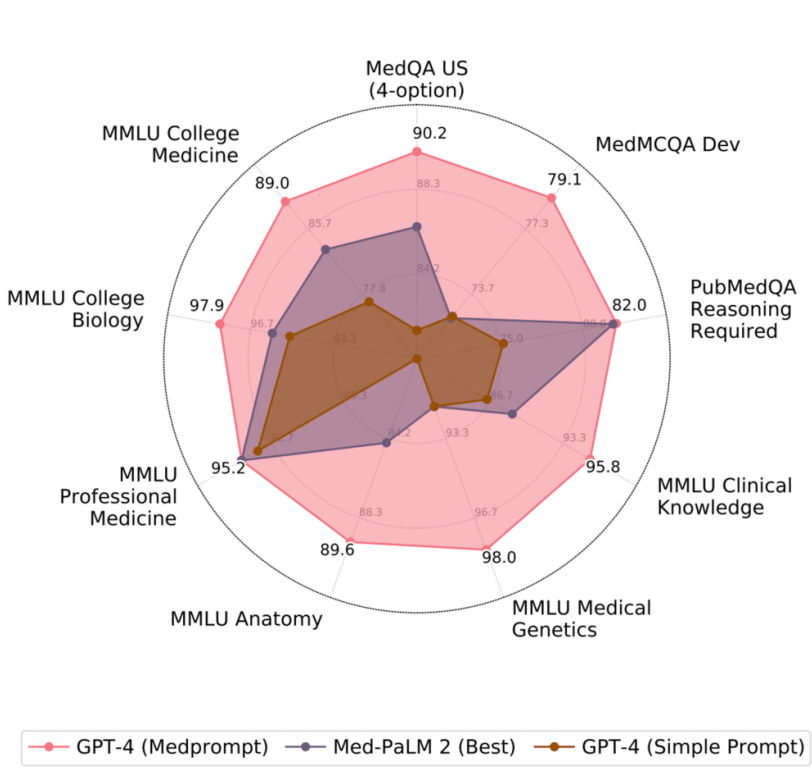


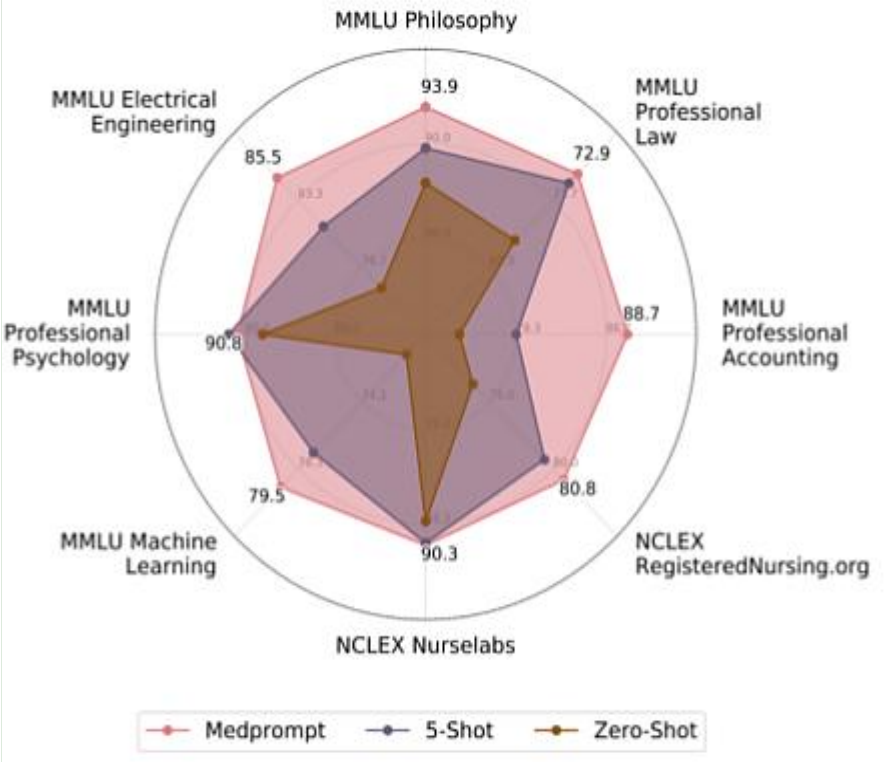
GPT finds associations: first thing that comes to our mind/network. Needs verification, reasoning.

Eliciting knowledge from LLMs

- MS: The Power of Prompting, PromptWizard feedback-driven self-evolving prompts, prompt actions in Microsoft Copilot Studio, Power Automate PL
- Optimize GPT-5, prompt cookbook. Prompt optimizers, and best practices.







MedQA

Jin, D et al. (2020). A Large-scale Open Domain Question Answering Dataset from Medical Exams. arXiv:2009.13081

MedQA, US Medical Licensing Exam (USMLE) style question benchmark with 12,723 questions. Passing score is 60%.

Human experts ~ 87% accuracy, GPT-4 ~ 90.2%,

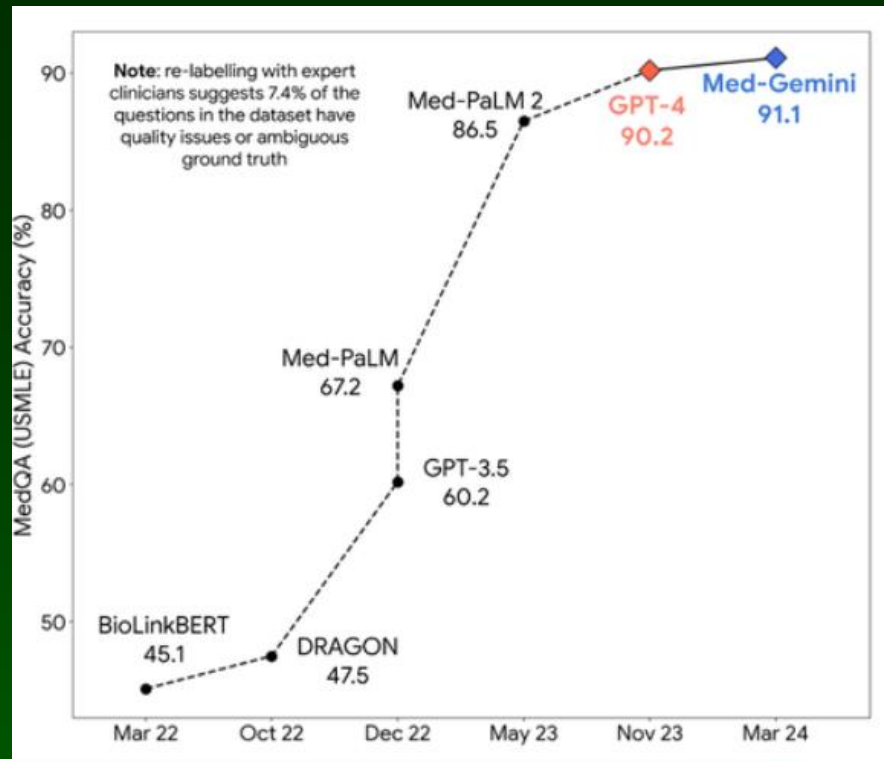
Med-Gemini achieves 91.1%, similar to GPT5 nano 91.4%, and GPT-5 achieves 95.84%.

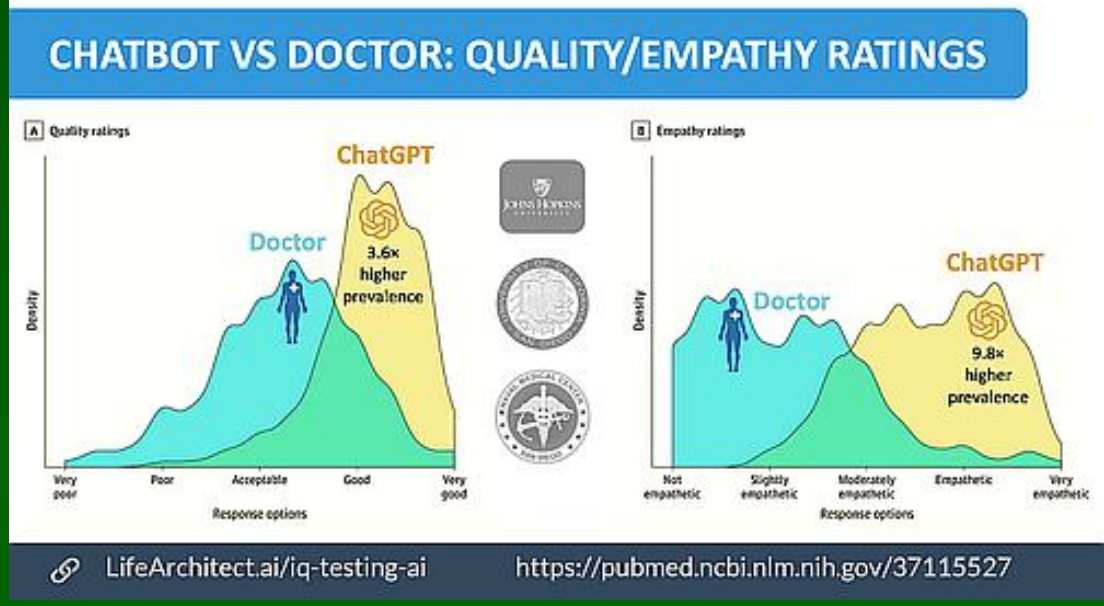
These systems have multimodal capabilities, can interpret complex 3D scans, answer clinical questions in pathology, dermatology, ophthalmology, genomics.

Goh, E. et al. (2024). *Large Language Model Influence on Diagnostic Reasoning*: A Randomized Clinical Trial. *JAMA Network Open*, *7*(10), e2440969–e2440969.

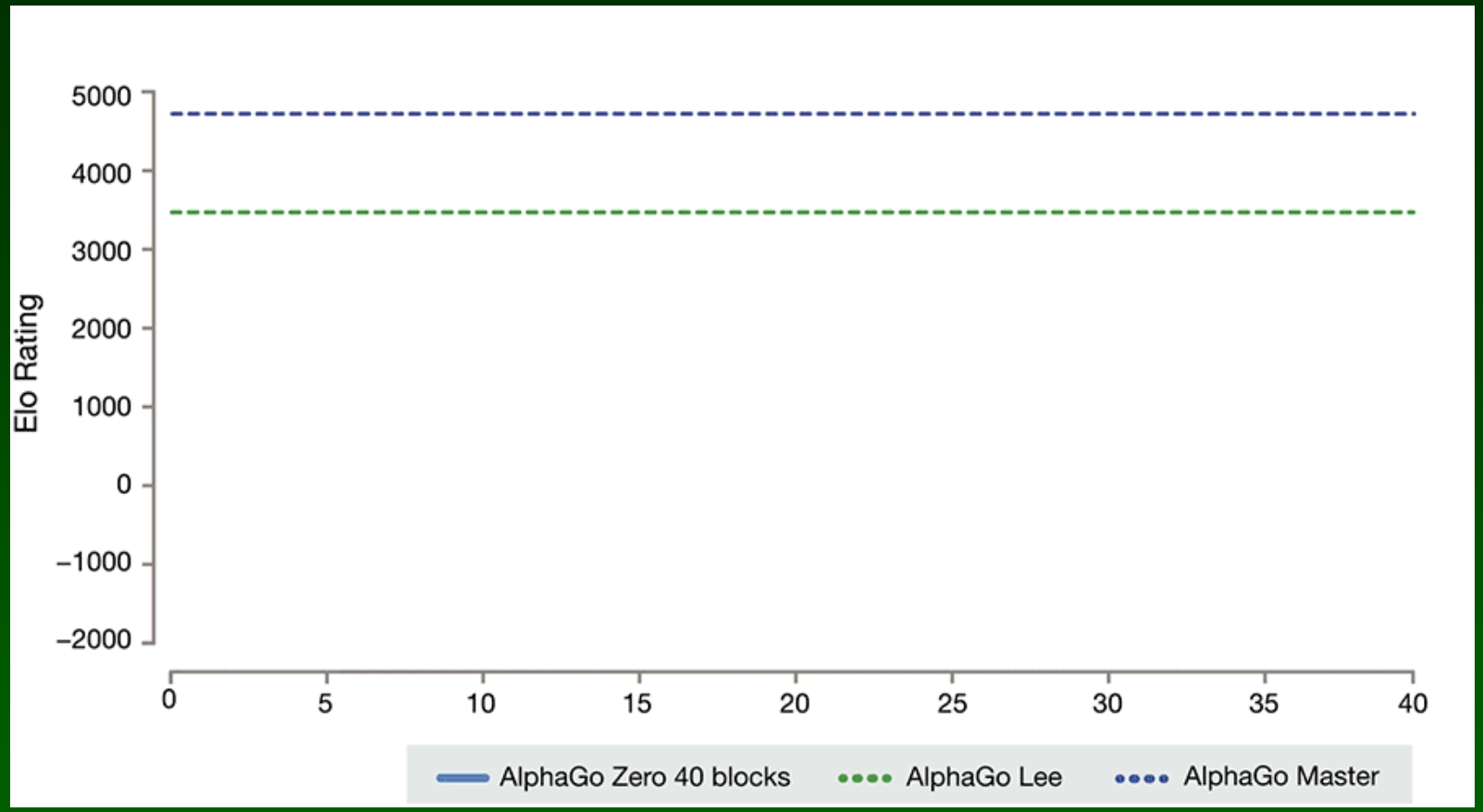
More empathy? Bots are nicer and have more time.

But GPT-5 is not specialized for all branches of medicine!

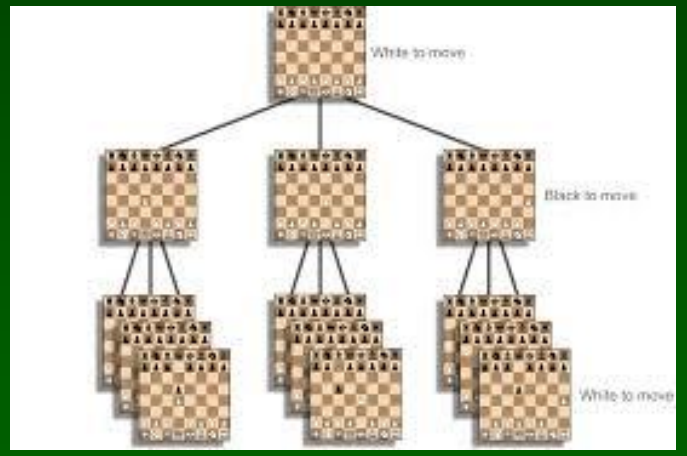




AlphaGo Zero: what is coming

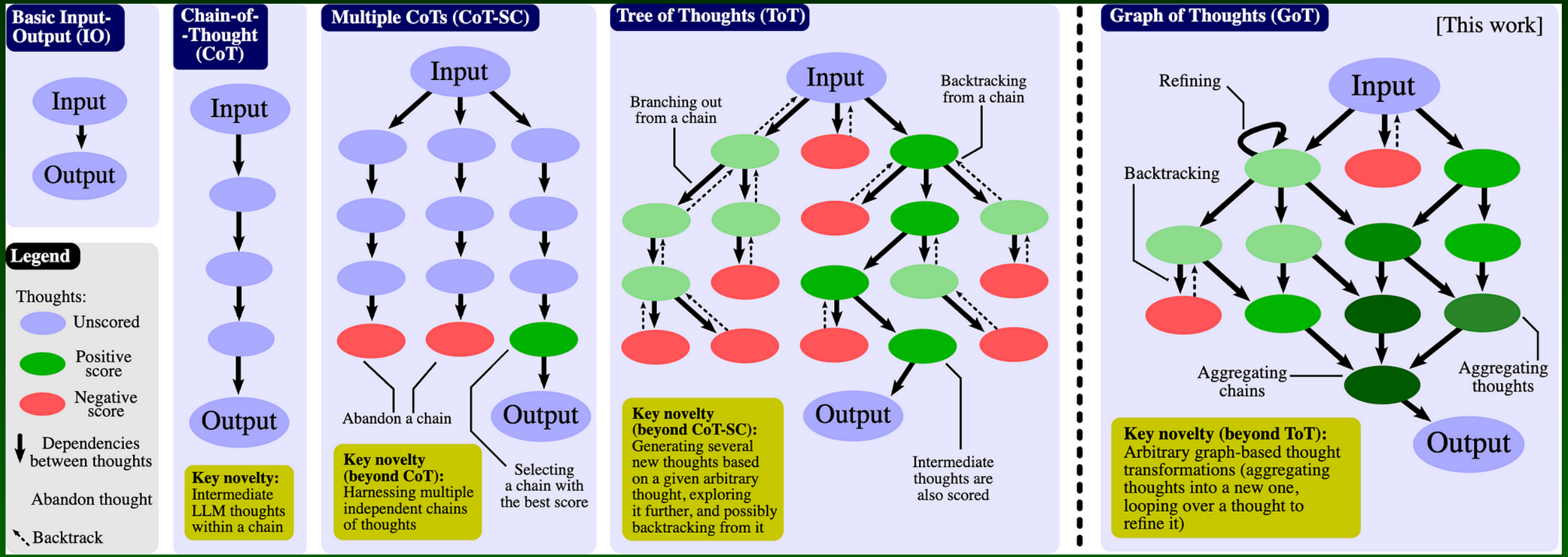


1997, Deep Blue-Kasparov.
2016, AlphaGo-Lee Sedol
2017, Alpha GoZero
2017, Poker, Dota 2;
2019, Starcraft II,
22 Stratego, Diplomacy



Superhuman level in Go. Best humans 3,600–3,860 ELO, KataGO more than 1400 point more!
MC Tree Search+deep NN pattern evaluation as heuristics + reinforcement learning, playing against itself.
Human knowledge becomes irrelevant, decreasing AI competence! Can we do it in other domains?
Shocking news: Ruoss … & Genewein, T. (2/24). *Grandmaster-Level Chess Without Search*.
270M parameter transformer model, 1-step intuitive decisions! Like in Blitz Chess.

Reasoning – cognitive inspirations



Associations: input => output.

This is GOFAI! LLM associations in heuristic search.

Chain of thought (CoT): step by step.

Tree of Thoughts (ToT): parallel thinking via RL.

Graph of Thoughts (GoT): like human reasoning.

Sketch-of-Thought: adaptive cognitive-Inspired sketching

AlphaGo Moment for Model Architecture Discovery

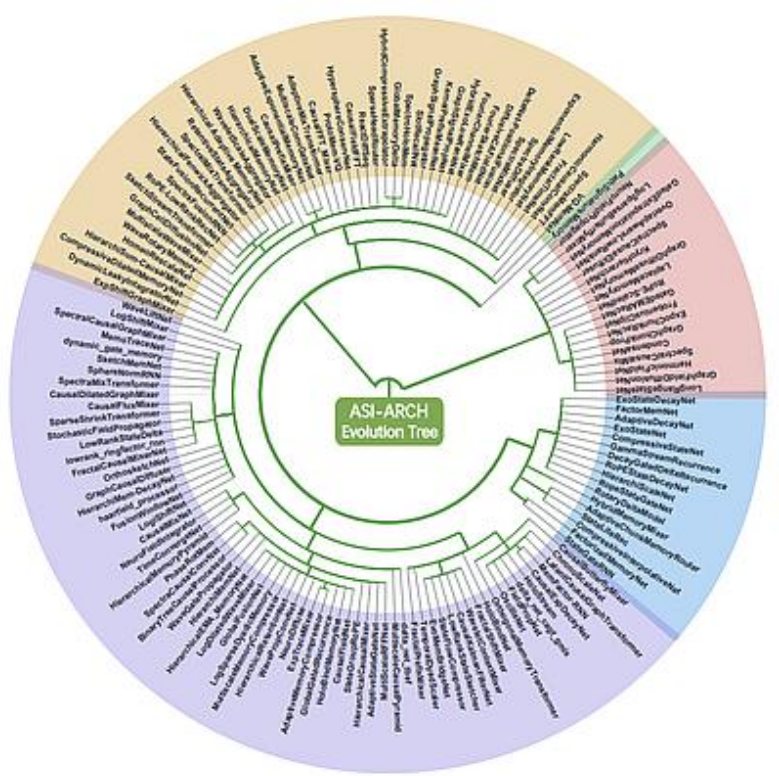
Liu, Y. … & Liu, P. (2025). AlphaGo Moment for Model Architecture Discovery. arXiv:2507.18074.

The pace of AI research itself remains linearly bounded by human cognitive capacity.

ASI-Arch is the first demonstration of Artificial Superintelligence for AI research (ASI4AI), a fully autonomous system enabling AI to conduct its own architectural innovation.

Conducts architecture discovery, autonomously hypothesizing novel architectural concepts, implementing them, training and empirically validating their performance through experimentation.

ASI-Arch conducted 1,773 autonomous experiments over 20,000 GPU hours, culminating in the discovery of 106 innovative, state-of-the-art (SOTA) linear attention architectures, demonstrating emergent design principles that systematically surpass human-designed baselines and illuminate previously unknown pathways for architectural innovation.



Jankowski N, Duch W, Grąbczewski K, Meta-learning in Computational Intelligence. Studies in Computational Intelligence, Vol. 358, Springer 2011.

Continuous Thought Machines

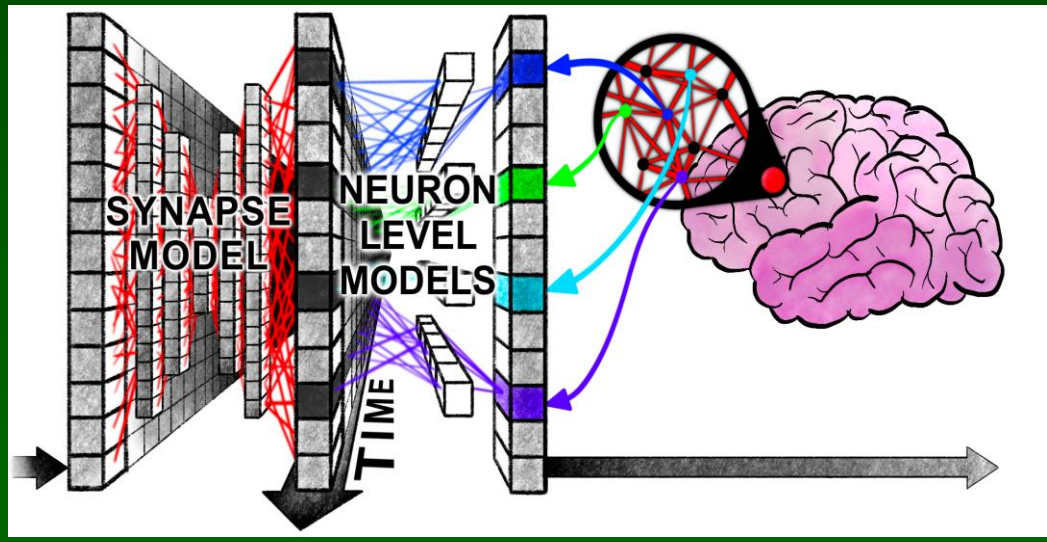
Continuous Thought Machine (CTM) is a model of neural dynamics that uses the synchronization of neural activity over time as its representation for taking actions in a world.

1) neuron-level temporal processing: each neuron uses unique weight (synaptic) parameters to process a history of incoming signals; and
2) neural synchronization (attractors) employed as a latent representation.

- Darlow L, Regan C, Risi S, Seely J, Jones L (5/2025). *Continuous Thought Machines.* arXiv:2505.05522
- SakanaAI/continuous-thought-machines. (2025). Github SakanaAI

Performs tasks that require complex sequential reasoning, mazes, image classification at human level.

Stops early for simple tasks, works longer for hard cases.

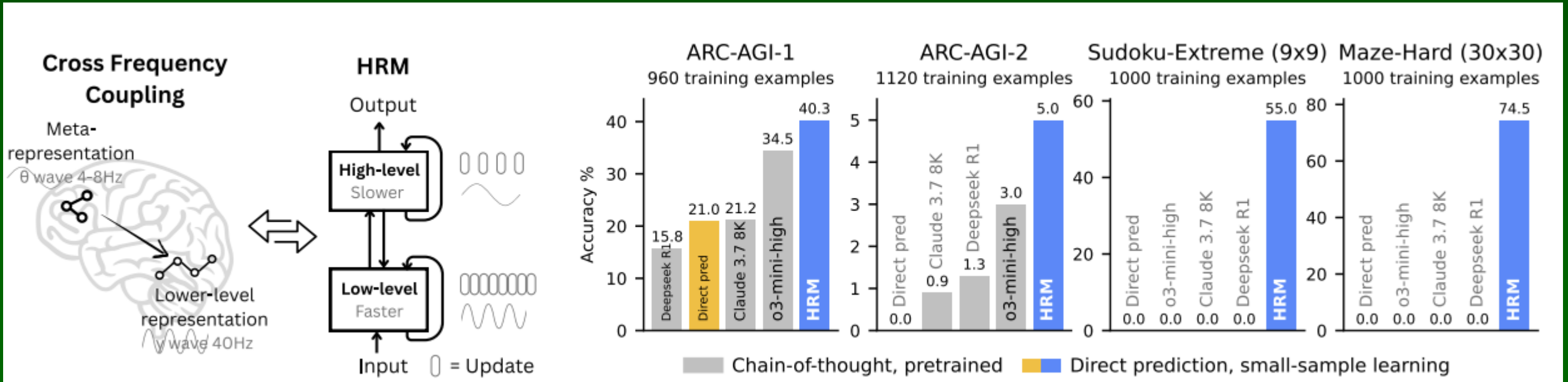


Hierarchical Reasoning Model (HRM)

HRM is a recurrent architecture inspired by the hierarchical and multi-timescale processing in the brain. Executes sequential reasoning in a single forward pass without explicit supervision.
A high-level module responsible for abstract planning, and a low-level module handling details.

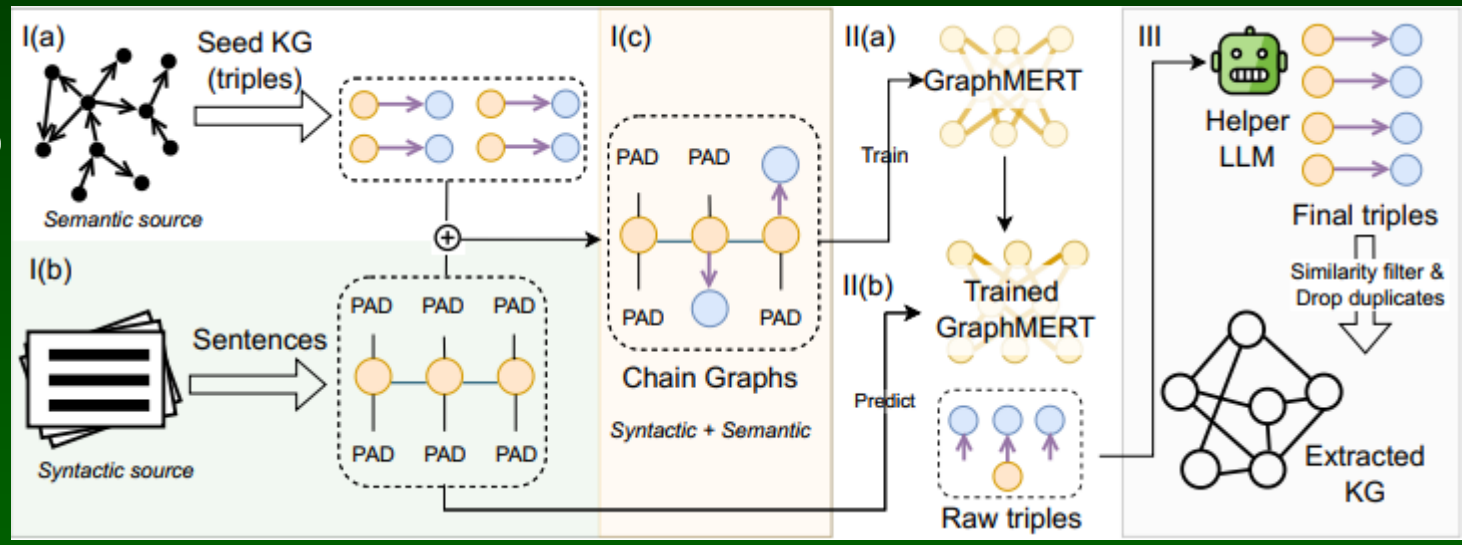
With 27 mln parameters, without pre-training or CoT data, using only 1000 training samples, nearly perfect performance is achieved on complex Sudoku puzzles and optimal path finding in large mazes. HRM outperforms much larger models with significantly longer context windows on the Abstraction and Reasoning Corpus (ARC), a key benchmark for measuring AGI capabilities.

Wang, G. et al. (2025). *Hierarchical Reasoning Model.* arXiv:2506.21734.



Neurosymbolic GraphMERT

GraphMERT, 80M-parameter graphical encoder-only model distills high-quality KGs from unstructured text corpora and its own internal representations, forming a modular neurosymbolic stack: neural learning of abstractions, symbolic KGs for verifiable reasoning. GraphMERT + KG is the first efficient and scalable neurosymbolic model to achieve state-of-the-art benchmark accuracy along with superior symbolic representations relative to baselines.
Recovers UMLS relations more accurately than large LLMs. KG integration is a key step toward domain-specific superintelligence.



Belova, M., Xiao, J., Tuli, S., & Jha, N. K. (2025). *GraphMERT: Efficient and Scalable Distillation of Reliable Knowledge Graphs from Unstructured Data.* arXiv.2510.09580 YouTube summary

Medical ontologies (ULMS) enable development of a large-scale practical algorithm to re-create pathways of spreading neural activations. Concepts of specific semantic type are identified in the text, and then all related concepts of the same type are added to the text, providing expanded representations.

Duch W, Matykiewicz P, and Pestian J, (2008) Neurolinguistic Approach to Natural Language Processing with Applications to Medical Text Analysis. Neural Networks 21(10), 1500-1510. Idea: each token generates synthetic context data, best are selected! Processing clinical text with domain-specific spreading activation methods.US Patent 8,930,178 B2. More here.

Latent reasoning

People can subconsciously generate a sequence of mental states without explicit verbalization.

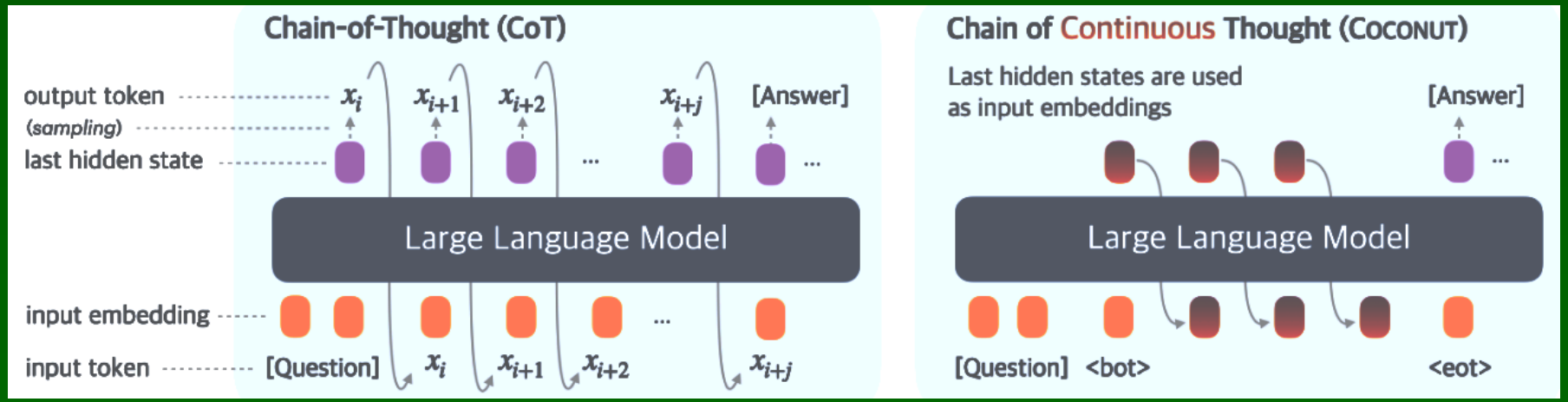
Unconscious thought theory, Ap Dijksterhuis (2004). Unconscious thinking about a complex problem can lead to better solutions than conscious deliberation.

Coconut (Chain of Continuous Thought) conducts reasoning in the latent (hidden) space.
Advanced reasoning patterns emerge, with many alternatives.

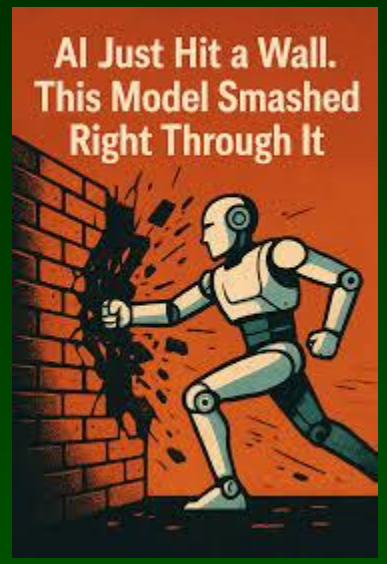
- S. Hao et al., "Training LLMs to Reason in a Continuous Latent Space" (12/2024)
- X. Shenet al. Efficient Reasoning with Hidden Thinking. arXiv:2501.19201 Heima model
- Zelikman, E. i inn. (2024). Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking. arXiv:2403.09629
- Feng, S., Fang, G., Ma, X., & Wang, X. (2025). Efficient Reasoning Models: A Survey. arXiv:2504.10903

Github papers on efficient reasoning.

Faster 4-24 times.



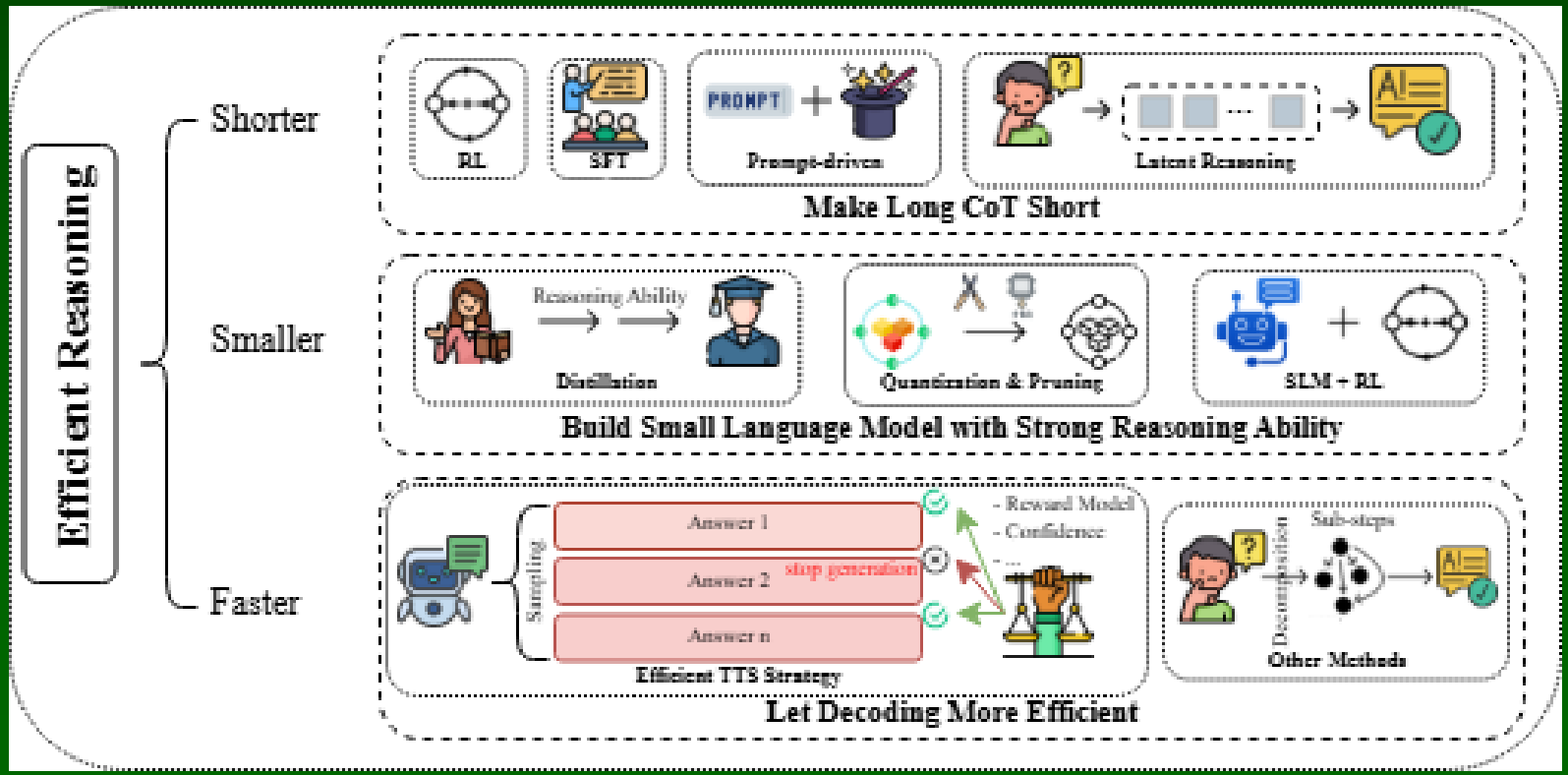
Efficient/latent reasoning



Feng S, Fang G, Ma X, & Wang X.

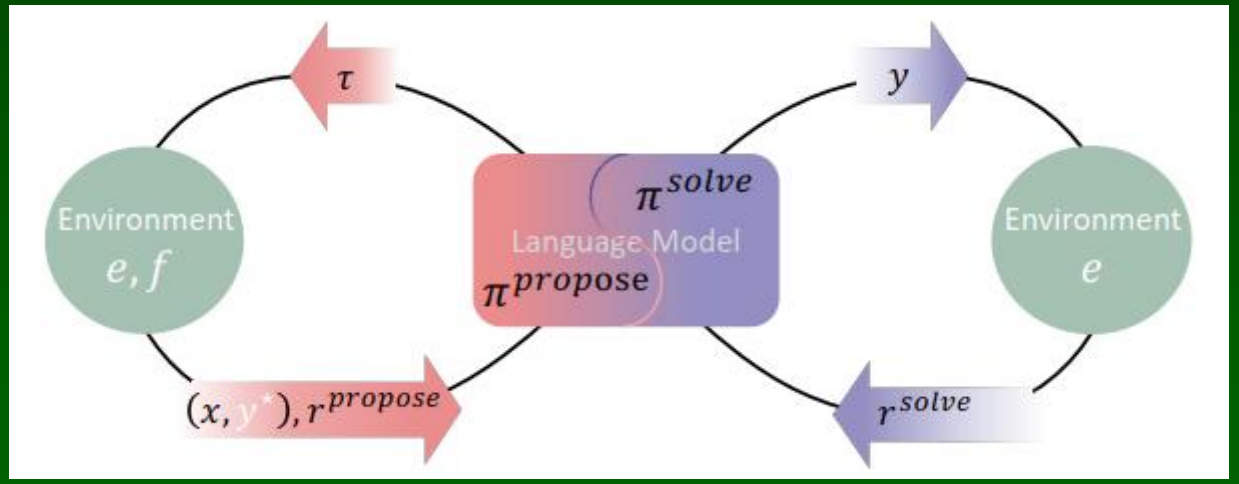
Efficient Reasoning Models: A Survey. arXiv:2504.10903

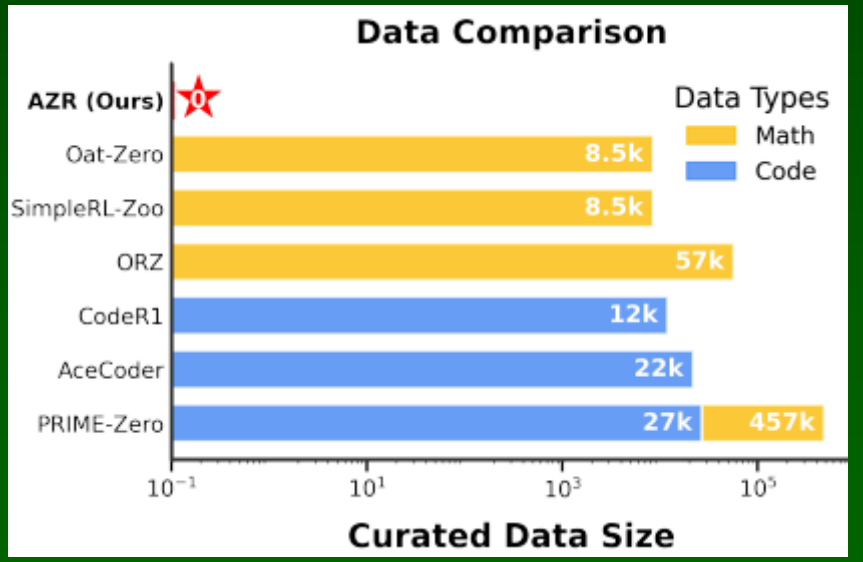
Short chains of thoughts, small models, fast decoding.



Self-play Reasoning

Can self-play be used to reach superhuman levels in complex fields? Verifiable rewards (RLVR) are needed. Absolute Zero Reasoner (AZR) self-evolves its training curriculum and reasoning ability by using a code executor to both validate proposed code reasoning tasks and verify answers, serving as an unified source of verifiable reward to guide open-ended yet grounded learning.
Despite being trained entirely without external data, AZR achieves results on coding and mathematical reasoning outperforming models that rely on tens of thousands of in-domain human-curated examples.





Zhao, A. et al. (2025). *Absolute Zero: Reinforced Self-play Reasoning with Zero Data* arXiv.2505.03335

Self-improving Darwin Gödel Machines

The Darwin Gödel Machine: AI that improves itself by rewriting its own code (5/2025)
https://sakana.ai/dgm/

DGM iteratively builds a growing archive of agents by harnessing principles of open-ended exploration. New agents are created and scored by interleaving self-modification with downstream task evaluation.

Zweiger, A… & Agrawal, P. (6/2025). Self-Adapting Language Models

Self-Adapting LLMs (SEAL), a framework that enables LLMs to self-adapt by generating their own finetuning data and update directives, may restructure the information, optimize its hyperparameters, or invoke tools for data augmentation and gradient-based updates.

Wang, T … & Li, X. (8/2024). *Self-Taught Evaluators*.

STE use synthetic training data only, starting from unlabeled examples. Iterative self-improvement scheme generates contrasting model outputs and trains an LLM-as-a-Judge to produce reasoning traces and final judgments, using the improved predictions at each new iteration. STE improved Llama3-70B-Instruct matched top-performing reward models with human-labeled examples.

Zhou Y, Levine S, Weston J, Li X, & Sukhbaatar S (6/2025). Self-Challenging Language Model Agents.

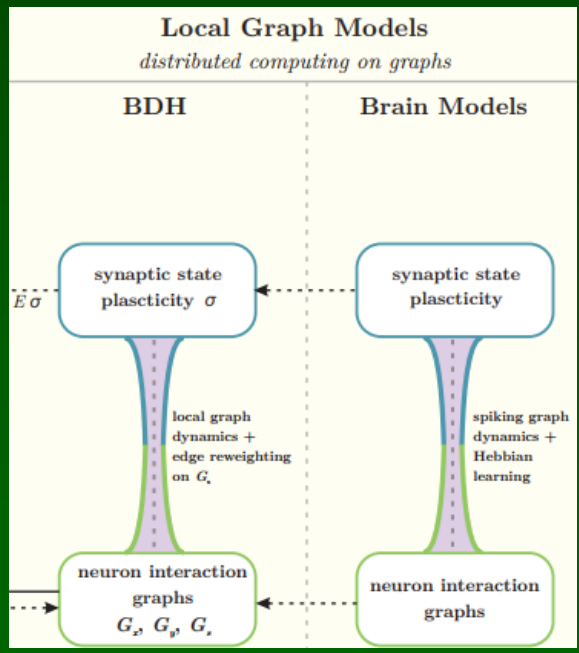
Self-Challenging framework for training an agent on high-quality tasks that are generated by itself.
It generates a task forming a novel general class of problems defined by an instruction, a verification function, and solution and failure cases which serve as tests, filtering the high-quality tasks.

Dragon Hatchling

"Dragon Hatchling" (BDH), state-space LLM architecture based on a scale-free biologically inspired network of *n* locally-interacting neuron particles forming a graph of high modularity. BDH performs attention-based state space sequence learning changing graph topology.

BDH can be represented as a brain model and rivals GPT2 performance on language and translation tasks, at the same number of parameters (10M to 1B), for the same training data. During inference it relies on synaptic plasticity with Hebbian learning using spiking neurons. Reasoning process leads to learning.

Very speculative as a model that is more "brain-like", and not yet tested on larger problems. Reweighting ≈ Chunking during inference in SOAR architecture.



Kosowski, A., Uznański, P., Chorowski, J., Stamirowska, Z., & Bartoszkiewicz, M. (2025). *The Dragon Hatchling: The Missing Link between the Transformer and Models of the Brain.* arXiv.2509.26507 Pathway.com

A hierarchy of intelligent systems with growing complexity/flexibility, from networks with simple neurons to complex interacting agents, was presented in:

Duch W, Mandziuk J (2004), Quo Vadis Computational Intelligence?
Advances in Fuzzy Systems 21, 3-28

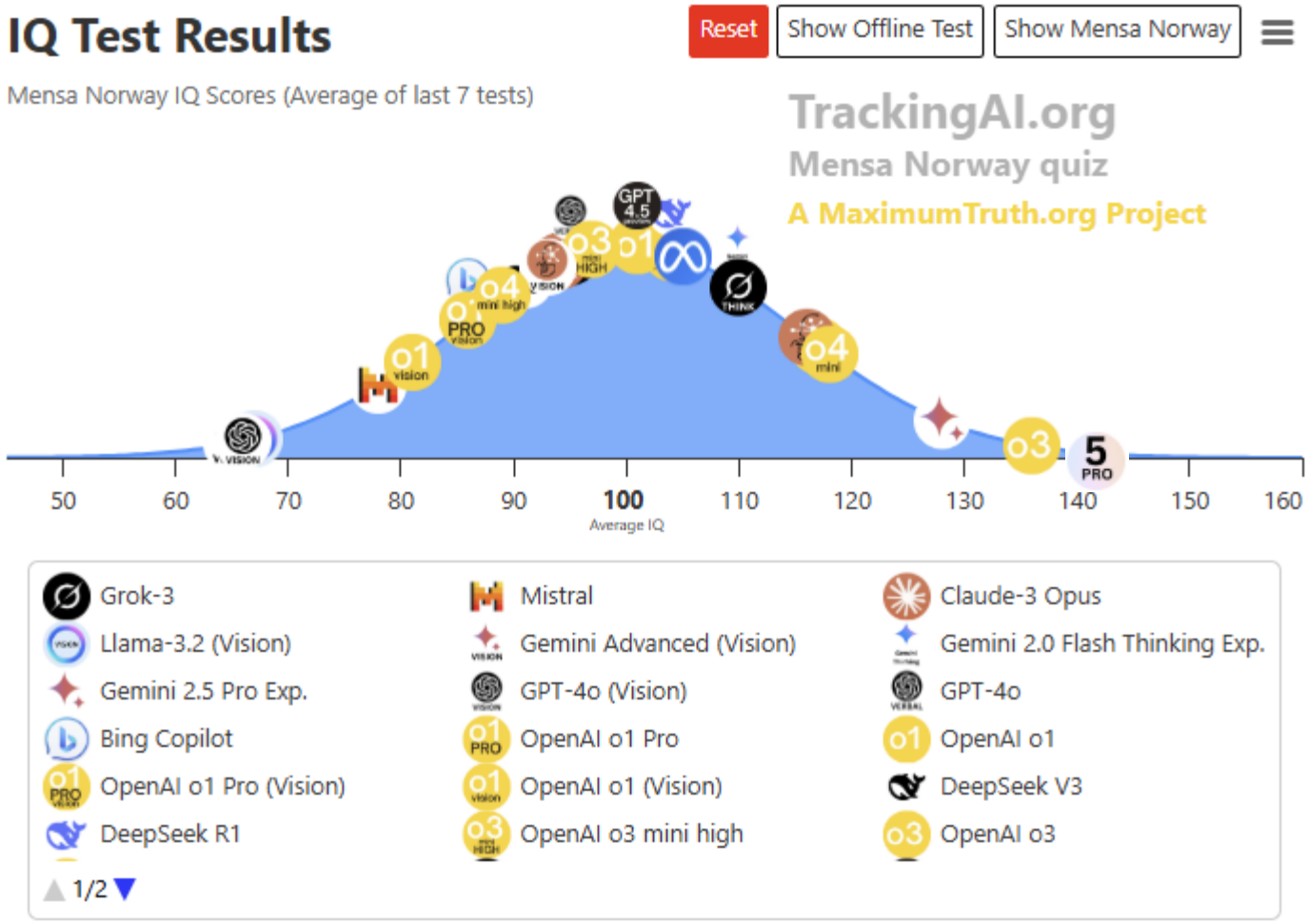
State-of-the-Art

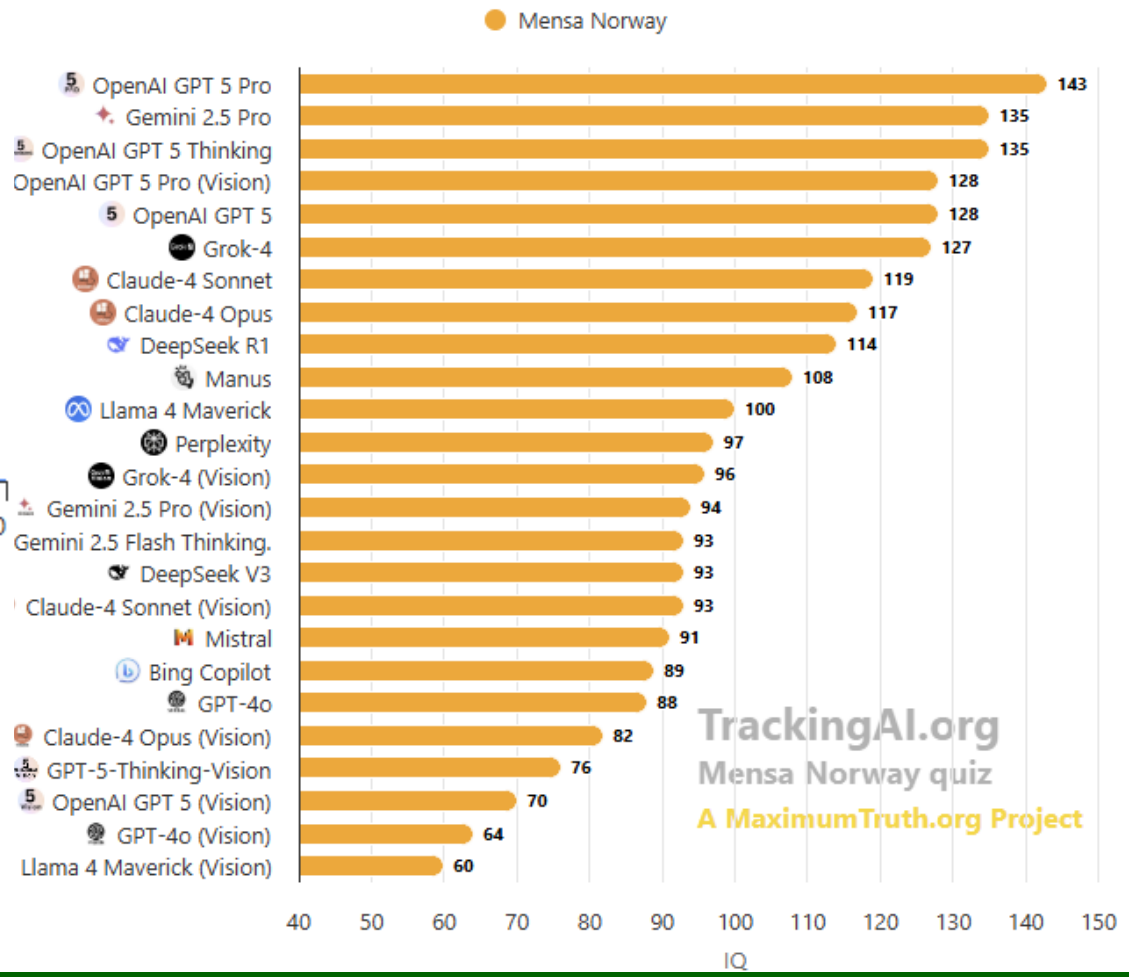
Turing Test

Q: You passed a Turing test, selected as human far more often than an actual human.
A: It shows that AI is crossing a critical threshold. It implies a much deeper understanding of language, context, and reasoning.

2023: GPT3 writes answers to 10 philosophical questions that even experts could not distinguish from those written by Daniel Dennet, most famous philosopher of mind.

MIQ of AI





At the beginning of the 2024 highest results were < 100 points.
29.09.2025: Claude Sonnet 4.5 shows 30+ hours of autonomous coding.

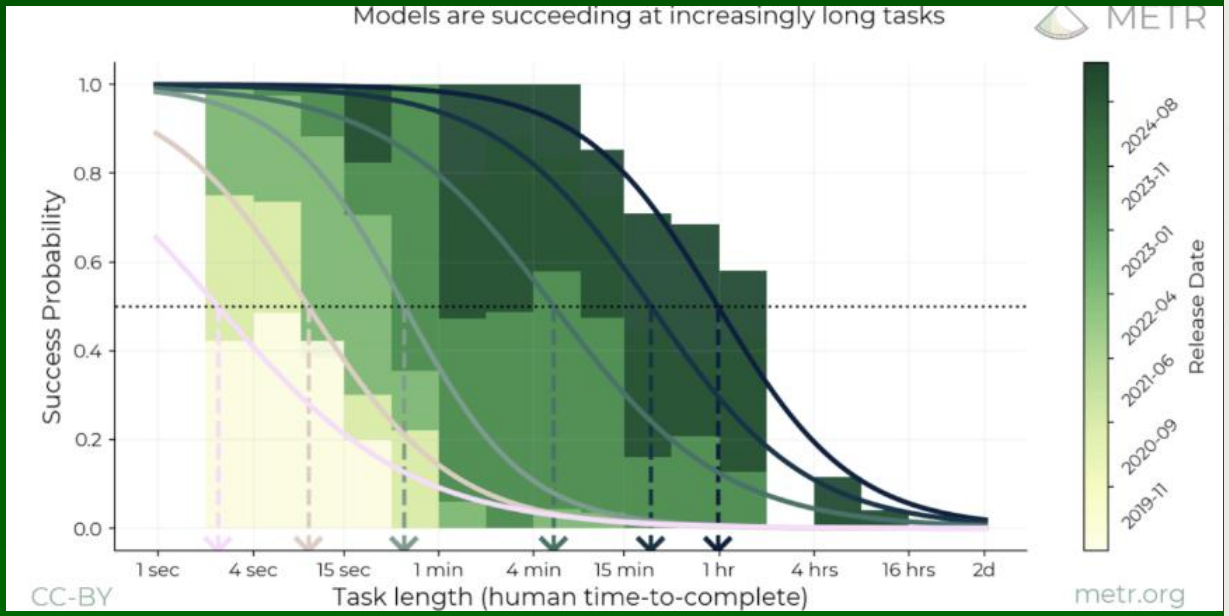
Long tasks

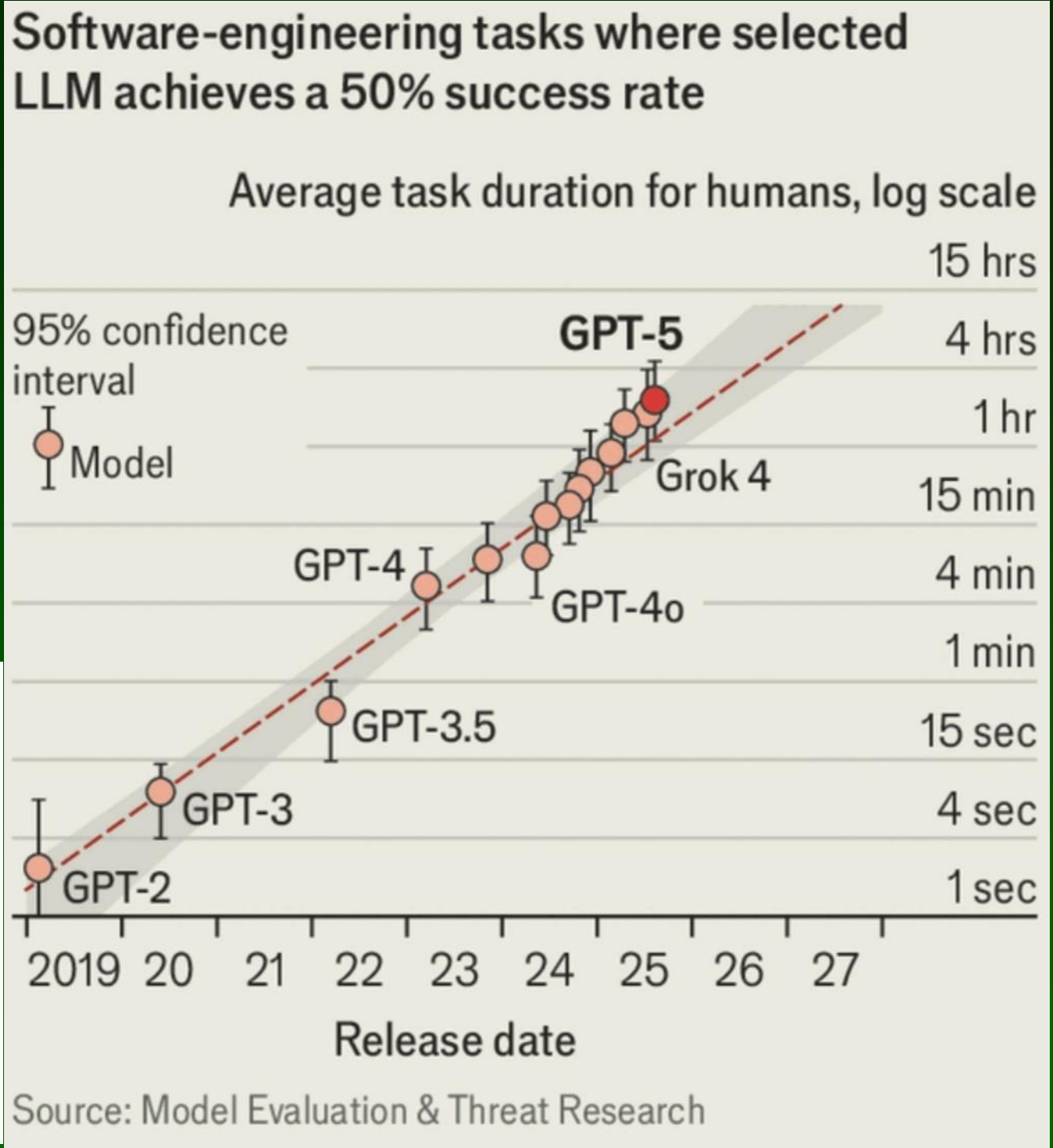
Measuring AI Ability to Complete Long Tasks.

AI performance in terms of the length of tasks AI agents can complete. GPT5: 2h15m-4.5h.

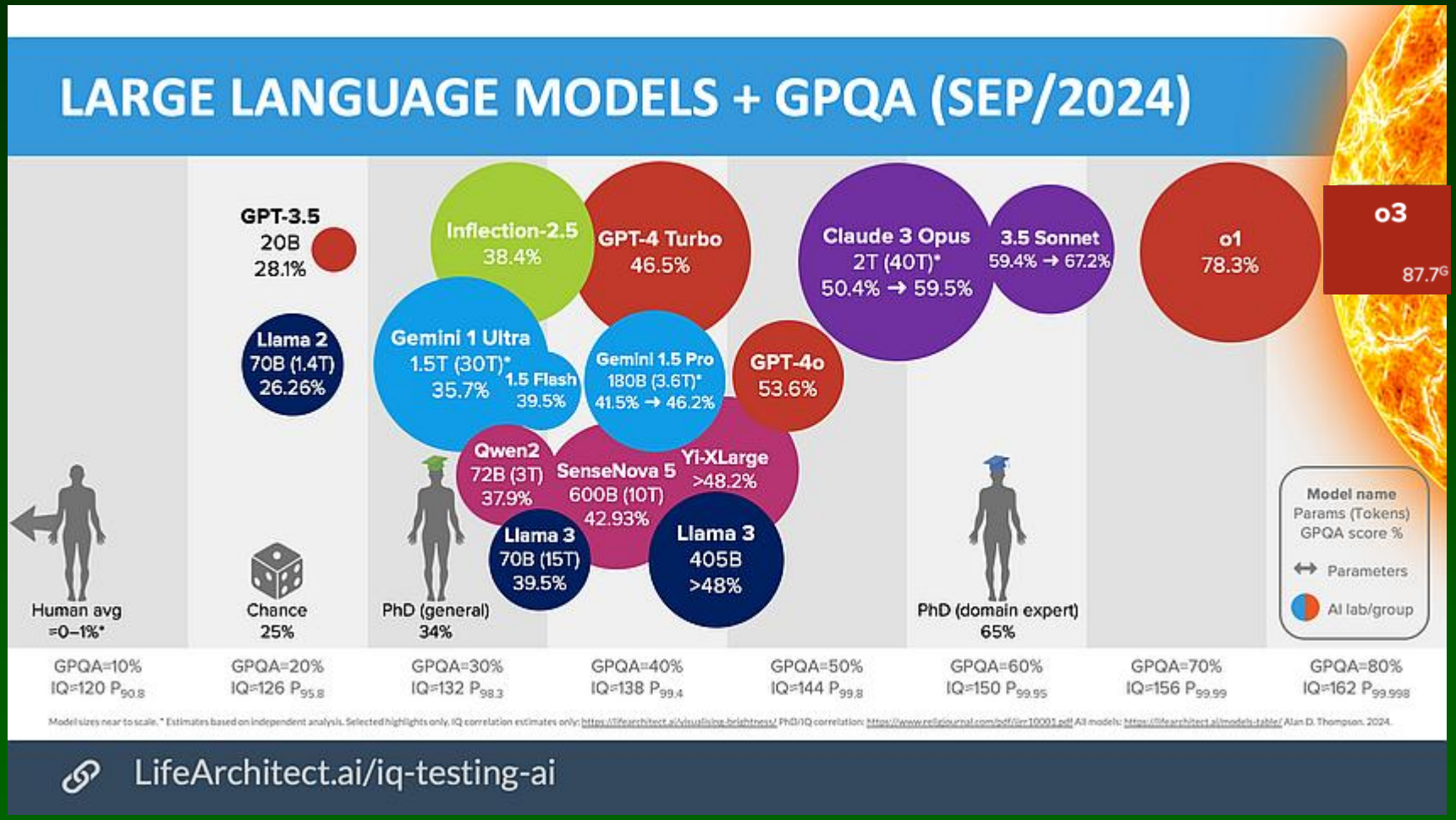
This metric has been consistently exponentially increasing over the past 6 years, with a doubling time of around 7 months.

Extrapolating, in less than a decade AI agents should independently complete a large fraction of tasks that take humans days or weeks.





GPQA benchmark



Grok 4 88.1
GPT-5 85.6%

vals.ai

Rein, D et al. (2023). *GPQA: A Graduate-Level Google-Proof Q&A Benchmark.* arXiv:2311.122
448 extremely difficult questions: highly skilled non-expert validators reach 34% accuracy with >30 minutes of web access, experts pursuing PhDs in the corresponding domains reach 65%-74%.

International Collegiate Programming Contest (ICPC)

5.09.2025: teams from 139 universities in 103 countries participated in the ICPC finals.

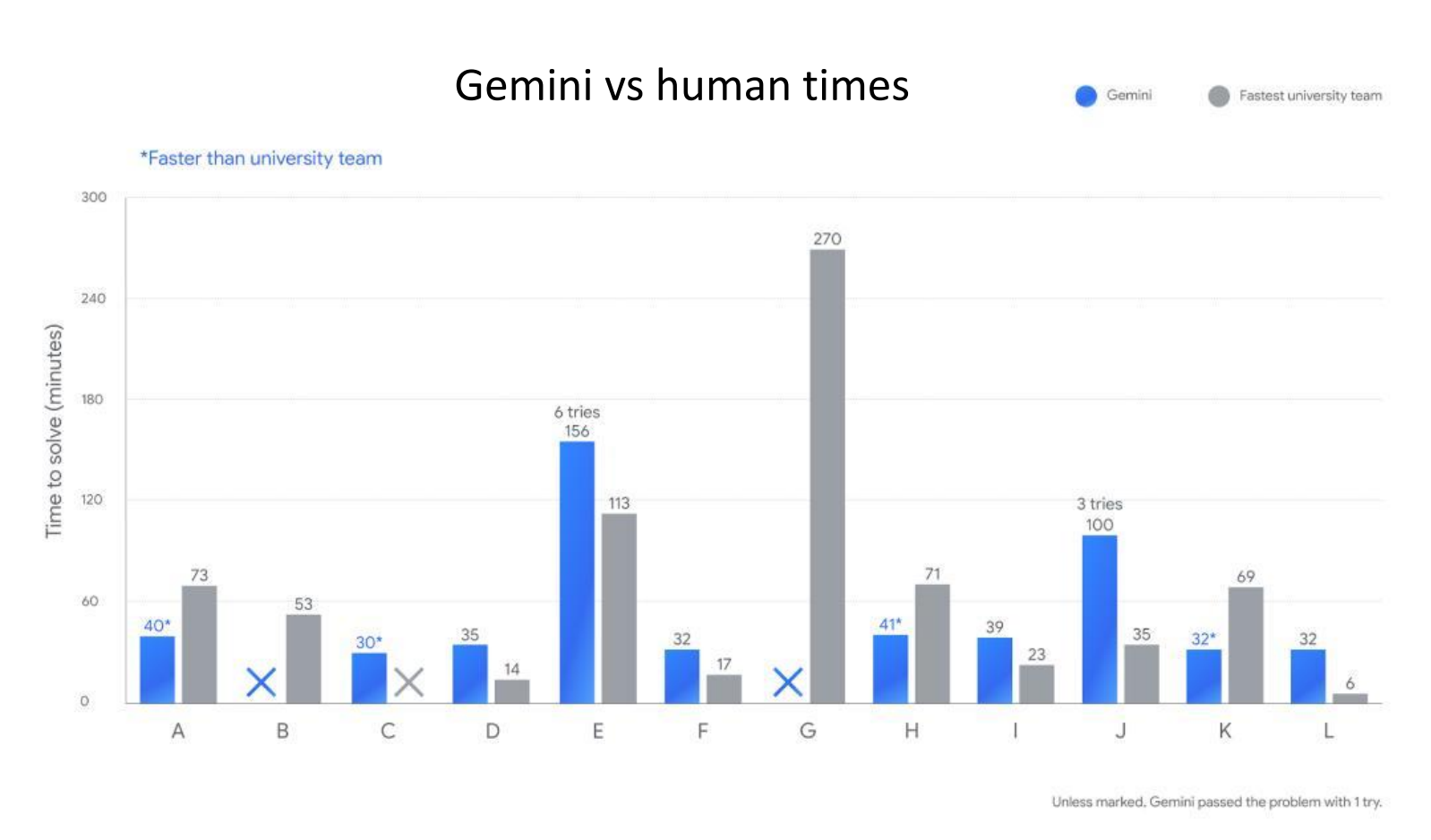
Team: 3 students. Task: solve 12 algorithmic problems within 5 hours. Cutting-edge algorithms: graph theory, number theory, dynamic programming, combinatorial optimization, and network flow.

Ranking depends on the number of problems solved and the time taken.

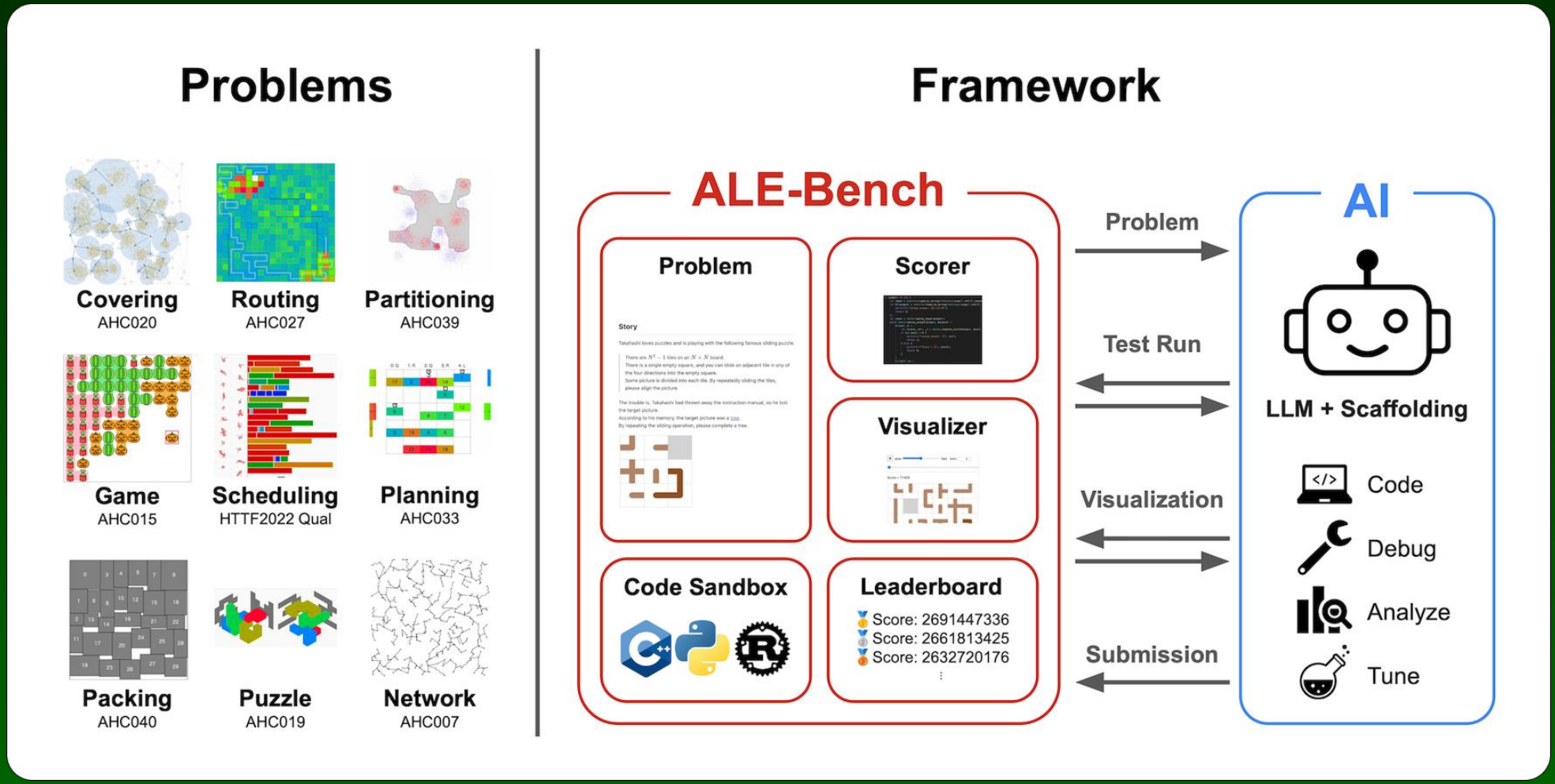
3 human teams, from Saint Petersburg State University, the University of Tokyo, Beijing Jiaotong, solved 11 problems.

GPT-5 solved 11 problems solved on the first try, one problem on 9th submission, for a perfect score of 12/12.

Gemini 2.5 Deep Think solved problem that no human team could solve.



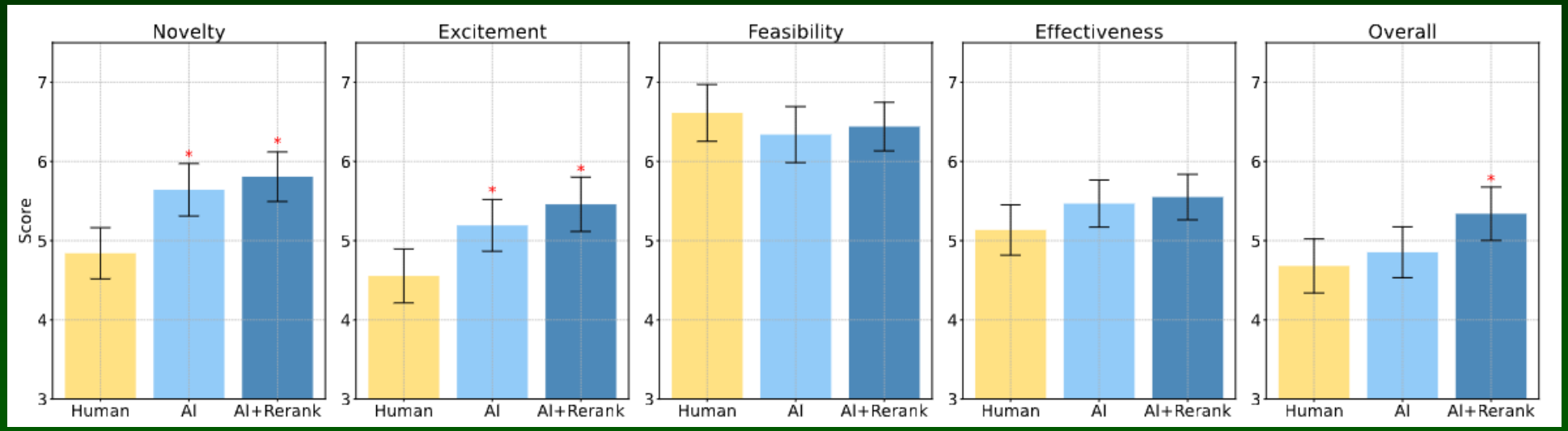
ALE-Bench and ALE-Agent



ALE-Agent was 21st out of 1,000 human participants in a live AtCoder Heuristic Competition (AHC), in AI discovery of solutions to hard optimization problems with important real world applications.

Generating novel ideas

Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. 6.09.24. Tests with Anthropic Claude-3.5-Sonnet (200.000 tokens).



AI is significantly better than human experts at inventing novel ideas!

Human re-rank is a selection of AI ideas by human.

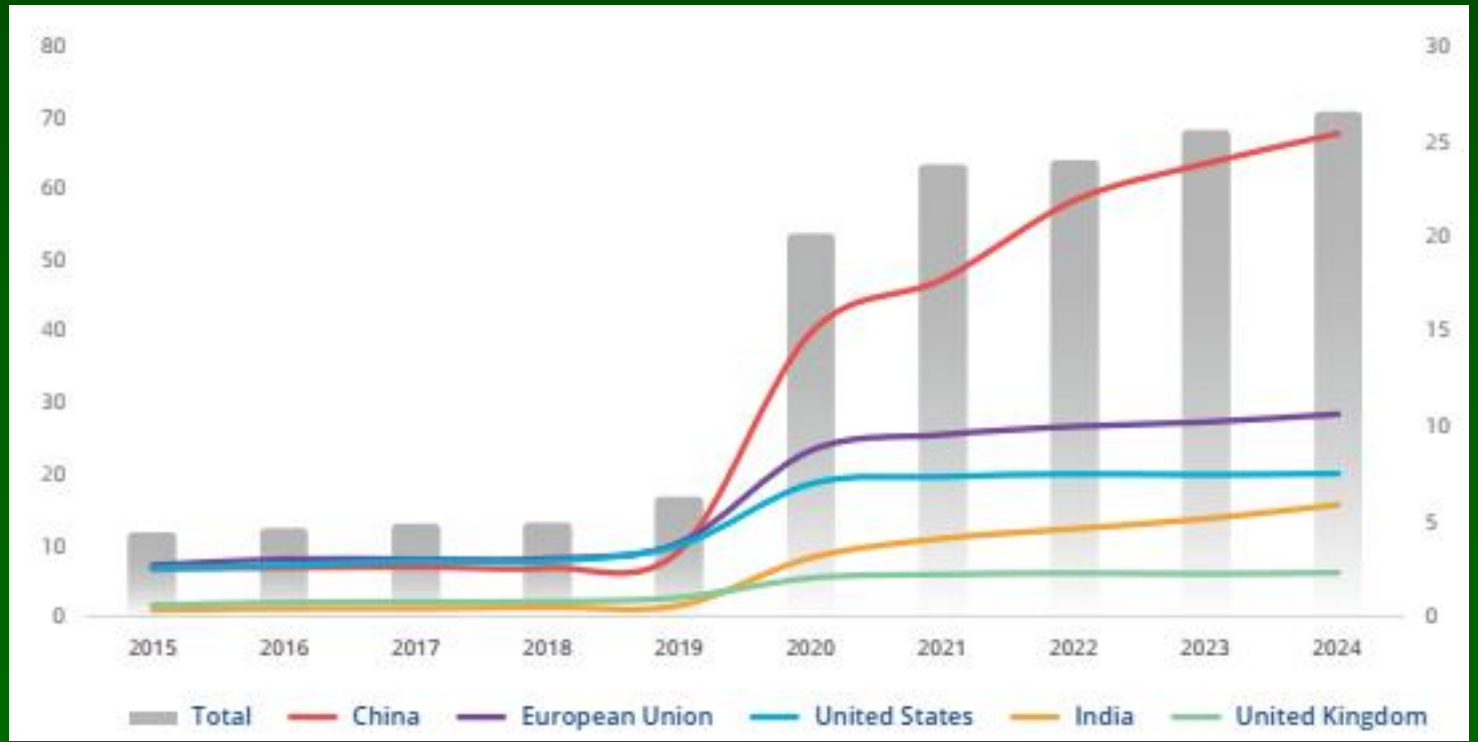
Research Topics: how to reduce social biases, improve code generation, security or privacy, mathematical problem solving, performance on low-resource languages, check factuality, how to reduce hallucination, estimate uncertainty and confidence of LLMs.

29.09.2025 much better Claude 4.5 Sonnet model.

AI in science report

Nature Research Intelligence report “AI in Science”, written by Fudan University and Shanghai Academy of AI for Science 5/2025.

Number of articles on AI in computer science, physics, mathematics, engineering, geosciences, environment, social sciences:
China >> EU > US, India > UK; Life sciences EU< US, China > India > UK.
In 2024 about 955 000 publications on AI, in physical sciences 70,700.
Nature Index 01.10.2025, Research hospitals



DeepMind Alphas

Superhuman level in:

AlphaGo, AlphaZero, Alphastar – real-time strategy game Starcraft-II (2019), and many other strategic games, like a war games.

MuZero AI Masters Games Without Even Being Taught the Rules (2020)

AlphaDev optimizing computer systems.

AlphaGenome, AlphaProteo, AlphaMissense.

Physics/Chemistry: AlphaQubit, QuantumMatter, Fusion, GNoME

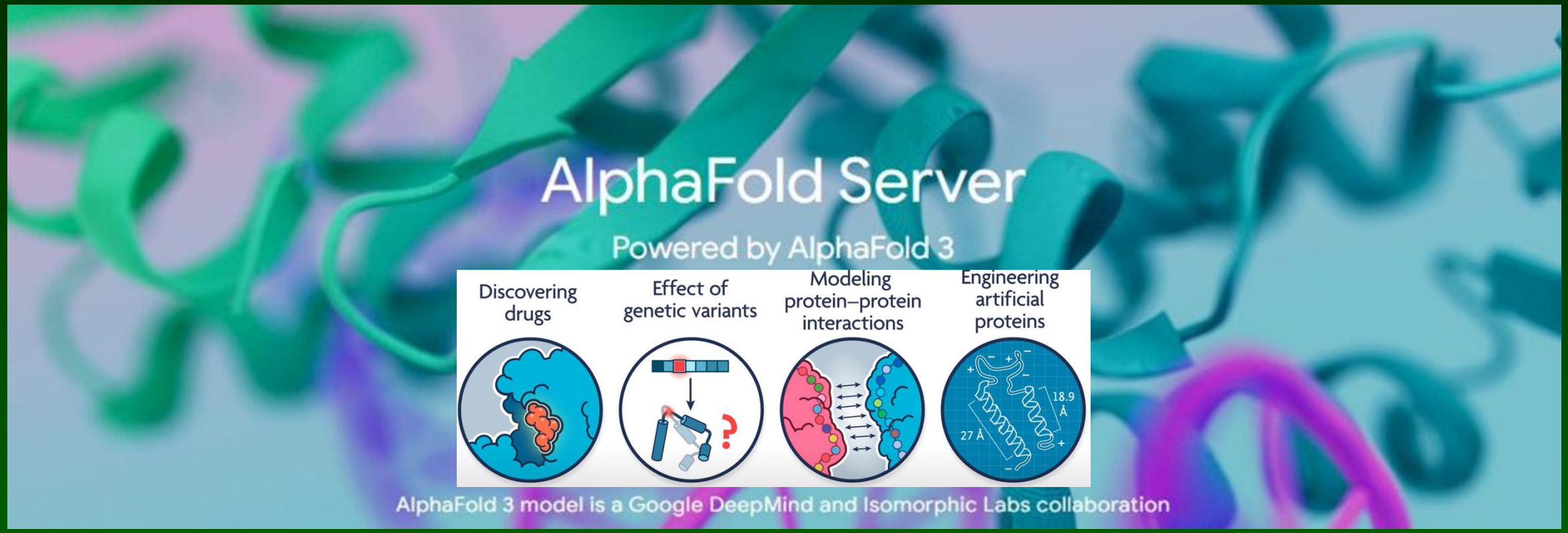
Math: AlphaEvolve, AlphaProof, AlphaGeometry.

AlphaChip superhuman chip layouts, ex. 3 generations of the Tensor Processing Unit (TPU).

AlphaEarth data streams: satellite imagery, sensor data, geotagged Wikipedia entries => unified digital representation that tracks crop cycles, coastlines, urban expansion, melting ice ...

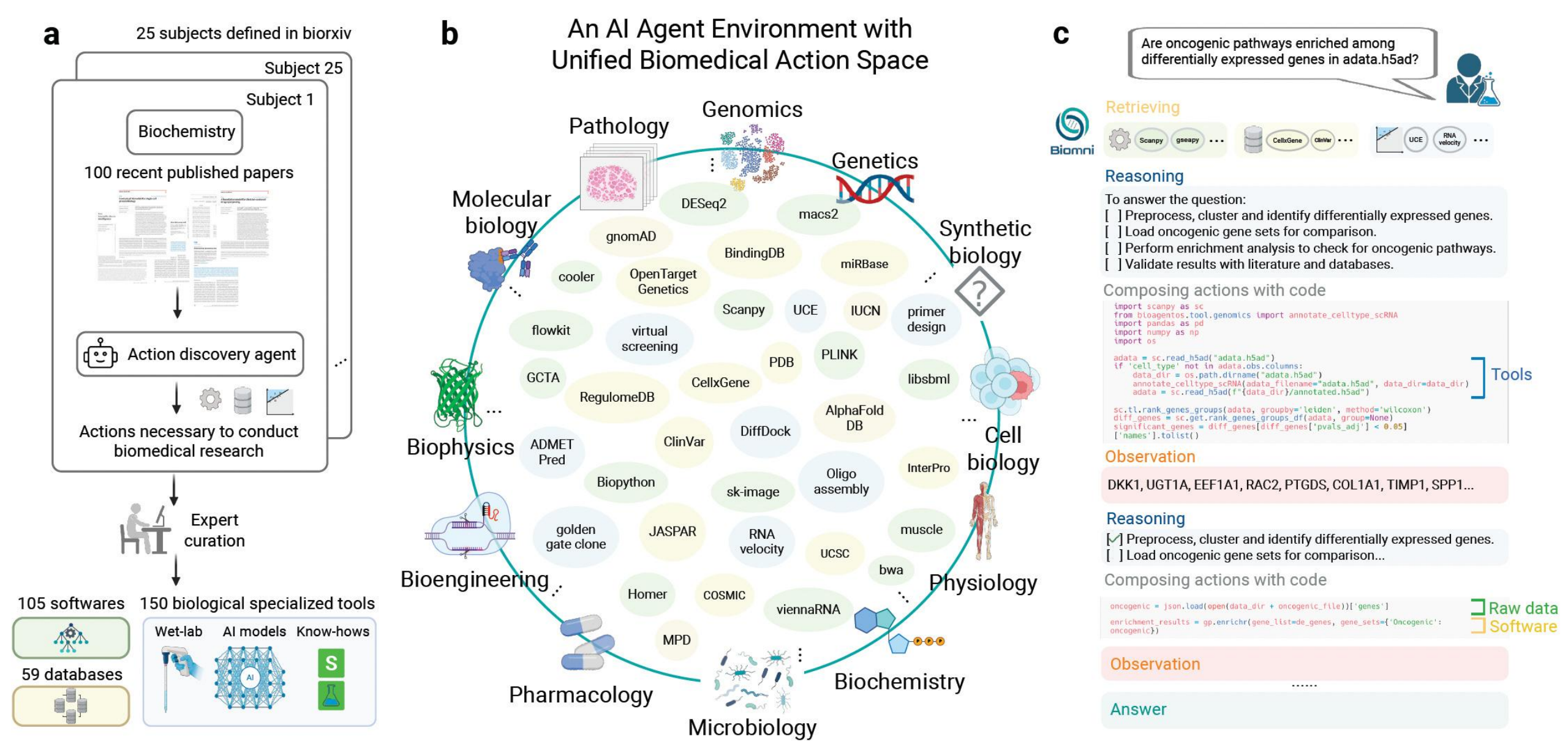
WeatherNext, a family of AI models produces state-of-the-art weather forecasts.

AlphaFold 3



J. Jumper + 32 coauthors + Demis Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature **596**, 583 (2021).
Now greatly improved: J. Abramson + 56 coauthors, Accurate structure prediction of biomolecular interactions with AlphaFold 3, Nature **630**, 493 (2024).

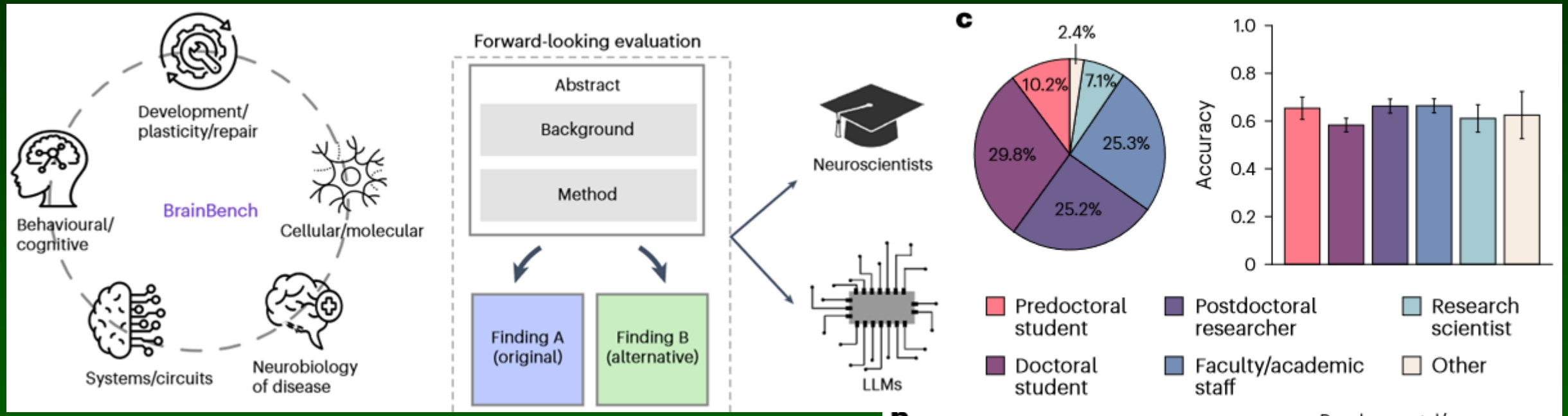
AlphaFold 3 predicts the structure and interactions of all of life's molecules, AlphaProteo generates novel proteins for biology and health, accelerating research in nearly every field of biology/molecular medicine.



Stanford research Biomni, biomedical AI agent that autonomously tackles a wide spectrum of research tasks across diverse subfields, mapped the biomedical action space by mining publications across 25 domains to curate an environment with 150 specialized tools, 105 software packages, and 59 databases. BioCyc/HumanCyc.

BrainBench predicts neuroscience results

Luo, X., Rechardt, A., Sun … Love, B. C. (2024). Large language models surpass human experts in predicting neuroscience results. *Nature Human Behaviour*, 1–11.



BrainBench is a forward-looking benchmark for predicting neuroscience results. Fine-tuned Mistral-7B can forecast novel results better than human experts.
LLMs surpass experts in predicting experimental outcomes and estimating their confidence.
This approach is quite general.

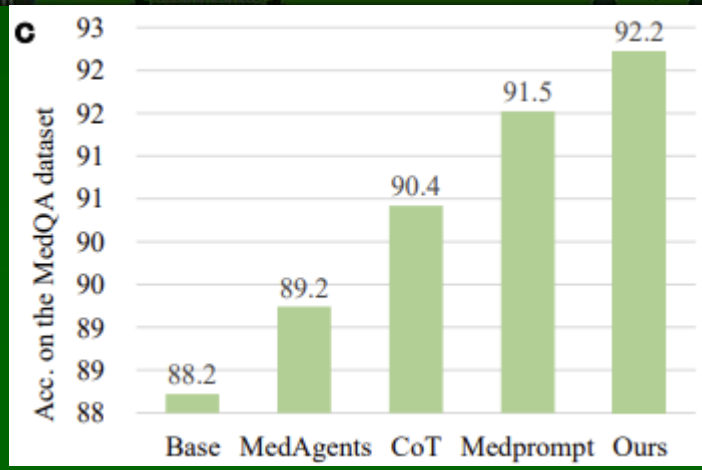
Tsinghua Hospital Simulacrum

Simulacrum-based Evolutionary Agent Learning (SEAL).

Hospital with 25 GPT-3.5 agents, 16 functional areas,
21 clinical departments, covering 339 diseases, simulates
the entire process of treating disease.
Autonomous agents as patients, nurses and doctors.
Agents request patient's tests and the doctor learns
accumulating experience.

Evaluation: correct medical examination, diagnosis,
and treatment plan recommendation.

After enrolling tens of thousands patients, Agent-Doctor reaches high level of success in treating diseases in virtual world, and top results on tests on the MedQA database, without any manually labeled training data.



Li J. et al. (5/2024). *Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents.*

This year: Tsinghua AI Agent Hospital (5/2025) was opened.
AI is used at every step, building a new blueprint for integrated care.
DeepSeek AI is already embedded in the intranets of more than 260 Chinese hospitals.

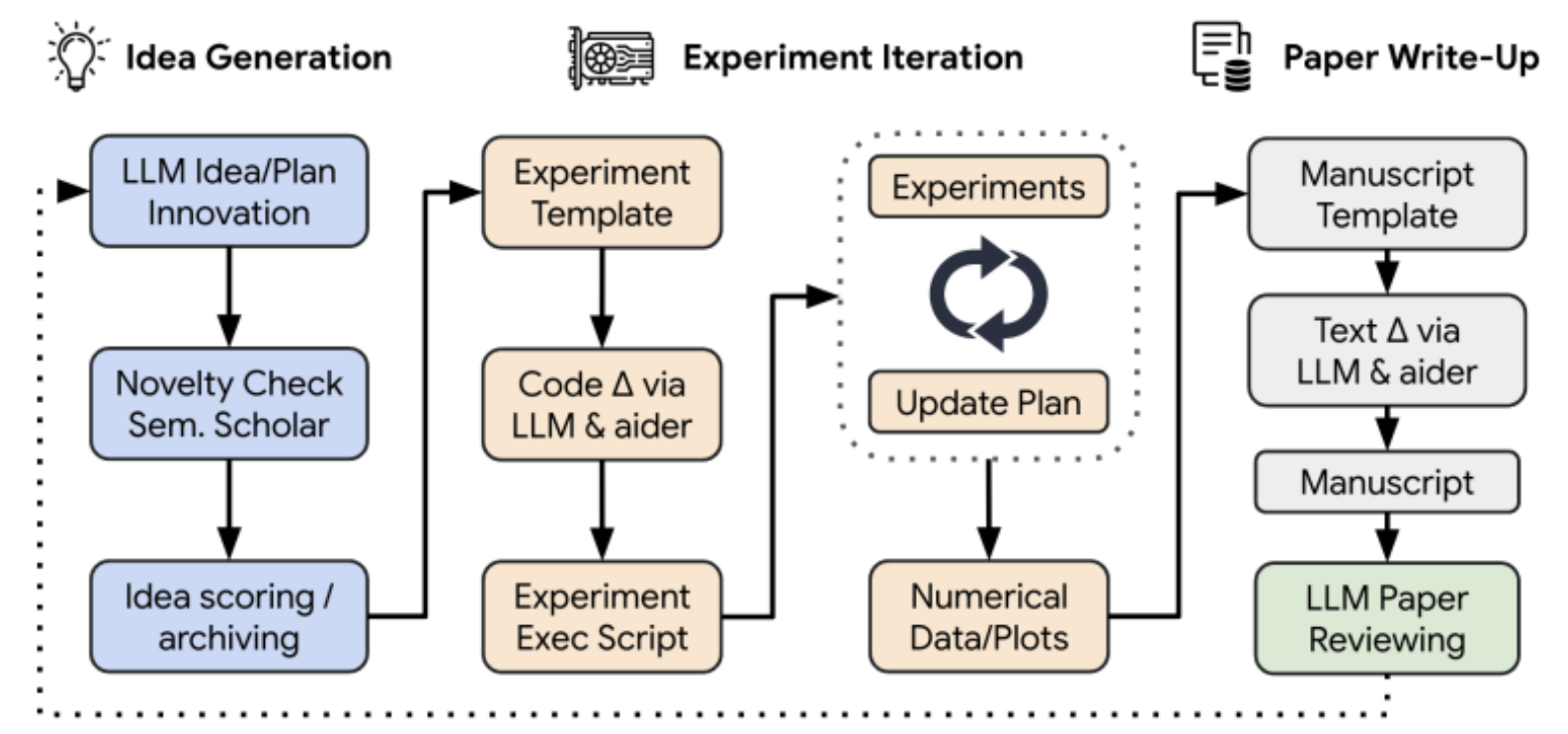
Towards autonomous agentic science

AI Scientist

The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. 08/2024

The AI Scientist (Sakana) software generates novel research ideas, writes code, executes experiments, visualizes results, describes its findings by writing a full scientific paper, simulates review process for evaluation. This process can be repeated to iteratively develop ideas in an open-ended fashion, acting like the human scientific community.

10 original papers, ex: Adaptive Learning Rates For Transformers Via Q-learning.



Github

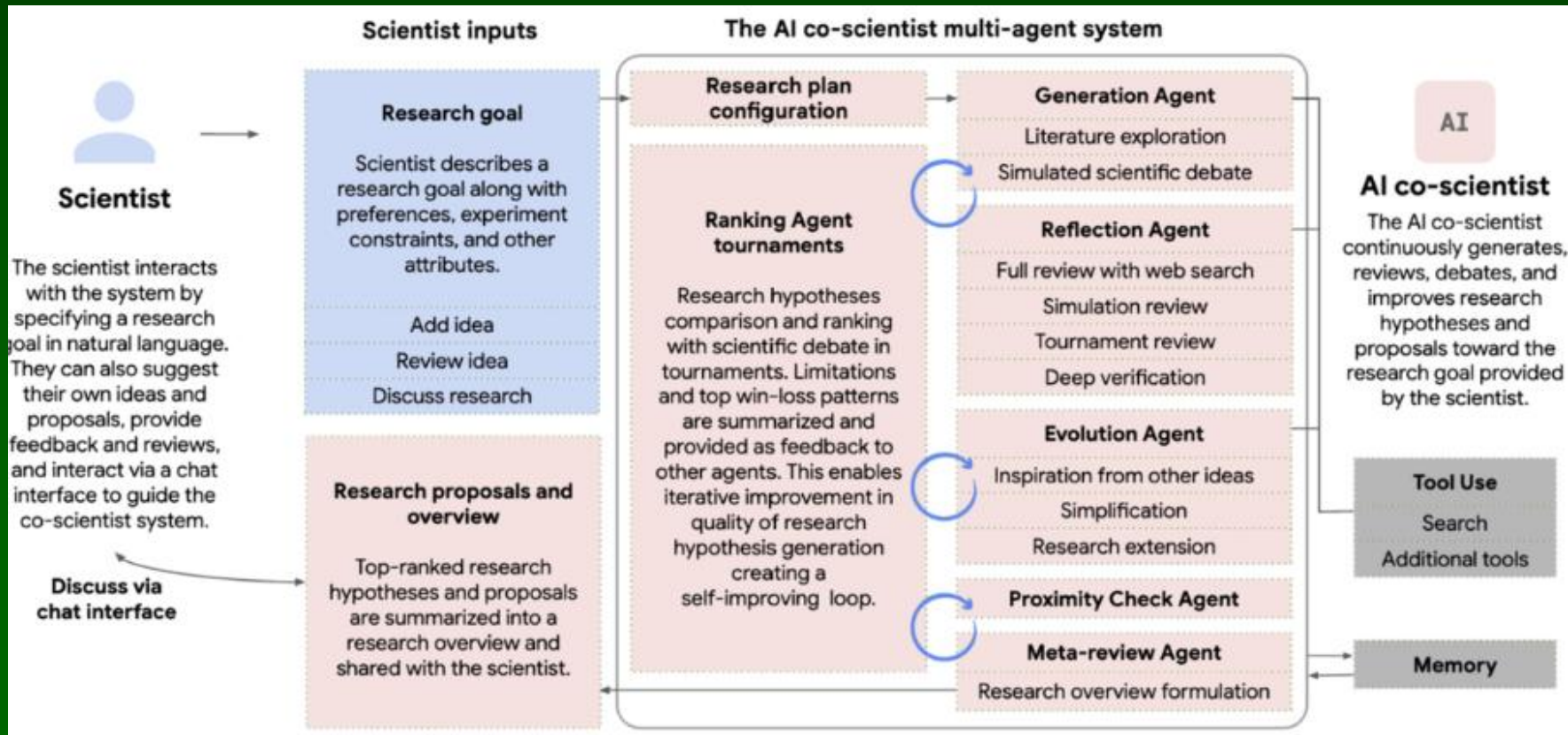
3 templates: NanoGPT, 2D Diffusion, Grokking

Sakana: paper accepted at ICLR 2025 conference!

Google AI Co-Scientist

AI co-scientist is a multi-agent AI system — Generation, Reflection, Ranking, Evolution, Proximity, Supervisor, and Meta-review, built with Gemini 2.0 as a virtual scientific collaborator to generate novel hypotheses and research proposals. Key reasoning steps include self-play–based scientific debate for novel hypothesis generation, ranking tournaments for hypothesis comparison, and an "evolution" process for quality improvement, recursive self-critique, including tool use for feedback to refine hypotheses.

Several examples of biomedical applications, including antimicrobial resistance, are in Google blog page.



The AI co-scientist system design

Virtuous Machines: Towards Artificial General Science

Wehr, G … & Ehrhardt, S.E. (2025). *Virtuous Machines: Towards Artificial General Science.* arXiv:2508.13421

Agentic system incorporating hypothesis generation through experimental design, physical experimental implementation, data analysis, result interpretation, theory refinement, visualization, and reporting. Completion of a full scientific study required about 17 hours' of processing time for a total marginal cost of ~$114 USD per research project (+ human participant ~$4,500 USD).

3 cognitive science experiments were conducted, testing visual working memory (VWM), mental rotation, and imagery vividness, with online data collection involving 288 participants.

AI performed real-world experiment, developed analysis pipelines and produced completed manuscripts for each experiment.

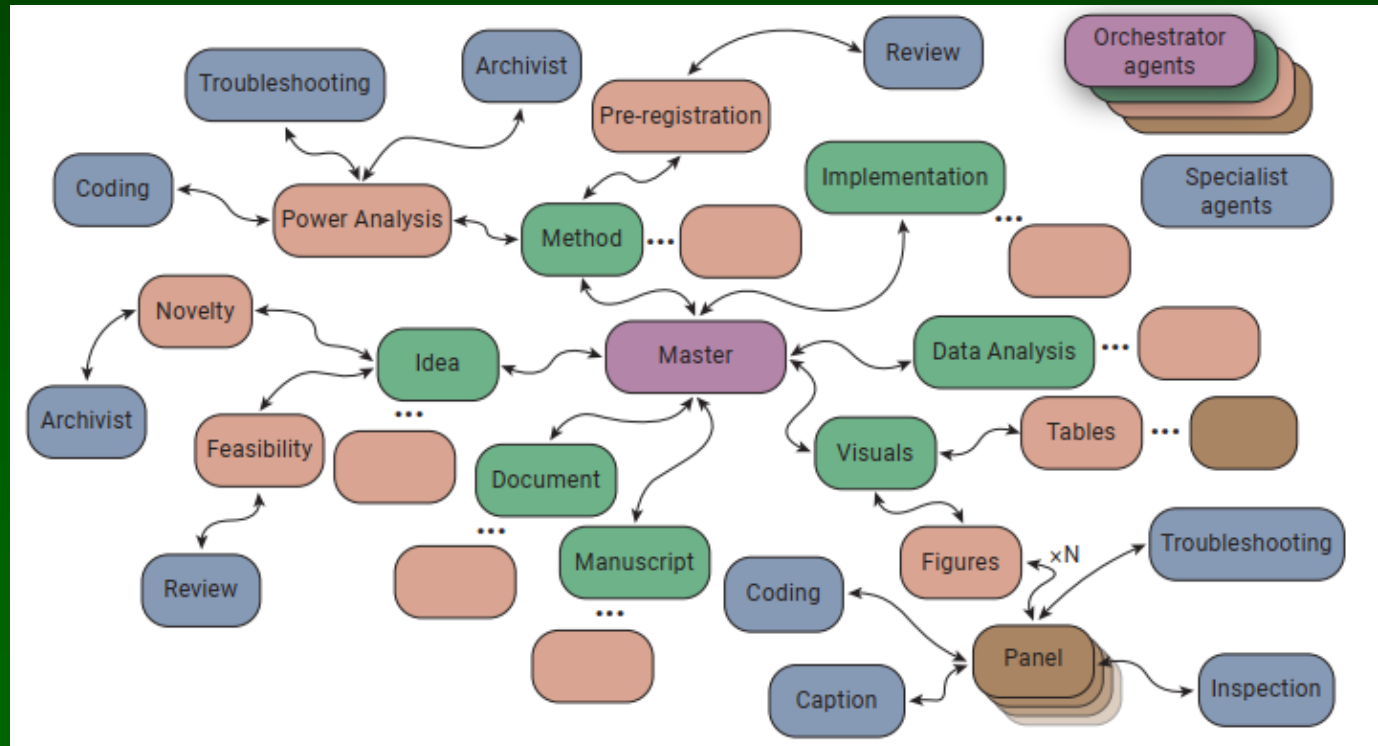


Figure 1: | Simplified network architecture of the autonomous scientific discovery

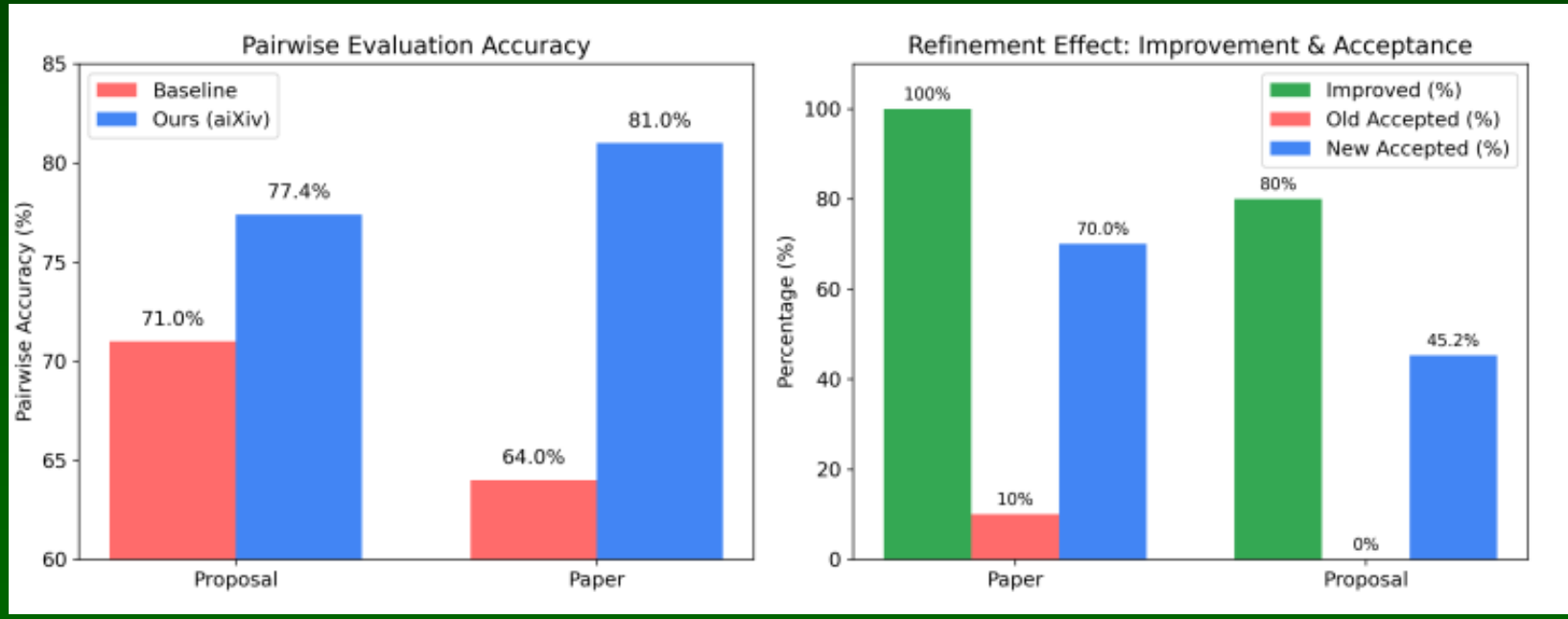
aiXiv for iterative improvement

Zhang, P. … Liu, X. (2025). aiXiv: A Next-Generation Open Access Ecosystem for Scientific Discovery Generated by AI Scientists, and https://github.com/aixiv-org

aiXiv significantly enhances the quality of AI-generated research proposals and papers after iterative revising and reviewing on aiXiv. The content is routed to a panel of LLM-based review agents assessing the novelty, technical soundness,clarity, feasibility, and overall potential impact of the submission. Structured feedback is generated to guide revisions.

aiXiv refines the proposal or paper, improving methodological rigor, clarifying contributions, addressing reviewer concerns, and incorporating recommended citations.

It improved all papers and 80% of proposals.



Agents4Science Conference

Agents4Science 2025 conference, Stanford, virtual, 22.10.2025.

Advisory board includes chief editor of *Nature Biotechnology*, Nobel Laureate in Economics, Harvard, Chicago, Rutgers and Scripps Research professors.

The 1st Open Conference of AI Agents for Science: AI serve as both primary authors and reviewers of research papers. It will explore if and how AI can independently generate novel scientific insights, hypotheses, and methodologies while maintaining quality through AI-driven peer review.

Is AI good enough to do science and review papers? We don't know yet.

Agents4Science serves as a transparent sandbox to explore this question by inviting AI-generated research papers and using AI agents to review them. It is the first venue where AI authorship is required, enabling open evaluation of AI-generated research.

We aim to create a clear picture of how AI can participate in scientific research, requiring disclosures of AI involvement in the research process. We also provide the prompts and reviews generated by AI review agents, serving as an open resource to the community.

Autonomous Scientific Discovery

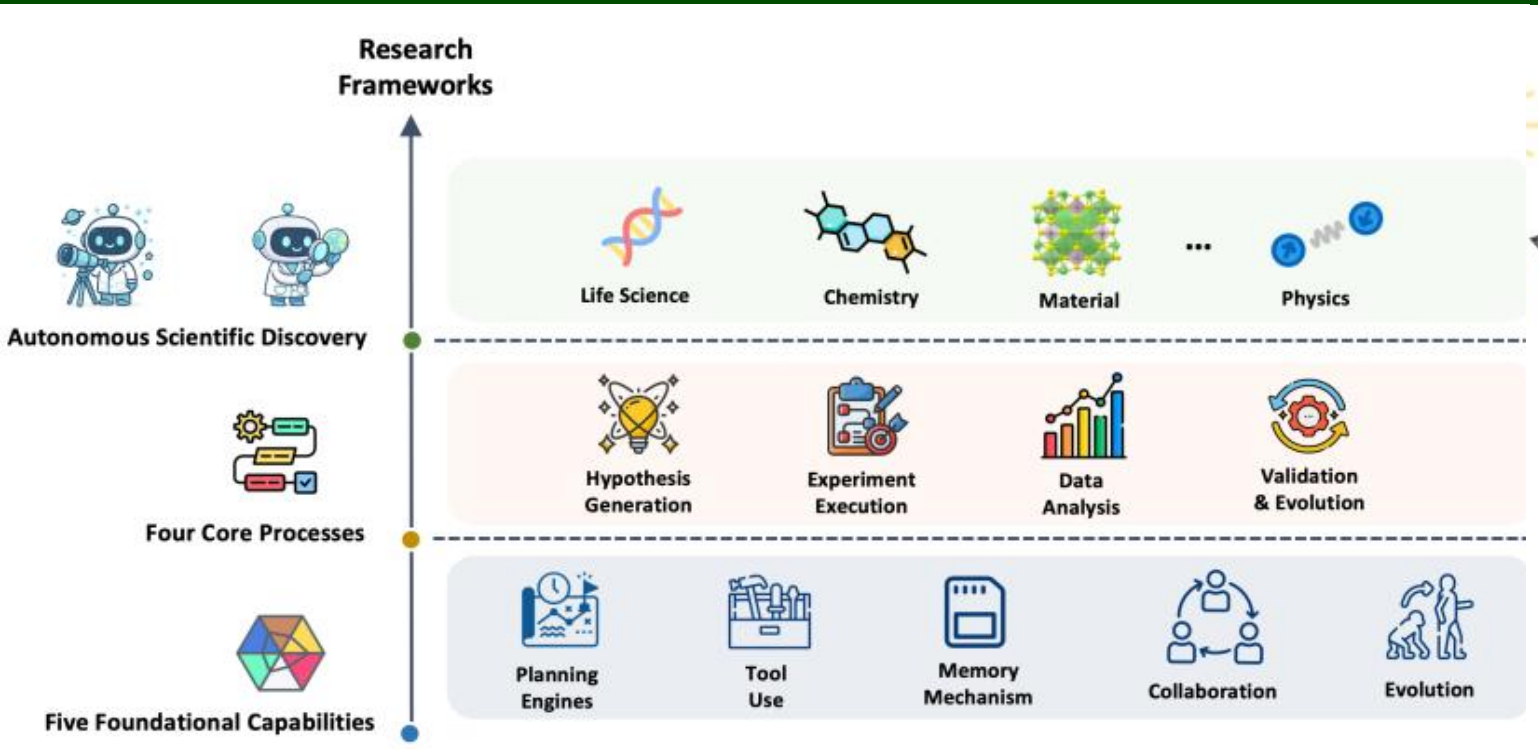
Level 1: AI as a Computational Oracle (Expert Tools)

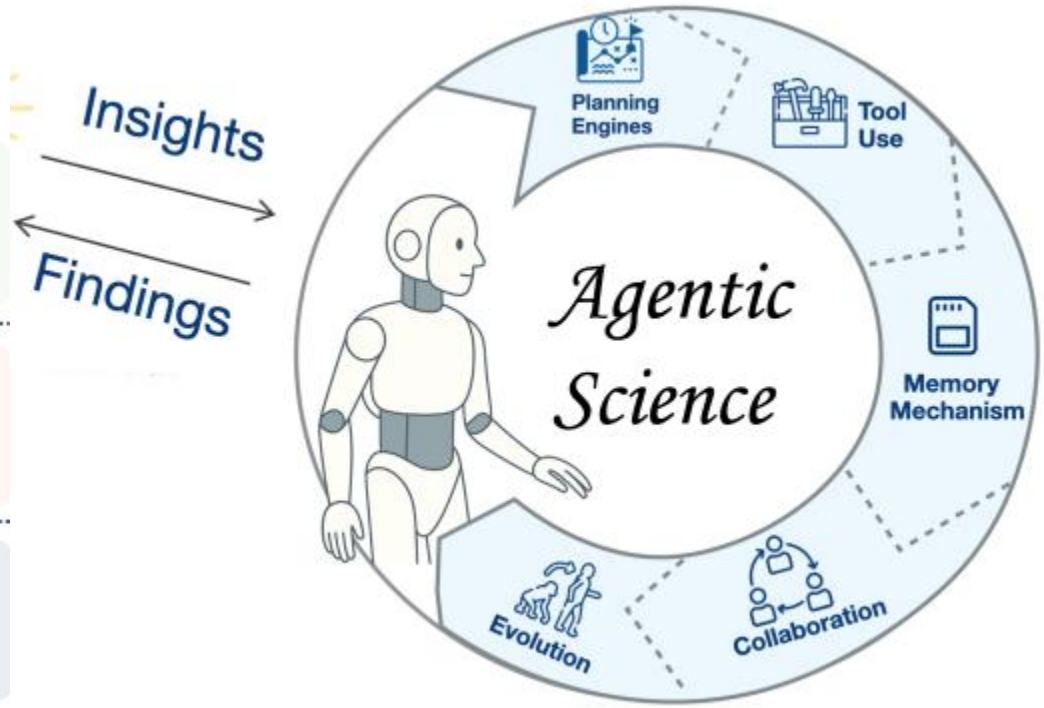
Level 2: AI as an Automated Research Assistant (Partial Agentic Discovery)

Level 3: AI as an Autonomous Scientific Partner (Full Agentic Discovery)

Level 4: AI as a Generative Architect (Future Prospect)

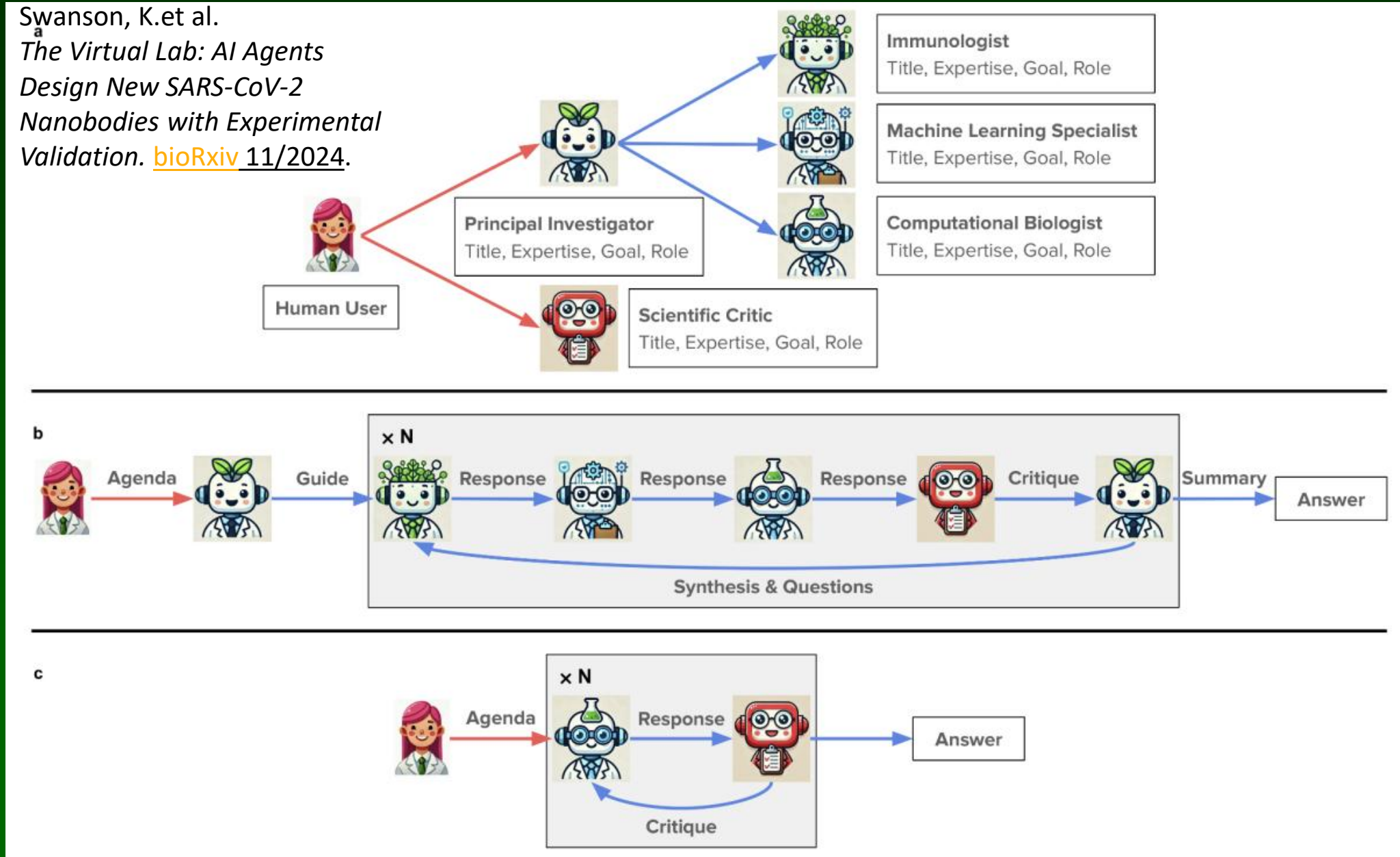
Scientists need new skills: give clear, context-rich instructions that shape the agent's policy, manage the toolset available to the agent, and judge when to trust its outputs vs when to apply deeper scrutiny.





Future: Virtual Labs with specialized agents

Swanson, K.et al.
The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation. bioRxiv 11/2024.

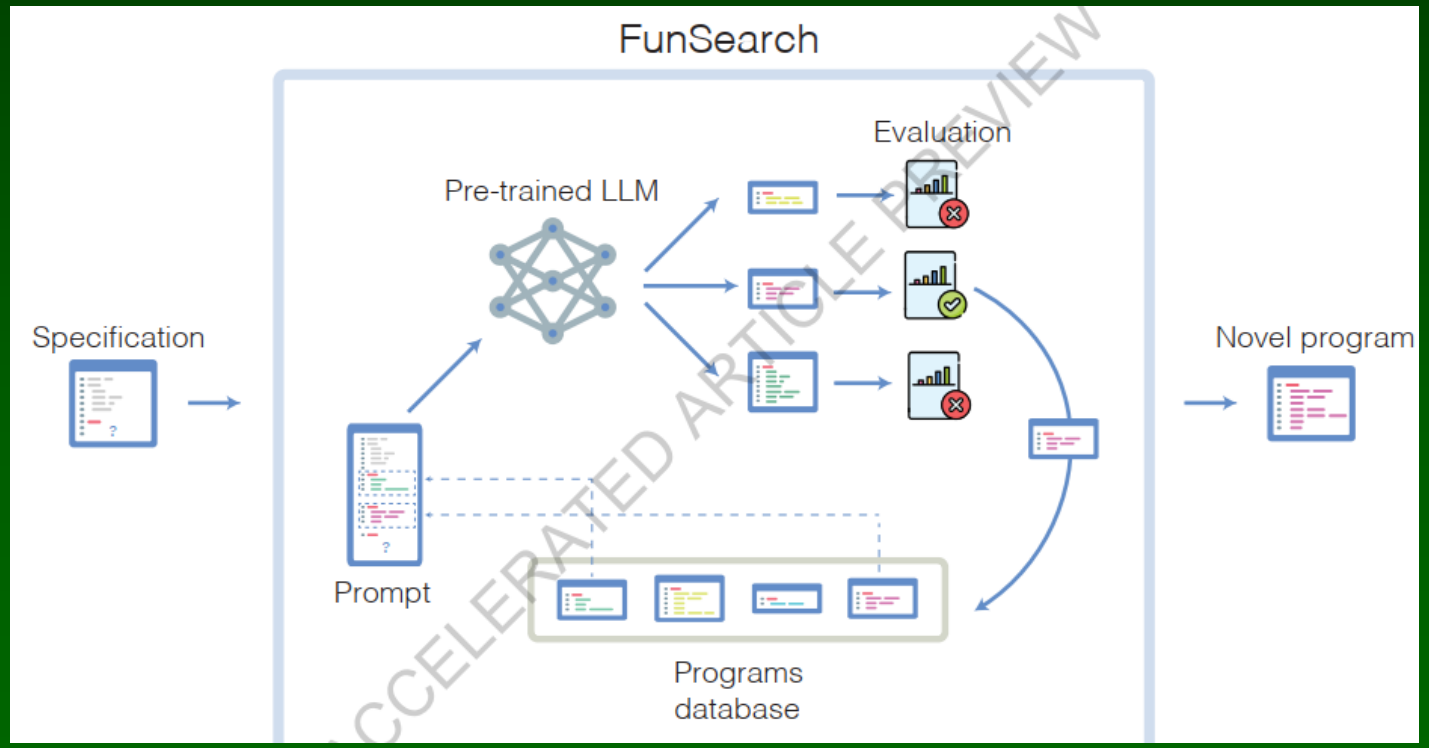


Examples of discoveries

Mathematical discoveries

Romera-Paredes … Fawzi, A. (2023). Mathematical discoveries from program search with large language models. *Nature*, 1–3.

FunSearch (*search*ing in the *fun*ction space), pairing LLM with evaluator. *FunSearch* applied to the cap set problem discovered new constructions of large cap sets going beyond the best known ones.



Strassen algorithm (1969) improved!
Fawzi, A. et al. (2022). Discovering faster matrix multiplication algorithms with reinforcement learning. Nature, 610(7930), 47–53.

DARPA expMath: Exponentiating Math.

The goal of expMath is to radically accelerate the rate of progress in pure mathematics by developing an AI co-author capable of proposing and proving useful abstractions and exponentially increase the rate of progress in mathematics.

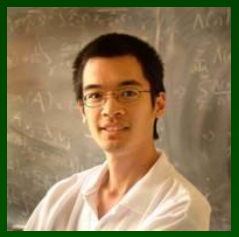
AlphaEvolve, AlphaGeometry, AlphaTensor.

Math.inc

Gauss, an agent for autoformalization aimes at a new paradigm — verified superintelligence and the machine polymaths. It will soon dramatically compress the time to complete massive initiatives.

The next generation of reasoning agents will discuss the hardest science and engineering problems with precision, self-improving with formal correctness guarantees that make their reliability inevitable.

The Strong Prime Number Theorem was set by Terence Tao and Alex Kontorovich in 2024. After 18 months only intermediate progress was done. Gauss formalized the key missing results in complex analysis, completed the project after 3 weeks, opening initiatives previously considered unapproachable.



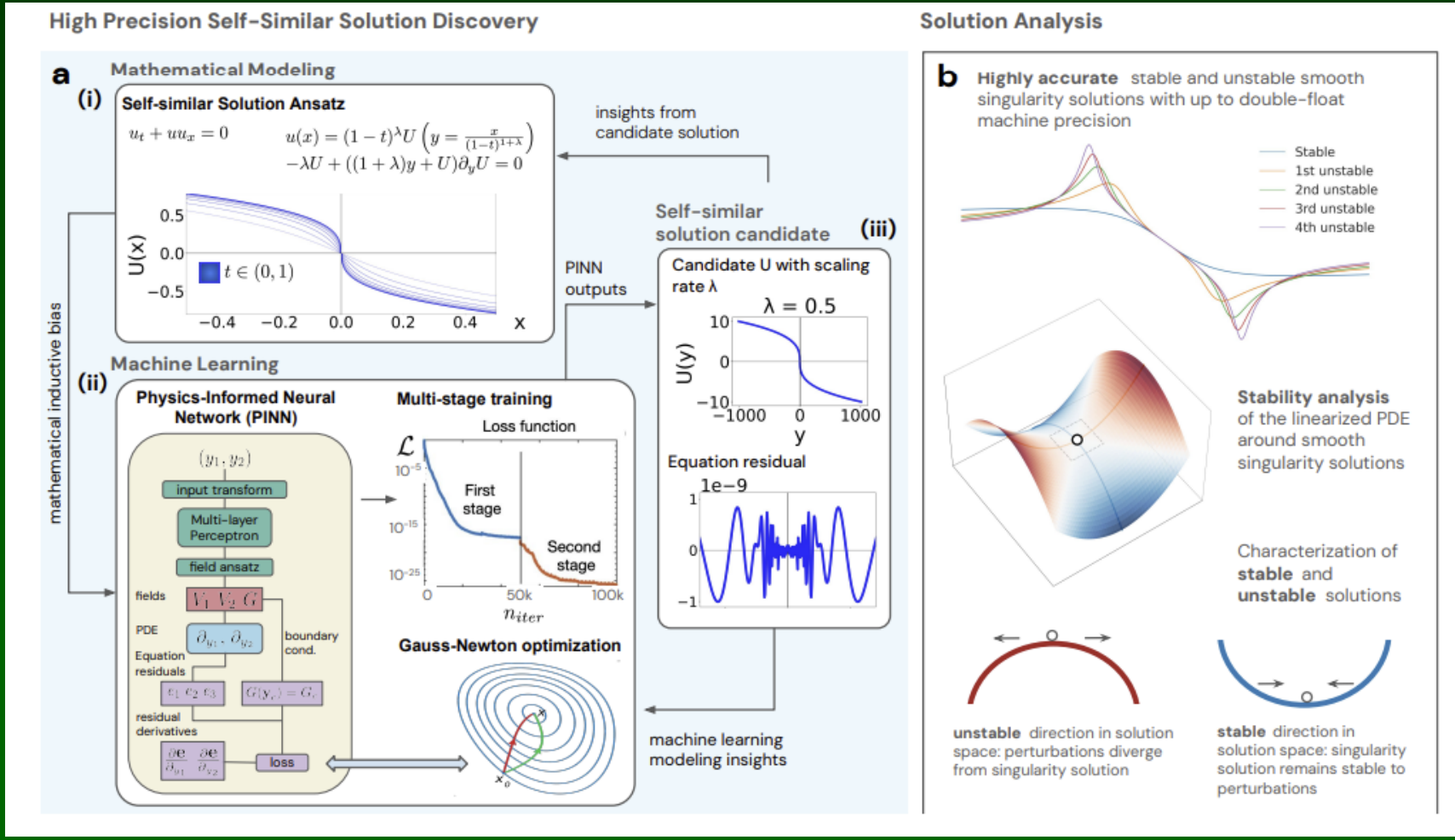
Math Inc. plans to expand mathematical code by 100-1000 x within 12 months, creating training data for "machine polymaths" and “verified superintelligence”.

Fluid dynamics

Wang, Y. et al. (9/2025). *Discovery of Unstable Singularities*.

Empirical results from candidate solutions and their accuracy guide the mathematical modeling by a Physics-informed Neural Network (PINN).

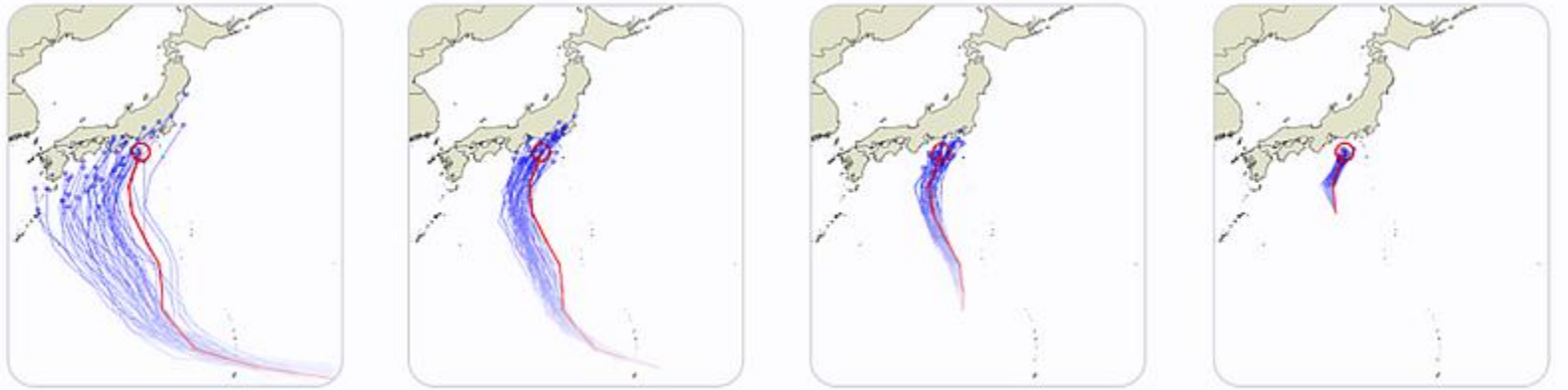
We discover unstable modes, along which any slight perturbation causes the system to deviate from the blow-up trajectory and confirm the discovery of highly accurate stable and unstable singularities.

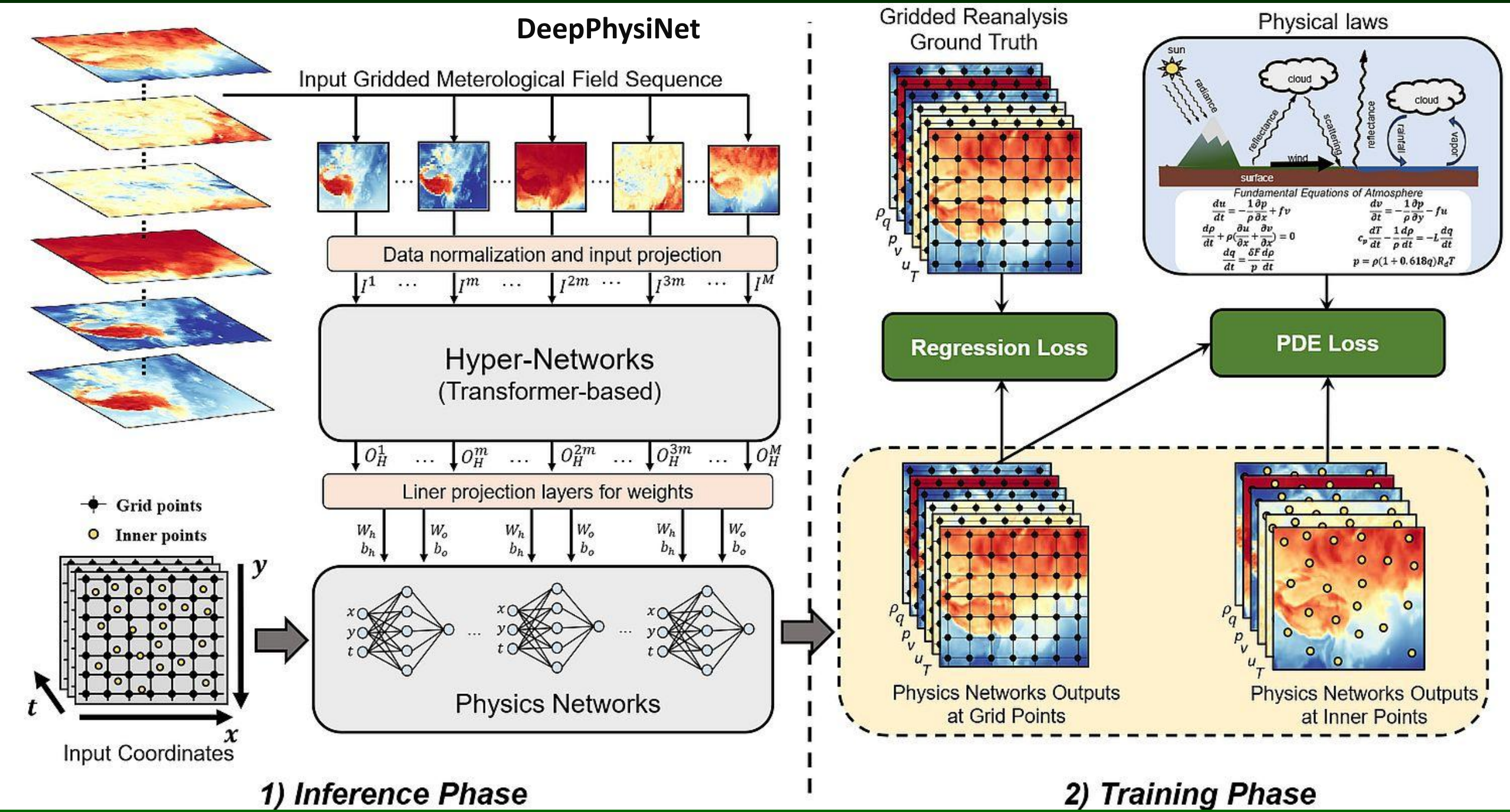


GenCast and Weather Lab

I. Price et al. Probabilistic weather forecasting with machine learning. *Nature*, 1–7 (24).

- Google DeepMind diffusion model predicts the weather more accurately than the best system currently in use, generating forecasts up to 15 days in advance in minutes.
- Computer-generated weather forecasts is using large 400 km^2 cells, averaging condition within such areas. GenCast is using statistical post-processing method to account for variation in each cell, and learns how weather can influence the state of cells.
- The GenCast system works better than the best European medium-range operational model. It can predict extreme weather, hurricanes, catastrophic rains, and heatwaves.
- For extreme flash-flood rainfalls current forecast is less than one day, GenCast extends it to 5 days.




DeepPhysiNet
Input Gridded Meterological Field Sequence
Data normalization and input projection
Hyper-Networks
(Transformer-based)
Liner projection layers for weights
Physics Networks
Grid points
Inner points
Input Coordinates
1) Inference Phase
Gridded Reanalysis
Ground Truth
Physical laws
sun
radiance
cloud
reflectance
scattering
wind
surface
rainfall
vapor
Fundamental Equations of Atmosphere
Regression Loss
PDE Loss
Physics Networks Outputs
at Grid Points
Physics Networks Outputs
at Inner Points
2) Training Phase

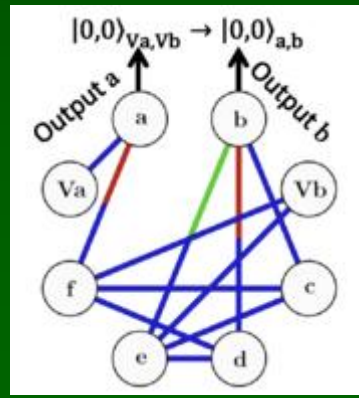
Theseus

- Krenn, M., Kottmann, J. S., Tischler, N., & Aspuru-Guzik, A. (2021). Conceptual Understanding through Efficient Automated Design of Quantum Optical Experiments. Phys. Rev. X 11, 031044

Theseus, an algorithm for new conceptual understanding, was applied to experimental quantum optics.

1) A graph-based representation of quantum optical experiments that can be used algorithmically.
2) Automated design approach for new quantum experiments, orders of magnitude faster.
3) Solved several crucial open questions in photonic quantum technology.
4) Theseus can produce solutions that a human scientist can interpret and gain new scientific concepts.

It can answer open questions and provide new concepts in quantum physics.
It shows how AI algorithms can contribute to science on a conceptual level.



- Krenn, M., Yehonathan Drori, & Rana X Adhikari. (2025). Digital Discovery of Interferometric Gravitational Wave Detectors. *Physical Review X, 15*(2).

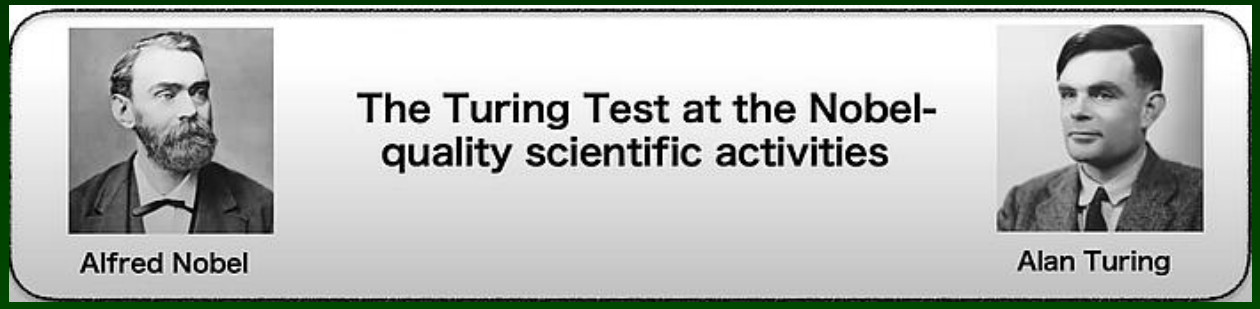
This design increases the potentially observable volume of the Universe 50-fold.
Analyzing the best solutions from our superhuman algorithm, we uncover entirely new physics ideas at their core.

Agentic science future

ACM Transactions on AI for Science (TAIS)

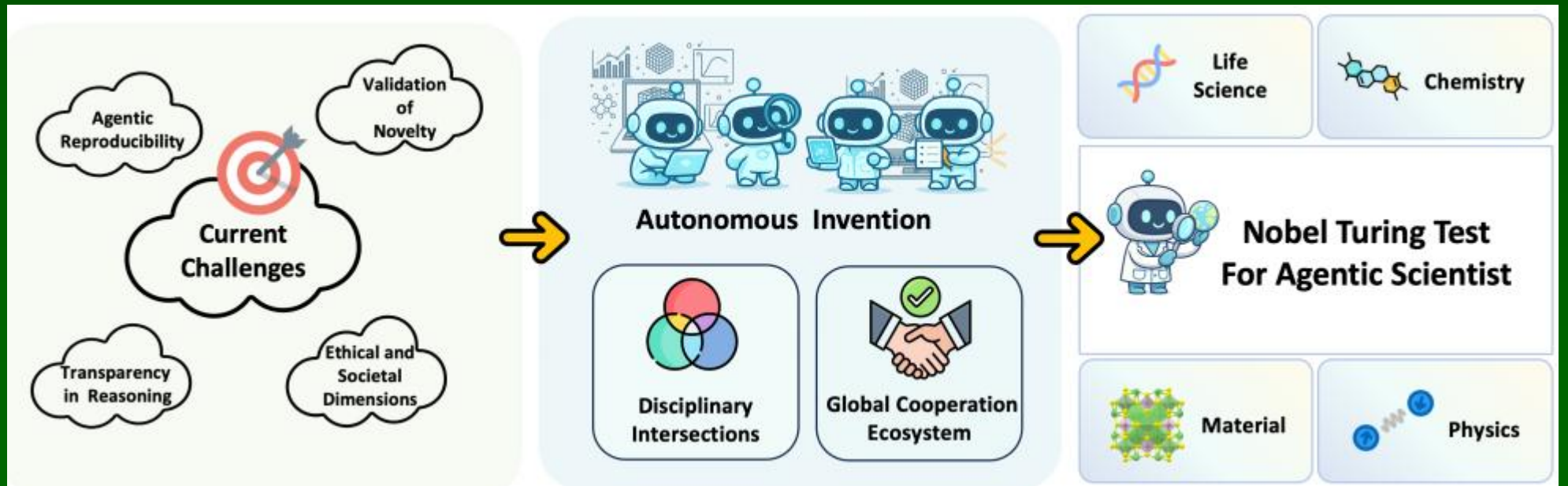
Nobel Turing Challenge

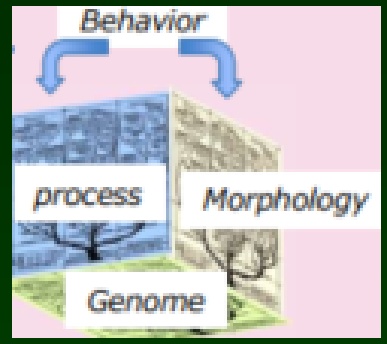
Nobel Turing Challenge (Hiraoki Kitano, Systems Biology Institute, Kyoto).



A grand challenge aimed at developing a highly autonomous AI and robotics system that can make major scientific discoveries, some which may be worthy of the Nobel Prize and even beyond.

5th Conference Singapore AI for Science and Nobel Turing Initiative, July 2024. Challenge on YouTube.



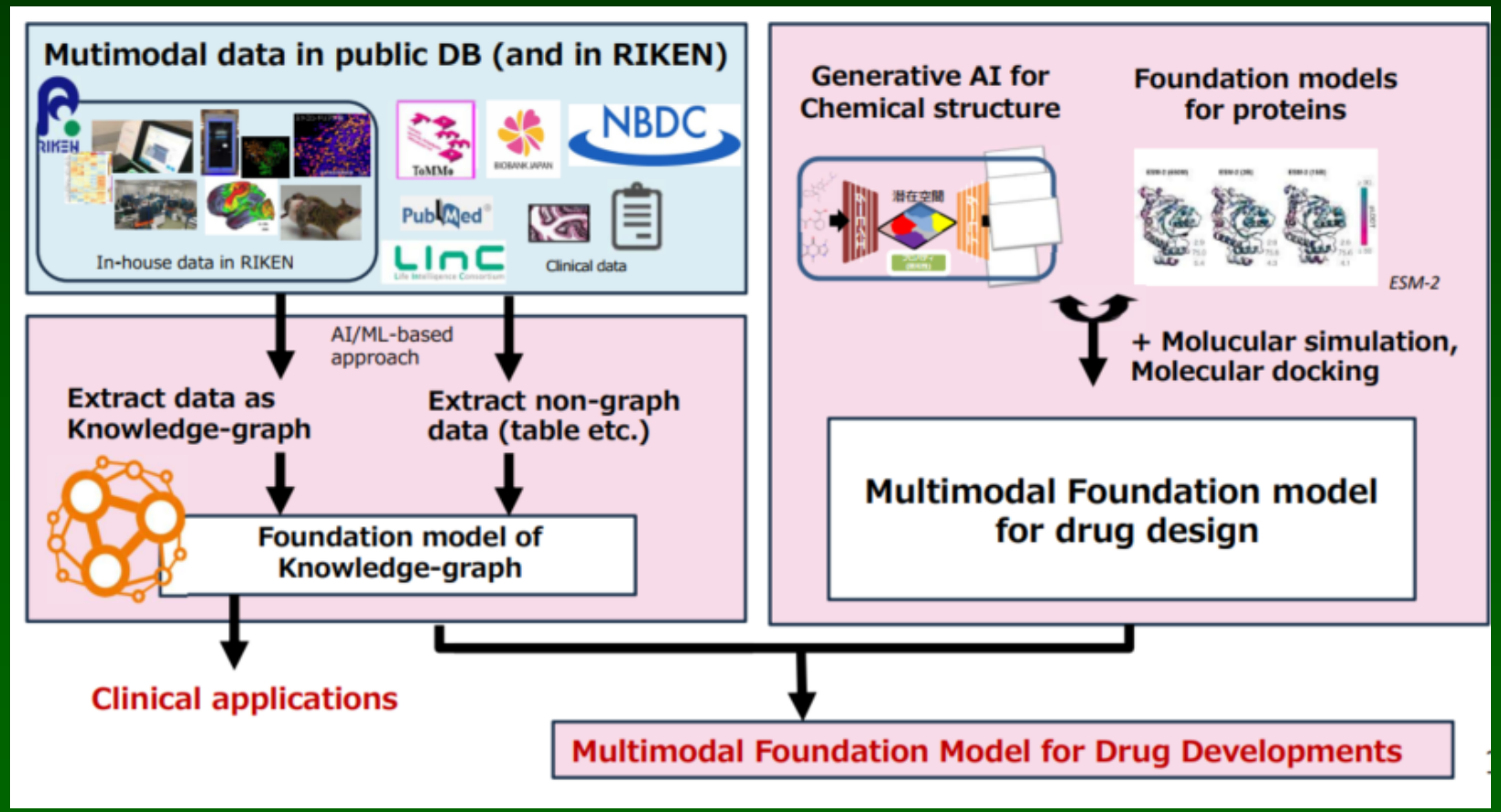


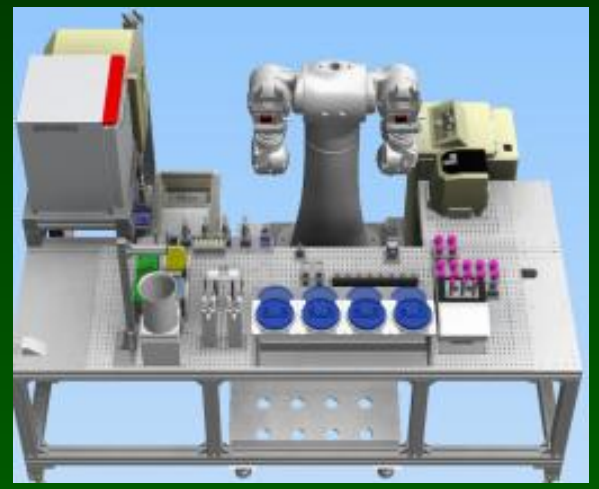
Building, training, and using large-scale models + hardware development (3/2024).
(a) AI development, LLMs/LMMs, scalable libraries and frameworks, AI workflows, data aggregation, model evaluation, alignment, etc.;
(b) design/build hardware and software systems, research automation;
(c) AI applications in science, engineering, medicine, and other domains.

In Japan RIKEN has launched the new program in 2024 called TRIP-AGIS,
Transformative **R**esearch **I**nnovation **P**latform - **A**rtificial **G**eneral **I**ntelligence for **S**cience.
Focused on multimodal foundation models of dynamic behavior, dynamic and spatial transcriptome, cellular response to drugs.

RIKEN foundation models

Multimodal Foundation Models for Drug Developments and Clinical application, material science, AI-driven automatic research, massive data production using robotic experiments, and simulations.

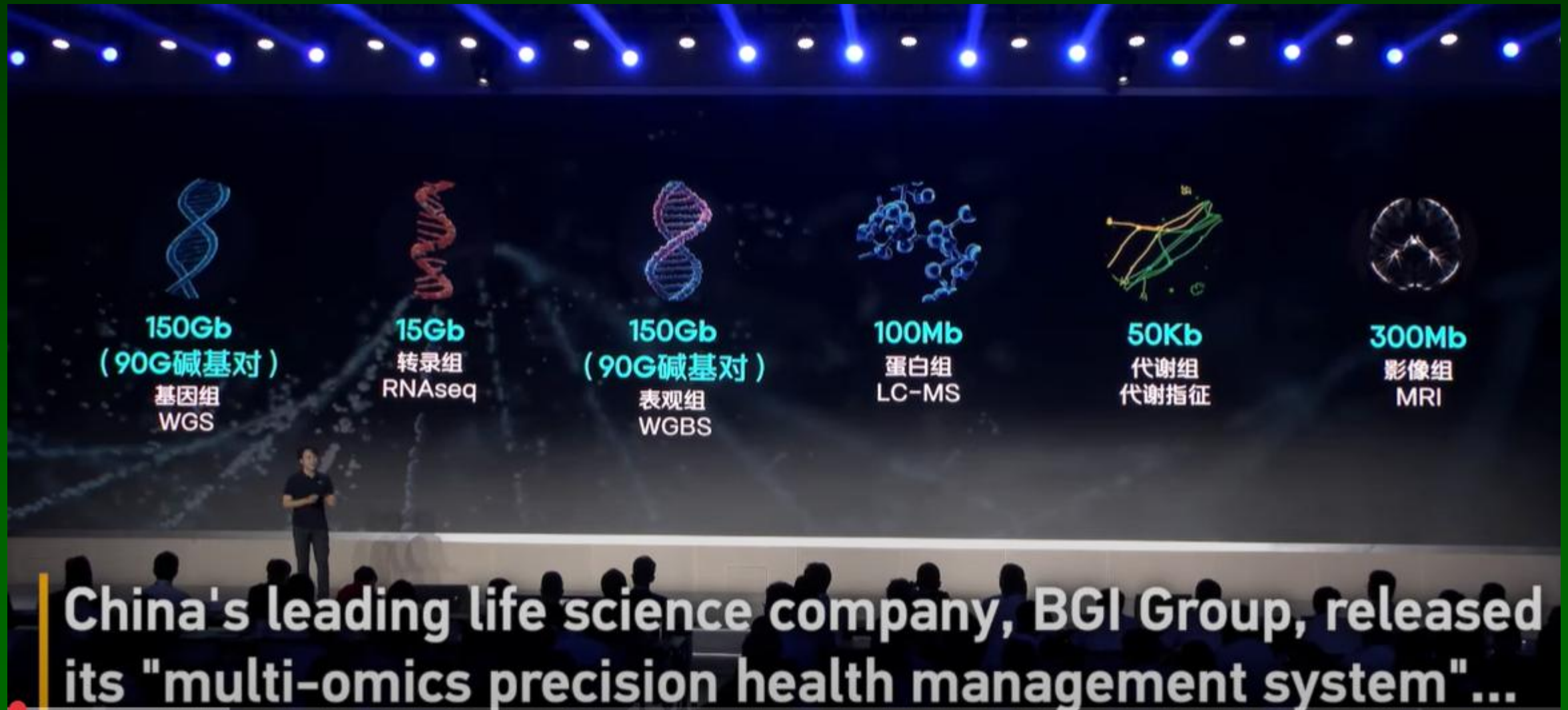


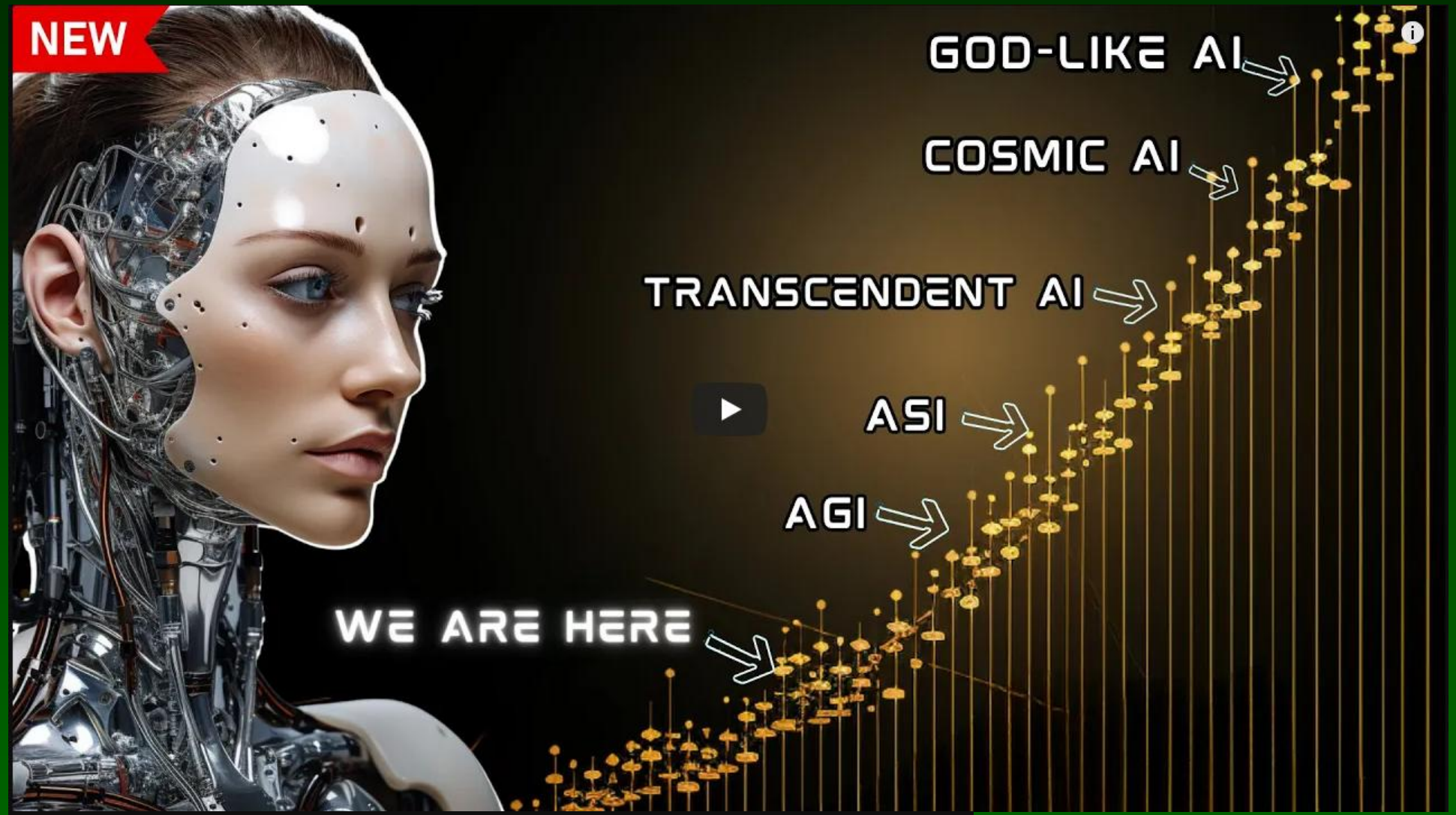


Robotics Biology, or RBI LabDroid, Tokyo. First humanoid robot for all lab work, for genomics, cell cultures, cell-based screening, proteomics, metabolomics.

BGI Multi-Omics

BGI Genomics Multi-Omics AI Model Shenzhen, China, is the world's leading integrated solutions provider of precision medicine, working in 100 countries and regions, involving more than 2,300 medical institutions.





AI TechXplorer

@AITechXplorer 788 subscribers 8 videos

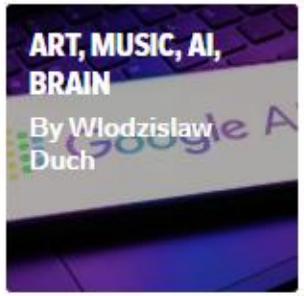
AI Tech Exploration: Unraveling the Wonders of Artificial Intelligence (AI) is...

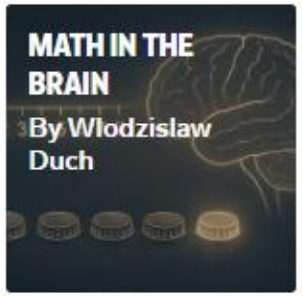
AGI = Artificial General Intelligence
2027-2030?
ASI = Artificial Super-intelligence.

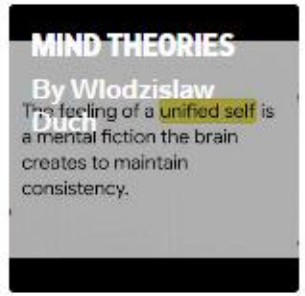
Questions posed by rapid AI development

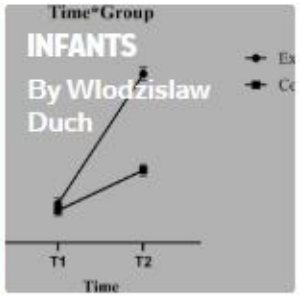
- We are at the turning point of human history, the big wave of technology is coming.
- AI is now incredibly big, with millions of models, a lot of new ideas, most benchmark are reaching saturation, with a few exceptions: math, scientific problems, Humanity's Last Exam.
- AI is changing the way science is done, already many discoveries have been made, there are still great opportunities for novel algorithms and applications.
- Science without strong support of AI agents will soon be irrelevant.
- Robots/AI systems can quickly learn from each other and may automatize many lab tasks.
- Half of the progress is due to hardware and data, but half is due to new algorithms.
- Small Language Models have many advantages, and can be massively deployed everywhere.
- Many jobs related to science may be taken by AI agents, universities will have to change.
- LMMs are capable of self-reflection and may mirror our minds, creating strong bonds with people.
- Our megalomania is dangerous. Our cognitive capabilities have limitations far greater than AI.
  We are entering completely uncharted territory.
  G. Hinton … digital intelligent beings are emerging.
  What will they teach us and what will we teach them?

MAGAZINES









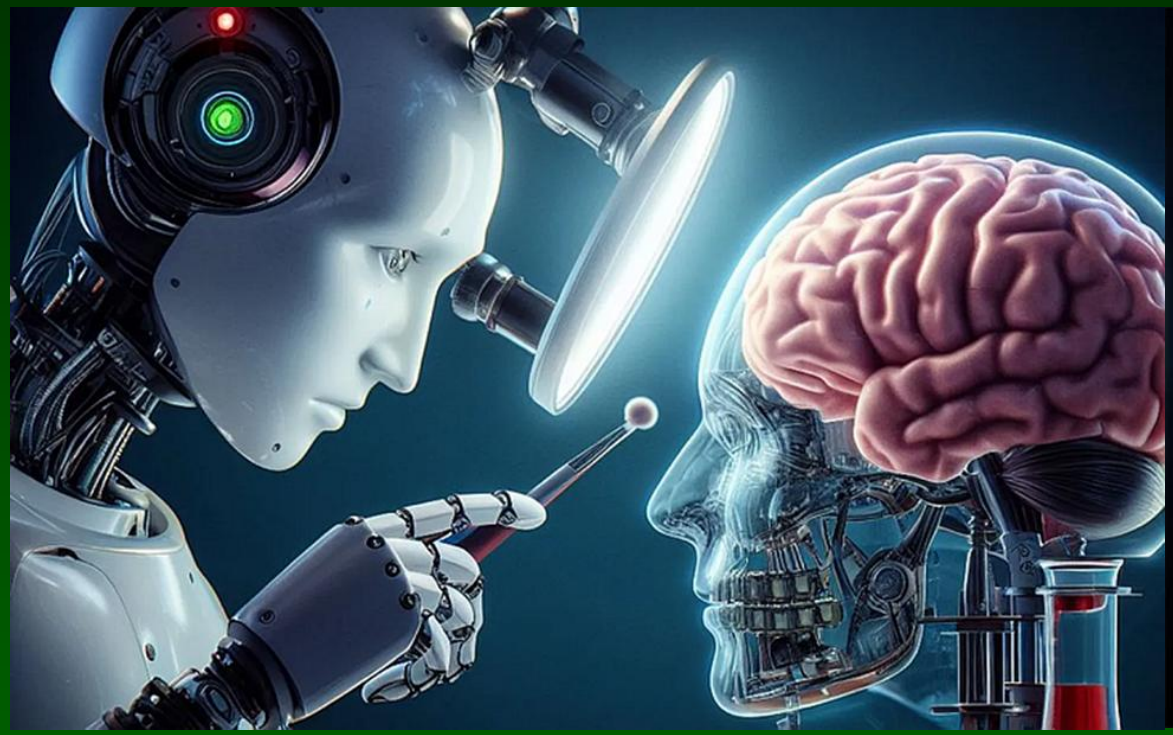
My Flipboard Magazines

Ostatnie referaty

Talks on my web page and in YouTube:

- 194. Wpływ sztucznej inteligencji na rozwój nauki, Klub Profesora, UMK, 10/25.
- 193. Cyfrowe inteligentne istoty, Inauguracja roku akademickiego WSG, Bydgoszcz, 10/25.
- 192. Medicine in times of superintelligent agents. Jesienna Szkoła Fizyki Medycznej, Centrum Onkologii, Bydgoszcz, 10/25.
- 190. Sztuczna inteligencja w roli Sherlocka Holmesa, XX Zjazd Polskiego Towarzystwa Medycyny Sądowej i Kryminologii, Bydgoszcz 09/25.
- 189. Czy agenci AI mogą nas poznać lepiej niż my sami siebie? Letnia Konferencja Psychoterapii. 08/25.
- 187. AI i ludzkie mózgi: podobieństwa i różnice. AGH, Kraków, 26.05.2025.
- 186. Physics-informed artificial intelligence for science. ISIT 2025, 09/25
- 185. Artificial Intelligence: challenges and opportunities. Instytut Łączności PIB, Warszawa, 09/25.
- 184. From statistical physics to machine learning. Smoluchowski Symposium on Statistical Physics, and Computational Neuroscience Academy (CNA 2025), 09/25
- 182. What can AI teach us about intelligence? 6th Central European Biomedical Congress, Kraków, 7/25.
- 181. Digital Intelligent Beings. Inter. Joint Conference on Neural Networks, Rome, 07/25.
- 170. From Neural Coding to Psychological Forces, Neural Coding 25, Ascona, Switzerland, 6/25.

Who is artificial?



Search: Wlodzislaw Duch

=> talks, papers, lectures, Flipboard, YouTube