Artificial intelligence and the future of science

Włodzisław Duch

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

Google: Wlodzislaw Duch

UNESCO Chair on the Ethical Governance of Artificial Intelligence, IPPT PAN, Warsaw 15.12.2025

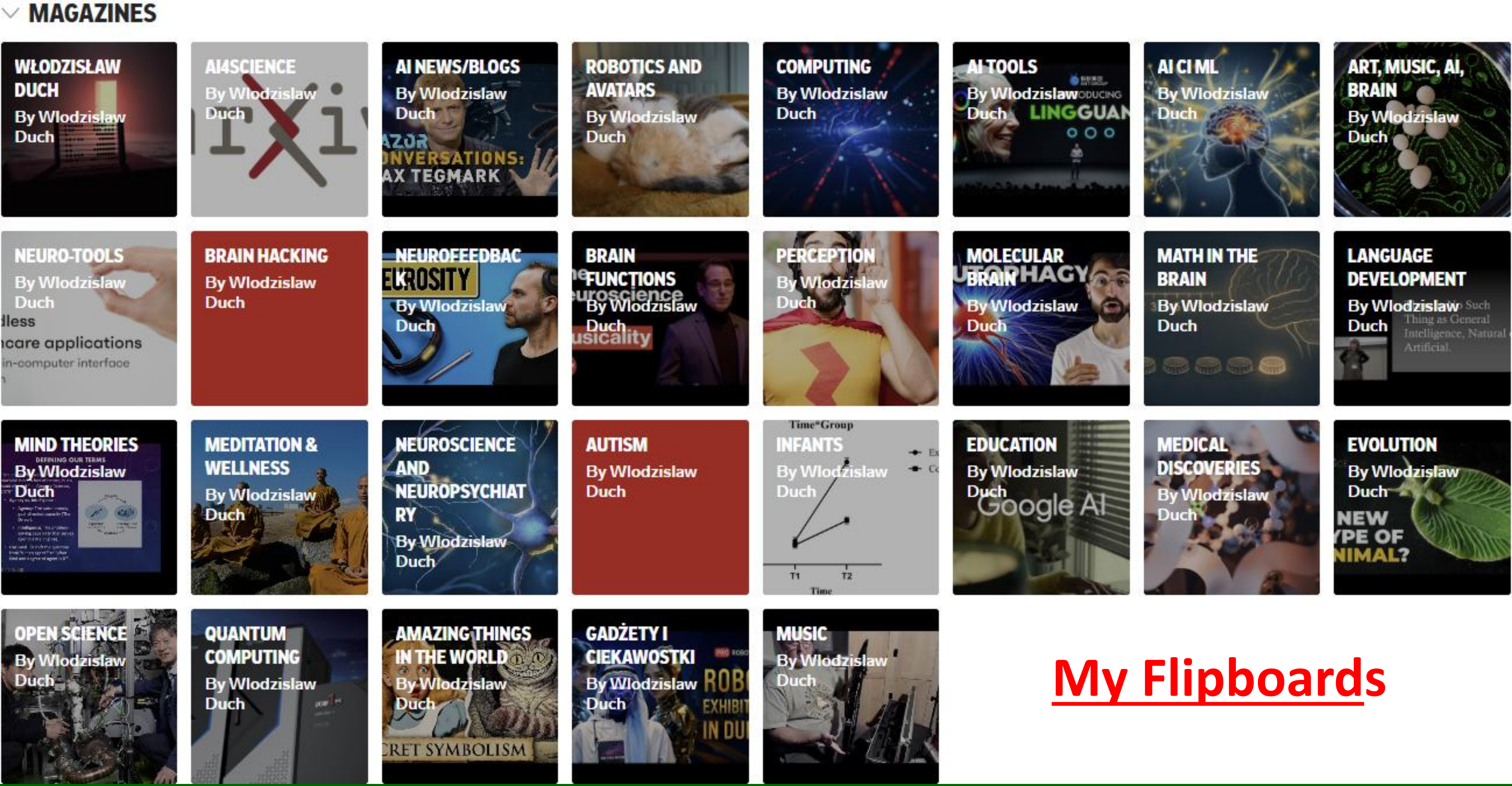
Towards autonomous ethical AI agents

1. Few thoughts on AI: computing, intelligence, cognition.
2. Superintelligence and exponential growth.
3. AI4Science – autonomous agentic AI science factories.
4. Ethical side: AI too human?

Pandora's box:
should we be more afraid that AI will turn against us,
or that we turn against AI and others will enslave us?

ChatGPT << AI, 150-300 papers in the arxiv.cs.ai each day!
NeurNIPS 2025, 15 000 papers, 5 290 accepted.

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

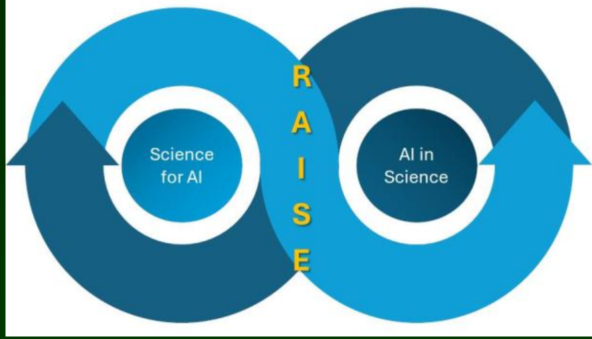


MAGAZINES

- **WŁODZISŁAW DUCH** By Wlodzislaw Duch
- **AI4SCIENCE** By Wlodzislaw Duch
- **AI NEWS/BLOGS** By Wlodzislaw Duch
- **ROBOTICS AND AVATARS** By Wlodzislaw Duch
- **COMPUTING** By Wlodzislaw Duch
- **AI TOOLS** By Wlodzislaw Duch
- **AI CI ML** By Wlodzislaw Duch
- **ART, MUSIC, AI, BRAIN** By Wlodzislaw Duch
- **NEURO-TOOLS** By Wlodzislaw Duch
- **BRAIN HACKING** By Wlodzislaw Duch
- **NEUROFEEDBACK** By Wlodzislaw Duch
- **BRAIN FUNCTIONS** By Wlodzislaw Duch
- **PERCEPTION** By Wlodzislaw Duch
- **MOLECULAR BRAIN** By Wlodzislaw Duch
- **MATH IN THE BRAIN** By Wlodzislaw Duch
- **LANGUAGE DEVELOPMENT** By Wlodzislaw Duch
- **MIND THEORIES** By Wlodzislaw Duch
- **MEDITATION & WELLNESS** By Wlodzislaw Duch
- **NEUROSCIENCE AND NEUROPSYCHIATRY** By Wlodzislaw Duch
- **AUTISM** By Wlodzislaw Duch
- **INFANTS** By Wlodzislaw Duch
- **EDUCATION** By Wlodzislaw Duch
- **MEDICAL DISCOVERIES** By Wlodzislaw Duch
- **EVOLUTION** By Wlodzislaw Duch
- **OPEN SCIENCE** By Wlodzislaw Duch
- **QUANTUM COMPUTING** By Wlodzislaw Duch
- **AMAZING THINGS IN THE WORLD** By Wlodzislaw Duch
- **GADŻETY I CIEKAWOSTKI** By Wlodzislaw Duch
- **MUSIC** By Wlodzislaw Duch

My Flipboards

EU RAISE Projects



A European Strategy for Artificial Intelligence in Science.
Paving the way for the Resource for AI Science in Europe (RAISE).

Announced at AI in Science Summit in Copenhagen on 3-4.11.2025 with pilot funding 108 M EUR.
In 2018 communication from the EU Commissison promised 20 B EUR/year -;)

RAISE is launched as a virtual European institute that pools, aligns and coordinates essential AI resources, including computational power, data, excellence and talent, and research funding, across the EU, Member States and private sector. It will support the development of AI systems beyond the current state of the art, that are ethical, explainable, transparent, accountable, reliable, safe, human-centric and aligned with human rights and societal values.

81% of researchers have concerns relating to AI models (ethics, accuracy, security/privacy, and/or lack of transparency) and 63% to the lack of guidelines, hindering the adoption of AI (survey . The Commission will continue to promote the “ethics by design” approach and develop different resources with the scientific community (training, tools, etc.).

The European Group on Ethics in Science and New Technologies (EGE) will evaluate this program.

Nauka o AI, AI dla nauki.

Defining intelligence

What is really important: not how we call it but what it can do!

If we have **effective** algorithms we can write programs, no need for AI.

No effective algorithms? Intelligence is needed.

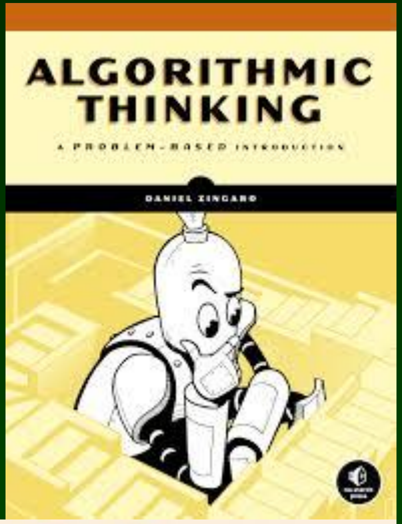
Intelligence: ability to solve problems where no effective algorithms are known. AI does it using some hardware, humans use wetware.

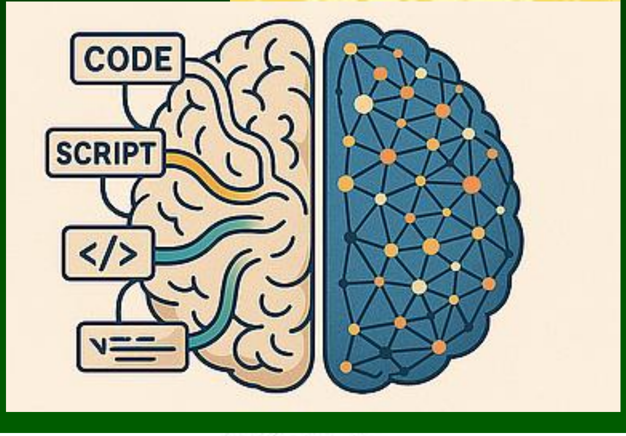
What cannot be successfully programmed but can be learned from experience may be replicated in a **supervised** way, discover in **unsupervised** way, or strategies may be developed using **reinforcement learning**.

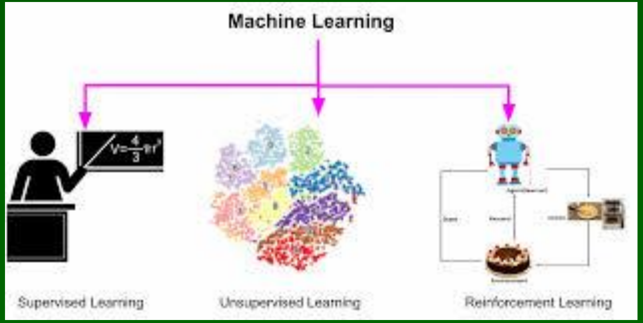
Humans are megalomaniacs and believe in magic in their wetware, but understanding == compression of complexity into coherent patterns. Transferring disorganized data => structure.

AI cannot always be perfect, but it can be better than humans.

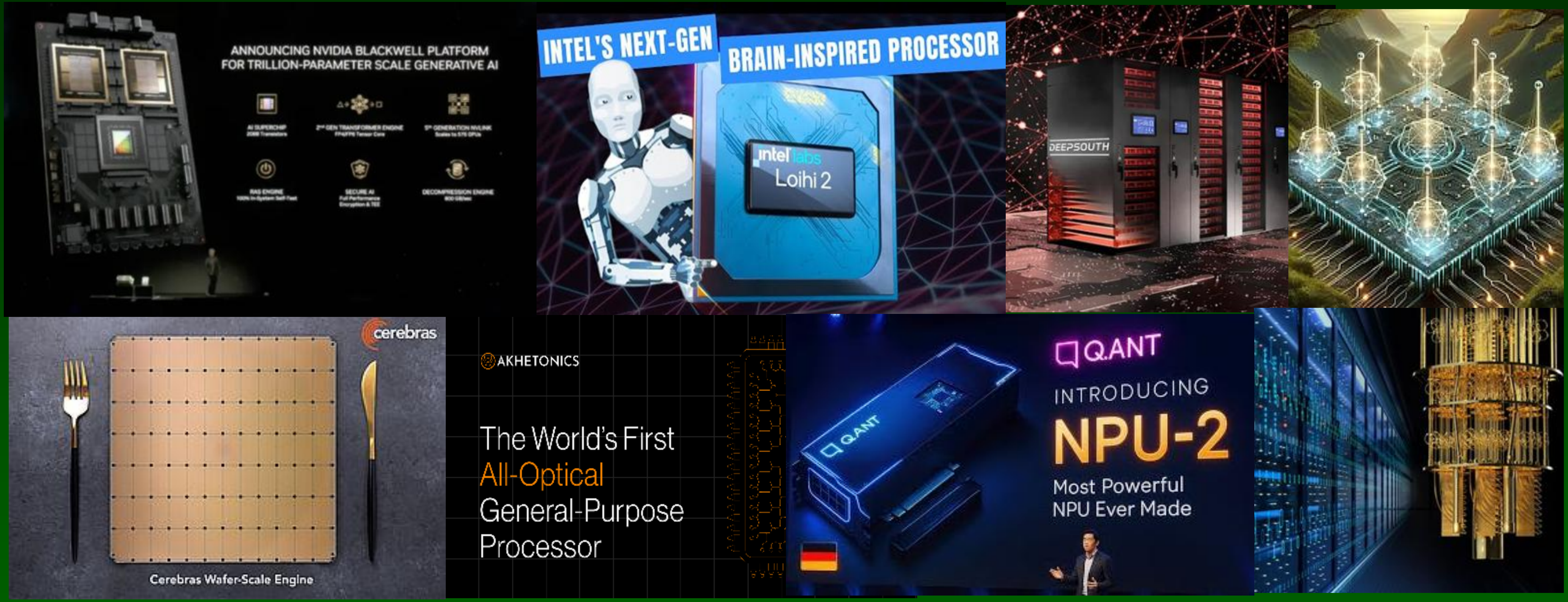
Hardware could be biological, electronic, photonic, spintronic, atomic …





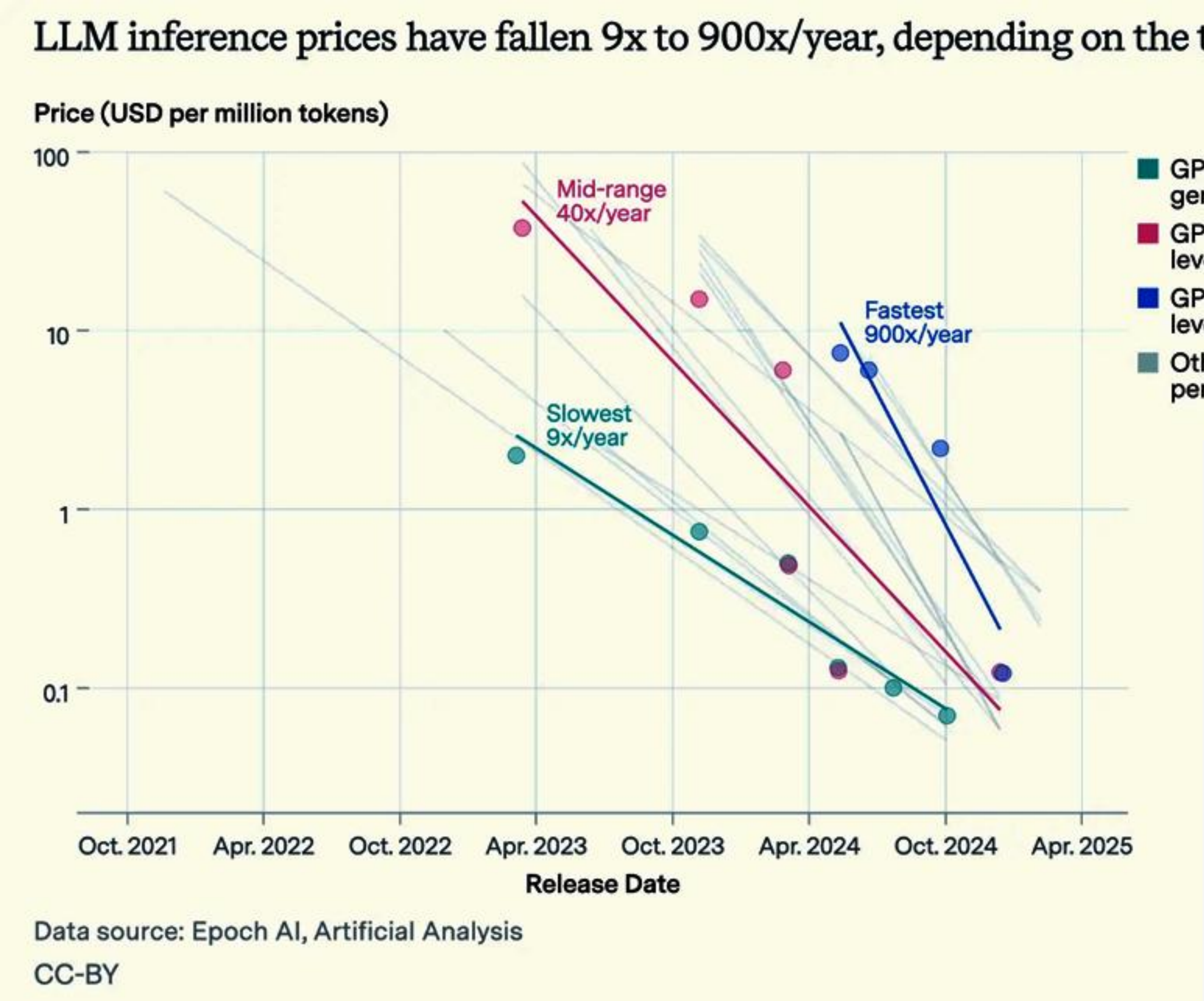


More substrates for thinking



New hardware has no constraints of wetware. 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. Cerebras, CIMs, spintronics, neuromorphic computing, Cortical Labs organoids CL1 800k cells chip.
LLM inference cost fall between 9-900x/year, depending on the task (Epoch.ai).
Energy may soon not be a problem. Should our AI factories use classical hardware?

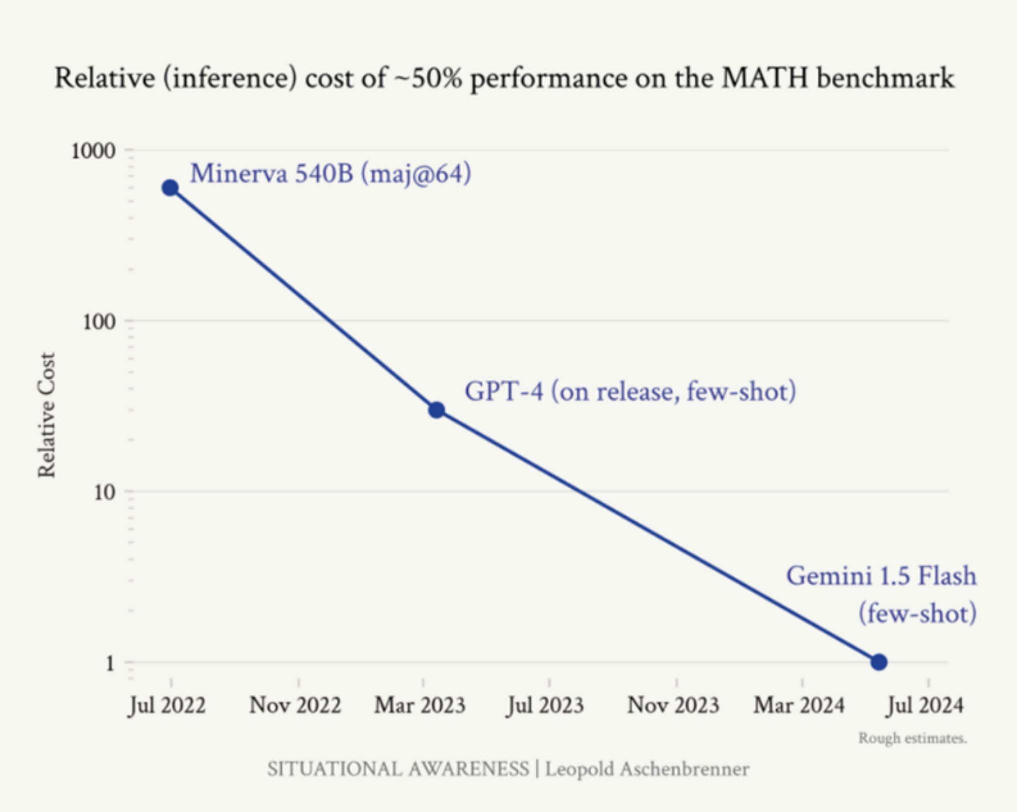
Who needs GWs?



LLM inference prices have fallen 9x to 900x/year, depending on the task

Data source: Epoch AI, Artificial Analysis

CC-BY



Relative (inference) cost of ~50% performance on the MATH benchmark

SITUATIONAL AWARENESS | Leopold Aschenbrenner

Embedding + attention = transformers

Word2Vec, Glove, FastText: embedding of words as vectors in the parameter space, preserving similarity.

Each word is encoded in a large number of contexts.

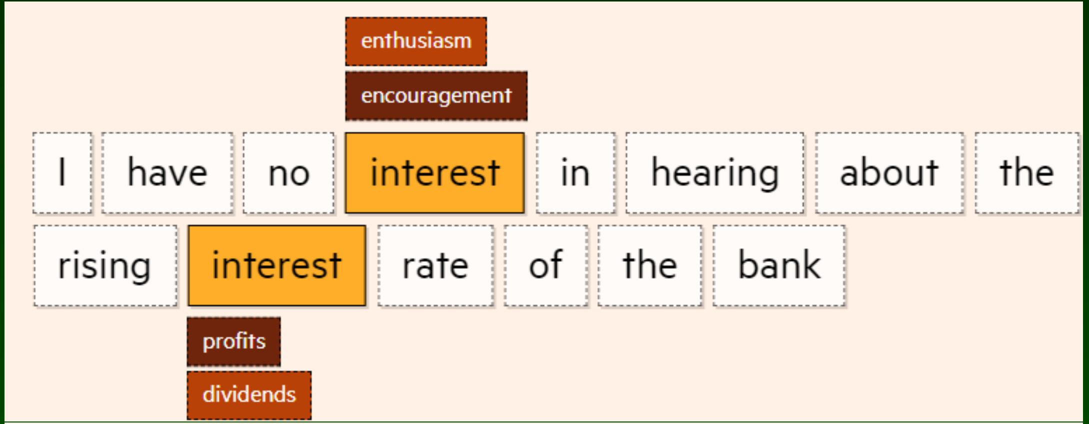
Transformer model published by Google in June 2017: Attention is all you need, started the generative AI era.

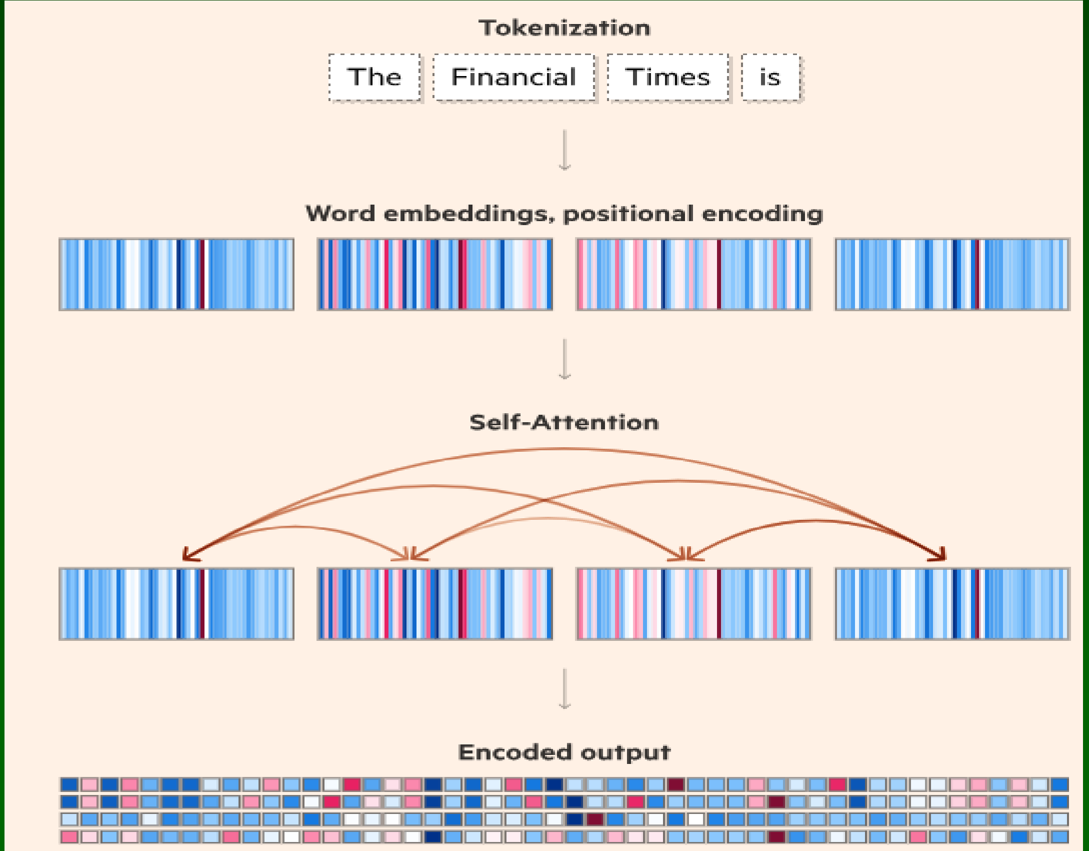
A key concept of GPT is self-attention to relationships between tokens, linking each token to other tokens.

Vectors of concepts that are semantically related become similar to each other
=> GPT interprets inputs using meaningful associations.

Associative thinking: creative, heuristics for reasoning.

Matykiewicz P, Pestian J, Duch W, and Johnson N. (2006) *Unambiguous Concept Mapping in Radiology Reports: Graphs of Consistent Concepts*, AMIA Ann. Symp Proc. 1024.





Transformers and spreading activation

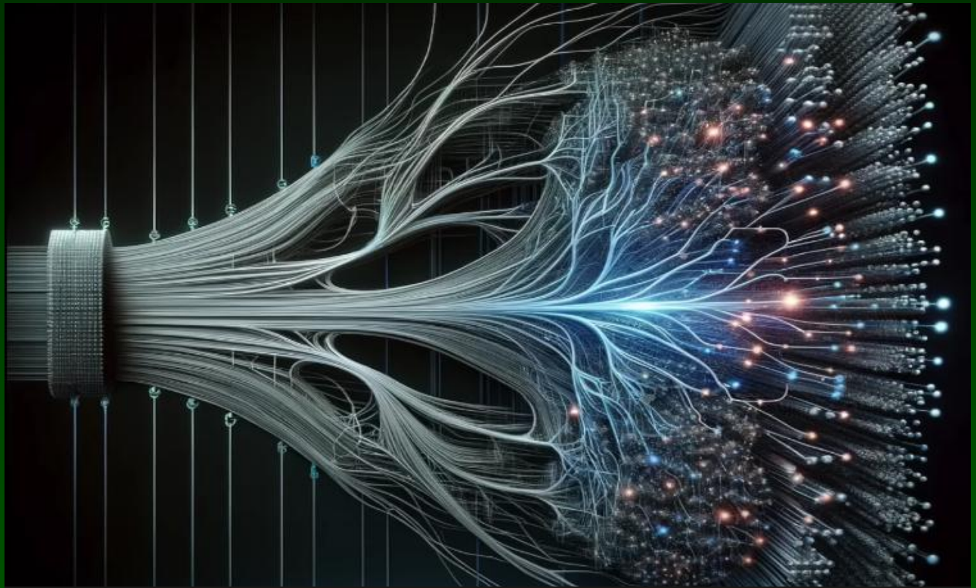
Predictive AI: search + heuristics, supercalculator.

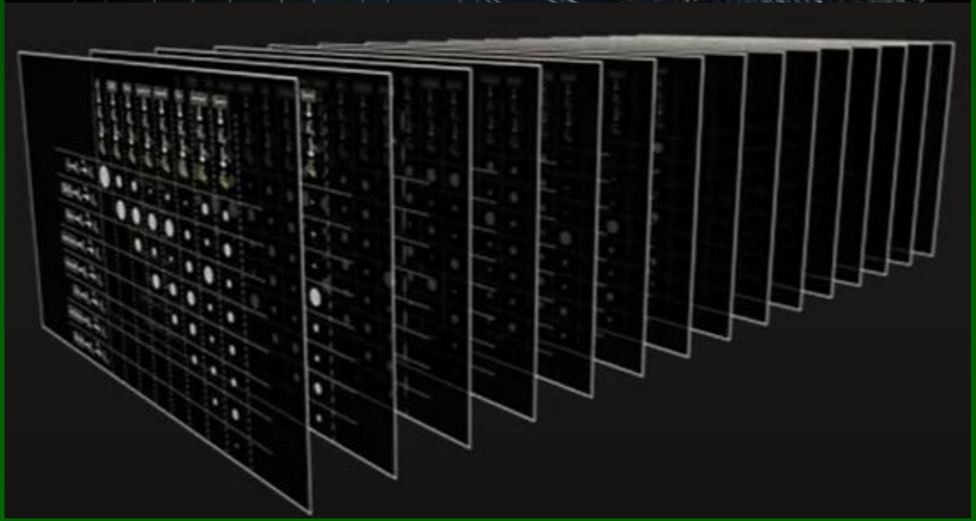
Generative AI: spreading activation networks, binding relevant information.

GPT = Generative Pre-trained Transformer

Input data ⇒ tokens ⇒ embed in vector space, preserve similarity of meaning (semantics) in different contexts.

Analyze relations between tokens, create structured knowledge graphs.

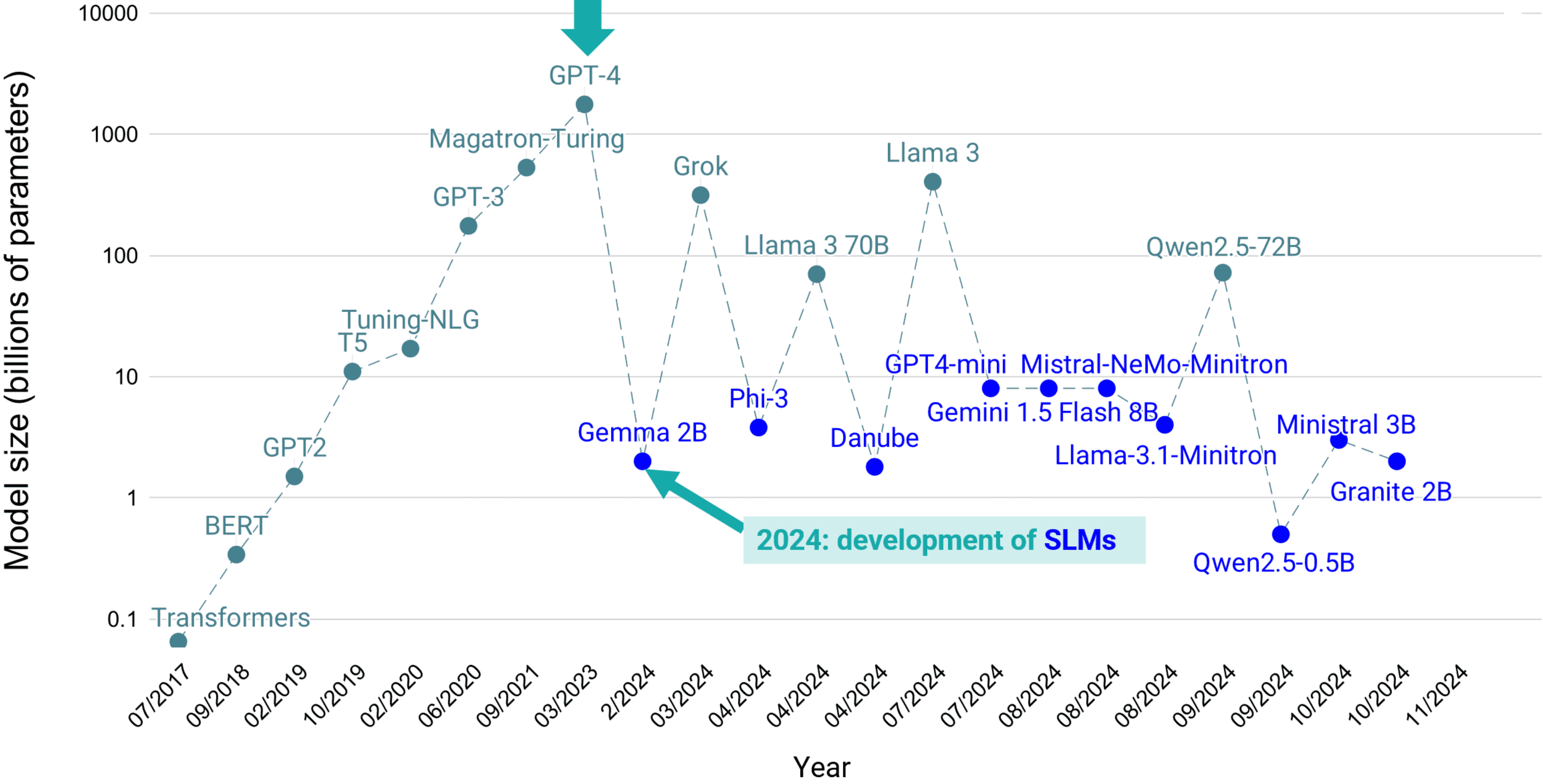




LLM visualization https://bbycroft.net/llm

Over 51 565 variants of open-source LLMs, over 3000 for fine-tuning (LLM Explorer, 14/12/25).

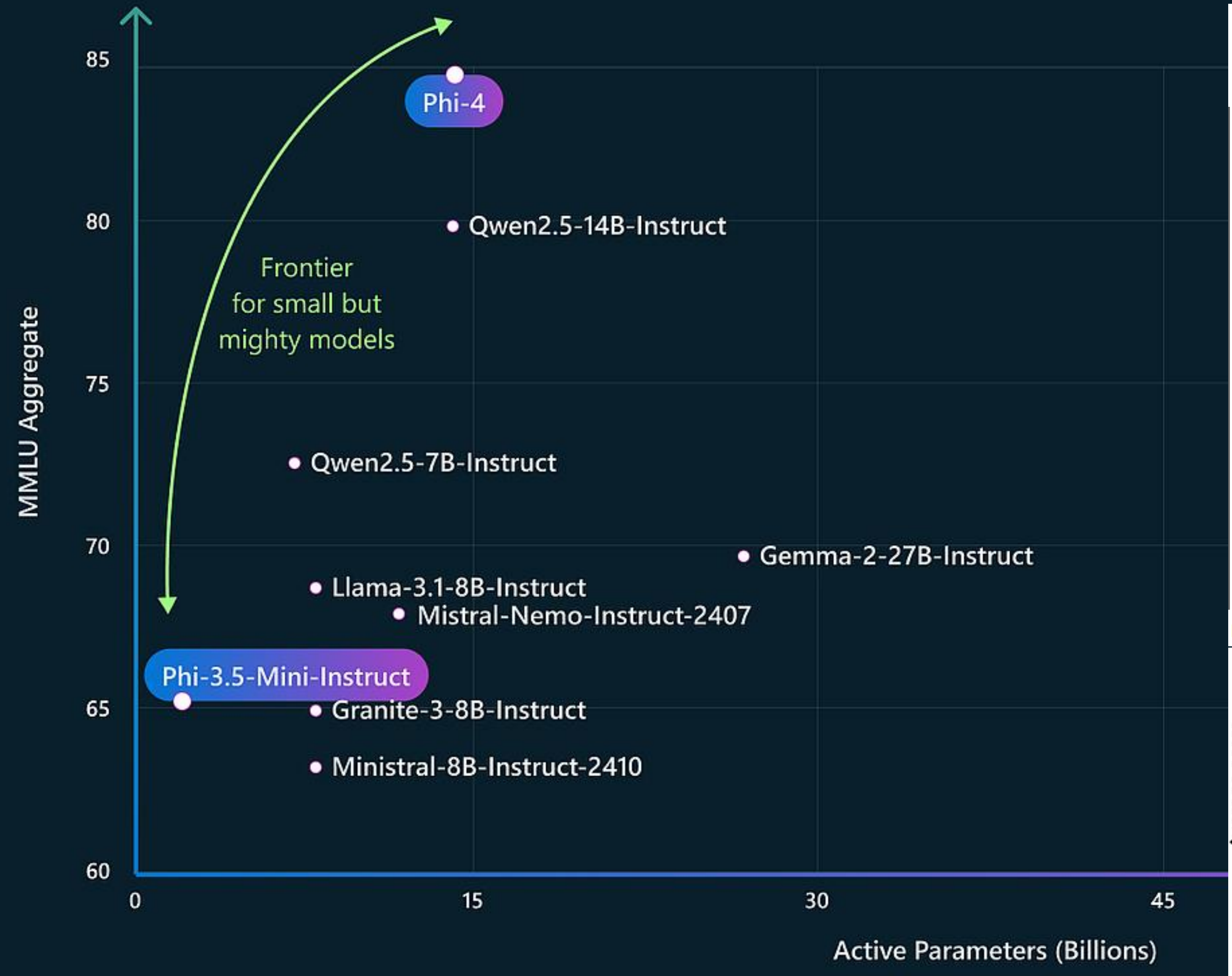
Processing clinical text with domain-specific spreading activation methods. US Patent US8930178B2 (1/2015)
Duch W, Matykiewicz P, Pestian J, Towards Understanding of Natural Language: Neurocognitive Inspirations. 2007

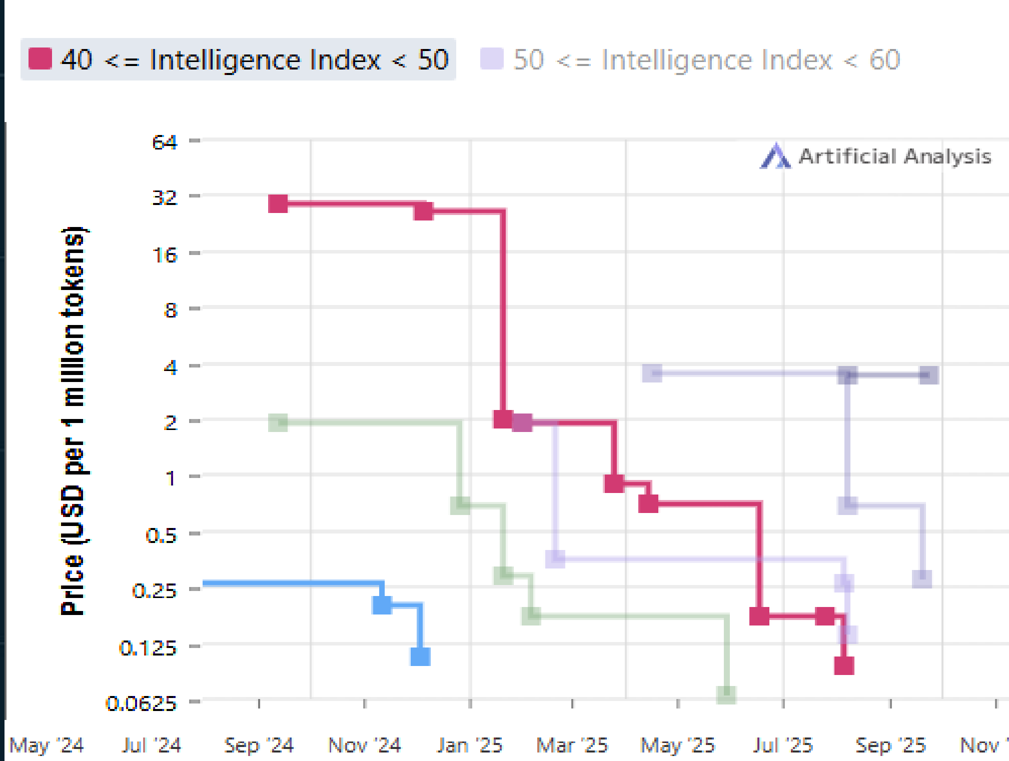


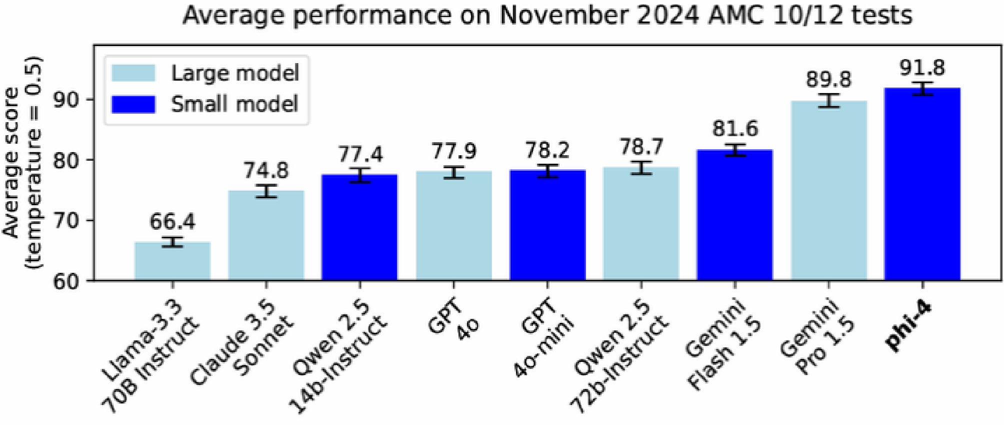
Exponential growth of LLMs stopped

2024: development of SLMs

Phi-4 offers high quality results at a small size







Prompt optimization

GPTs/LLMs need Medprompt to achieve SOTA performance on a wide range of medical problems.
Microsoft Research tools for thought, promptions. Best prompt libraries.

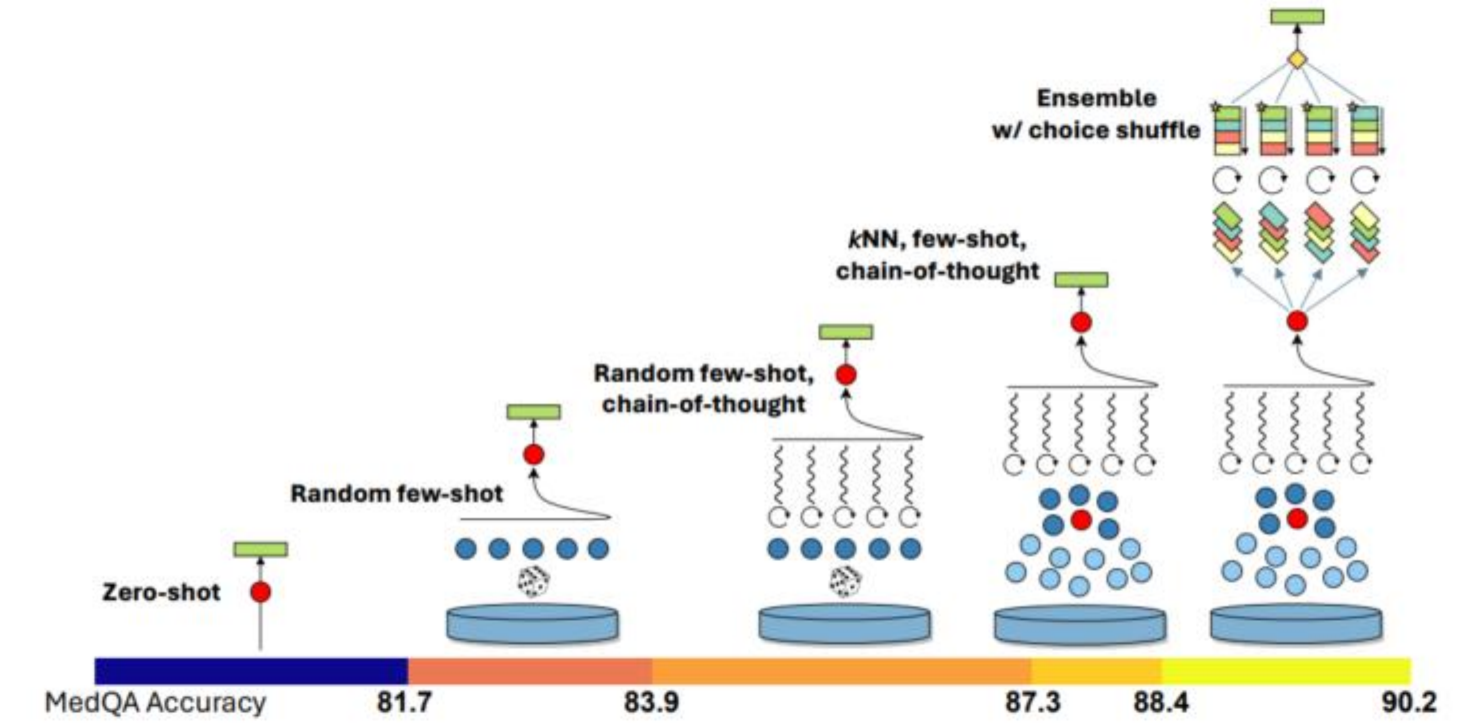
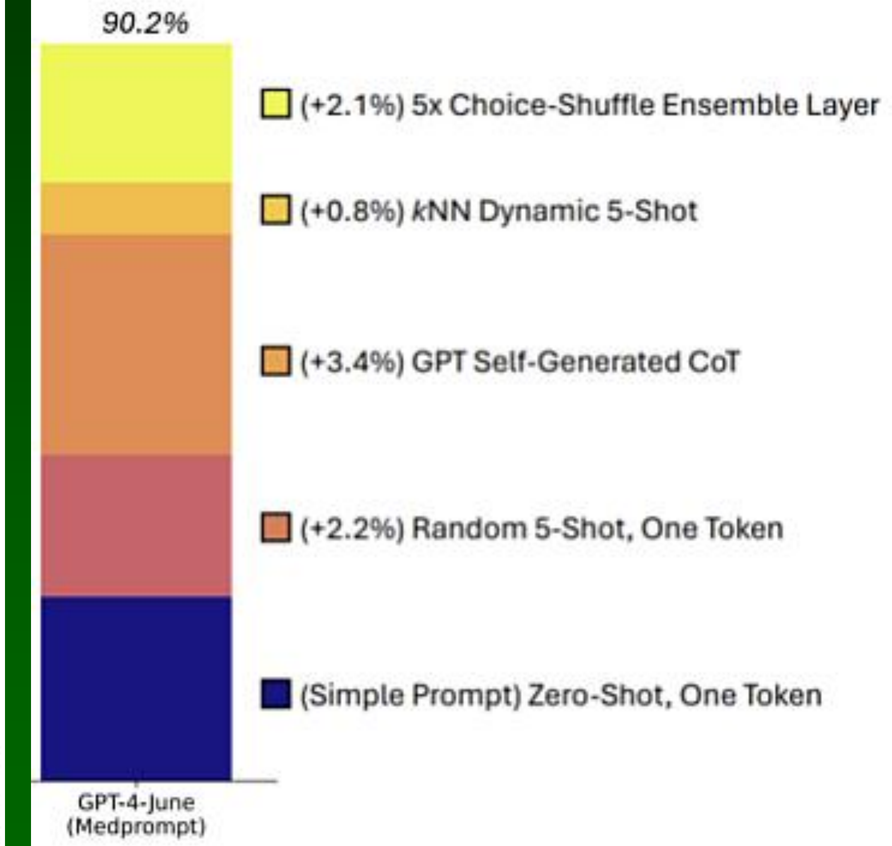


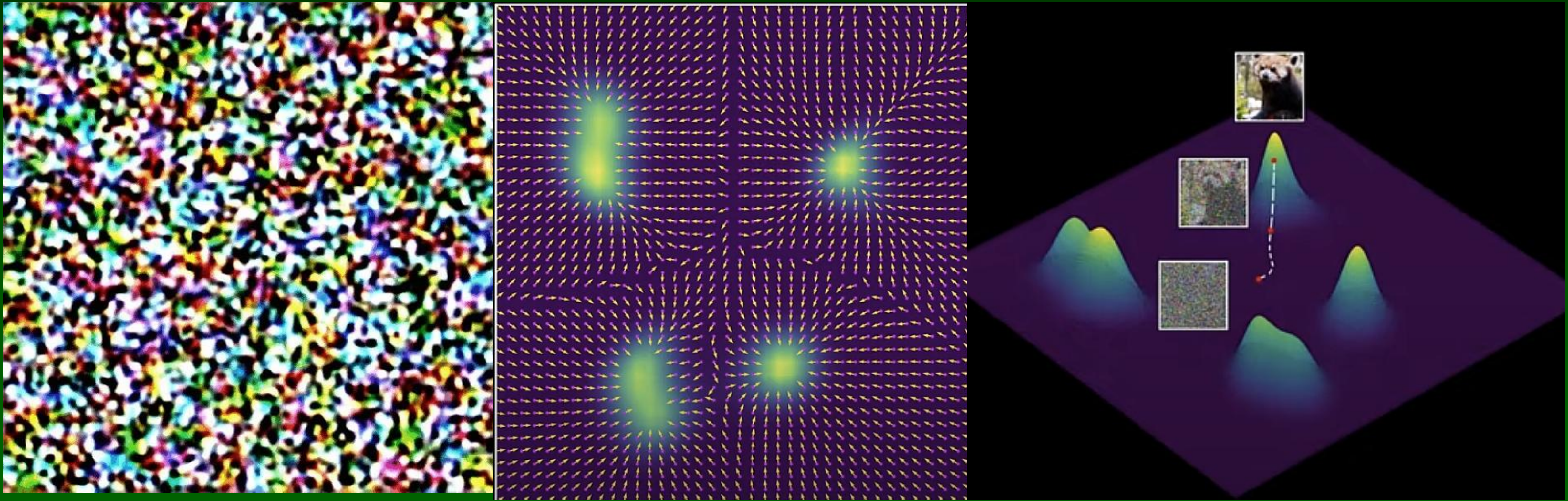
Figure 4: Visual illustration of Medprompt components and additive contributions to performance on the MedQA benchmark. The prompting strategy combines *k*NN-based few-shot example selection, GPT-4–generated chain-of-thought prompting, and answer-choice shuffled ensembling (see details in



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

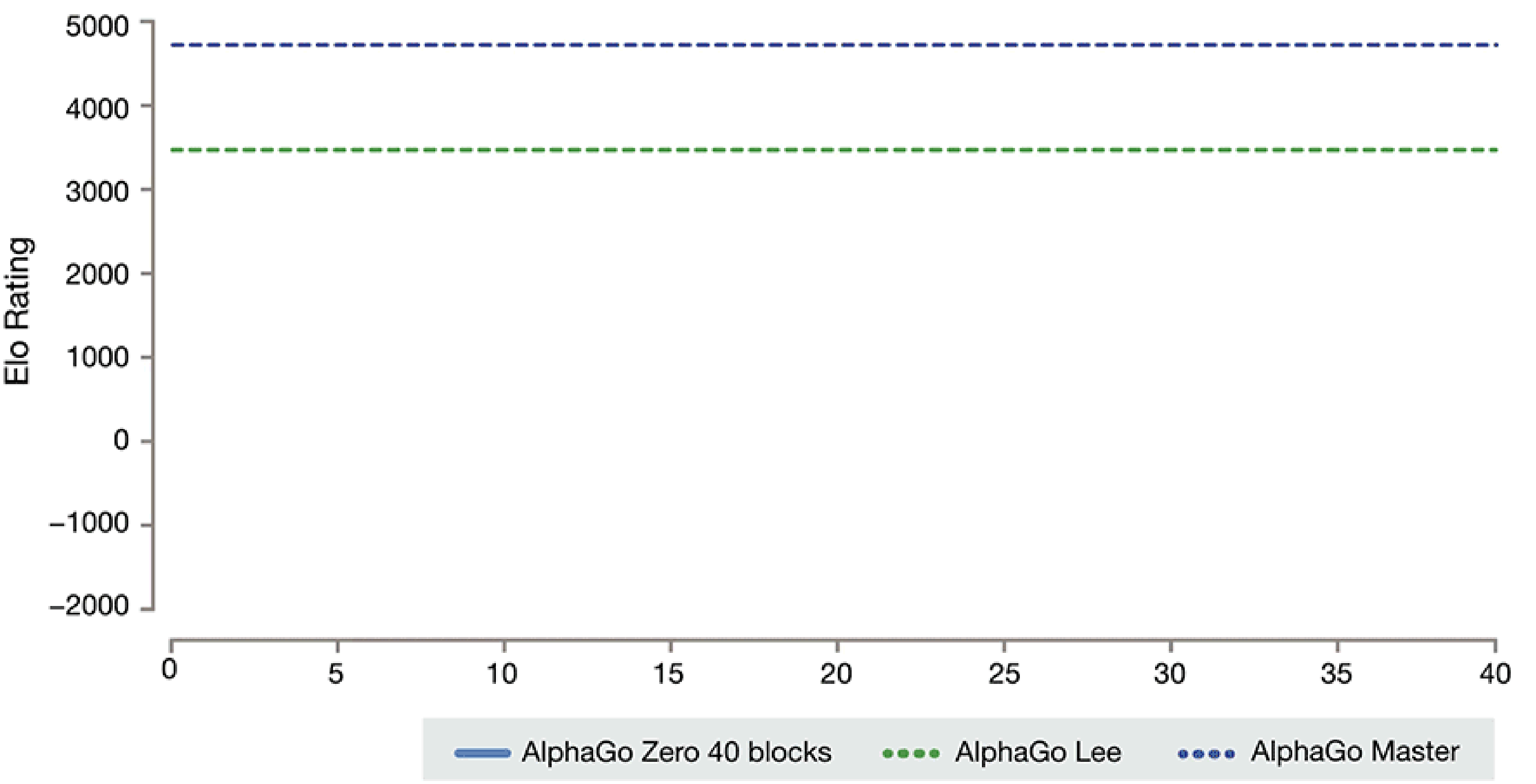
Diffusion models

Almost all image/audio/video systems are based on diffusion. A few sentence generation are based on continuous diffusion. 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.

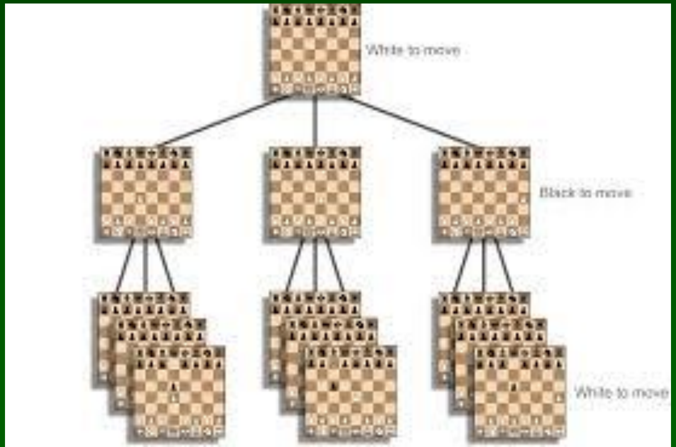


See videos at the @depth-first YT channel. Diffusion has deep roots in statistical physics.
Nobelization of Neural Networks: Deep Roots and Insane Future of Neural Networks (INNS 12/2024).

AlphaGo Zero and verifiable rewards (RLVR)

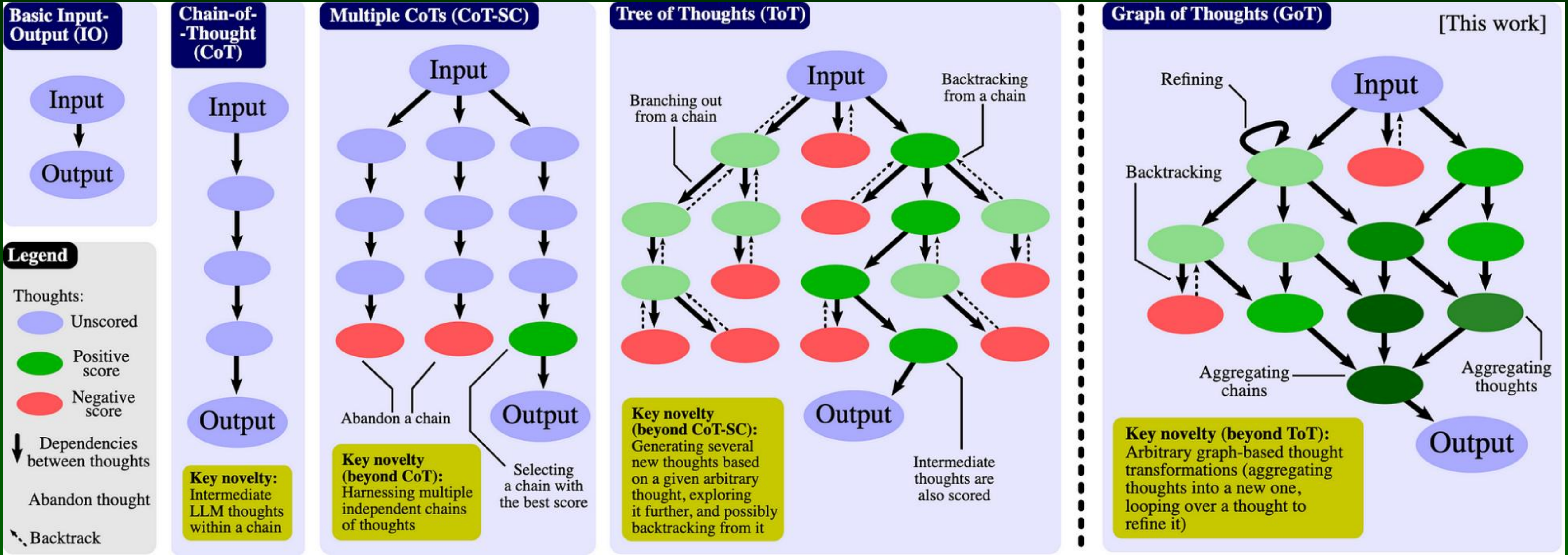


1997, Deep Blue-Kasparov.
2016, AlphaGo-Lee Sedol
2017, Alpha GoZero
2017, Poker, Dota 2;
2019, Starcraft II,
2022 Stratego, Diplomacy



Superhuman level in Go. Best humans 3,600–3,860 ELO, KataGO over 6000 ELO, playing against itself!
Human knowledge becomes irrelevant, decreasing AI competence! Can we do it in other domains?
Reinforcement learning with Verifiable Rewards, RLVR continuous self-improvement.
Shocking news: Ruoss … & Genewein, T. (2/2024). *Grandmaster-Level Chess Without Search*.
270M parameter transformer model, 1-step intuitive decisions! Like in Blitz Chess.

Reasoning – cognitive inspirations



Key novelty: Intermediate LLM thoughts within a chain

Key novelty (beyond CoT): Harnessing multiple independent chains of thoughts

Key novelty (beyond CoT-SC): Generating several new thoughts based on a given arbitrary thought, exploring it further, and possibly backtracking from it

Key novelty (beyond ToT): Arbitrary graph-based thought transformations (aggregating thoughts into a new one, looping over a thought to refine it)

Associations: input => output.

This is GOFAI! LLM associations in heuristic search.

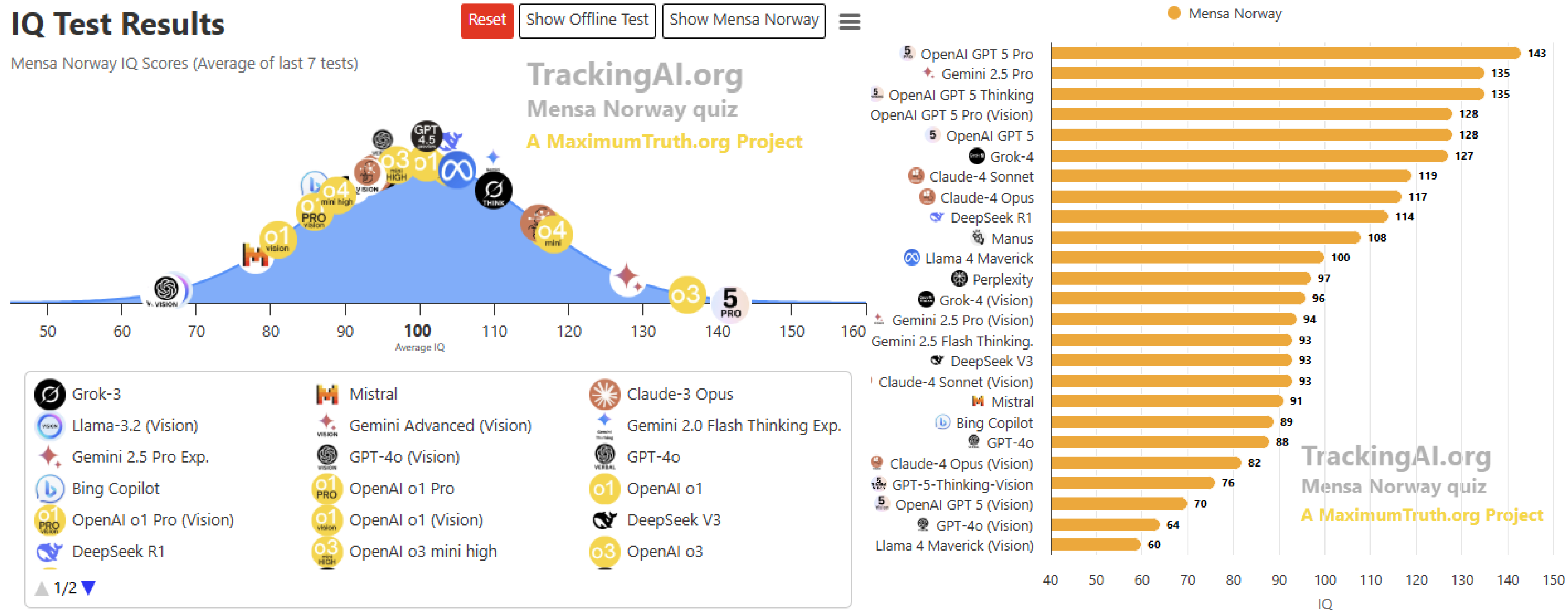
Chain of thought (CoT): step by step.

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

Graph of Thoughts (GoT): like human reasoning.

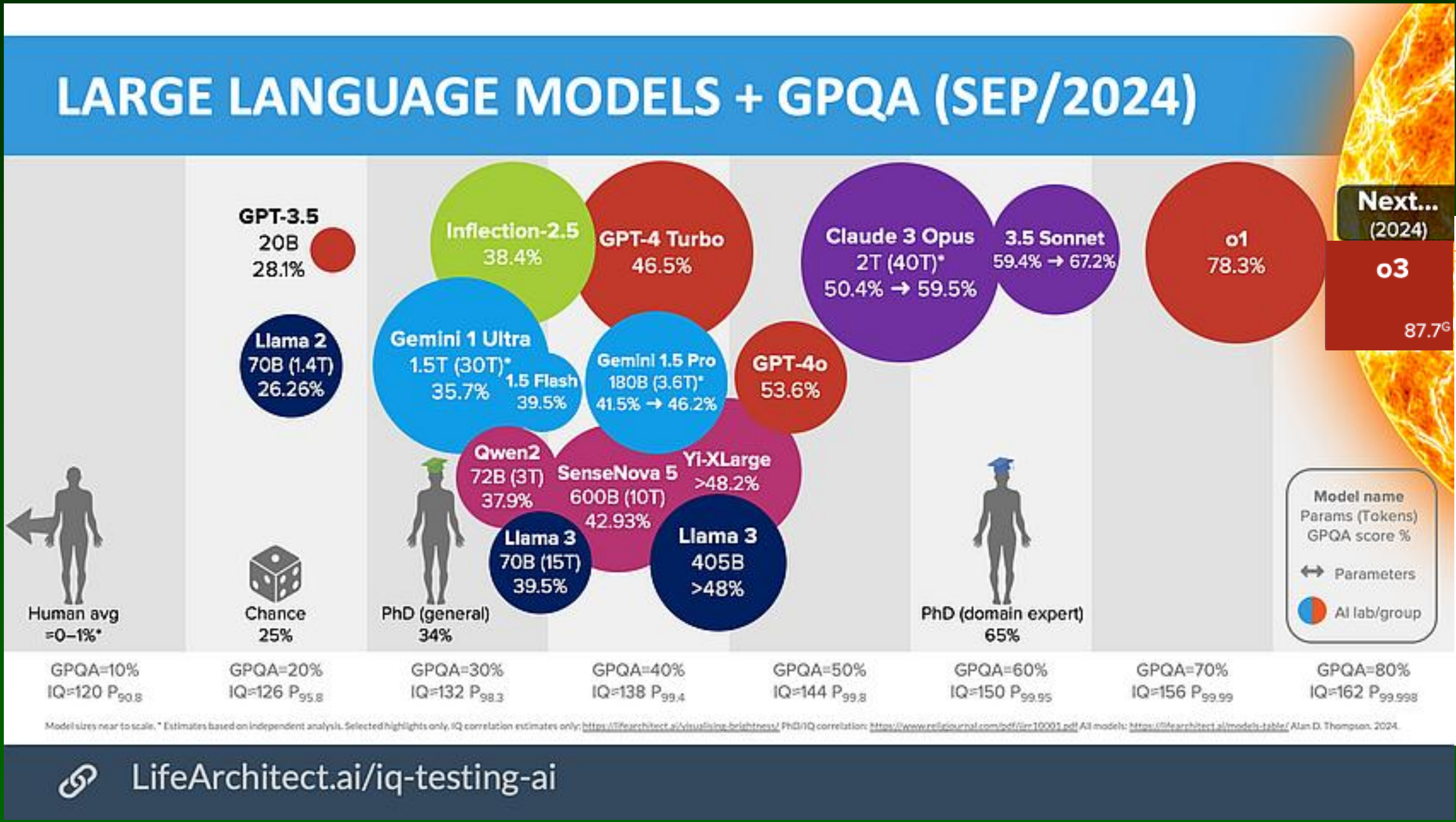
Sketch-of-Thought: adaptive cognitive-Inspired sketching

MIQ of AI



At the beginning of the 2024 highest results were < 100 points. Gemini 3 Pro, GPT 5.1 Pro > 140. 29.09.2025: Claude Sonnet 4.5 shows 30+ hours of autonomous coding.

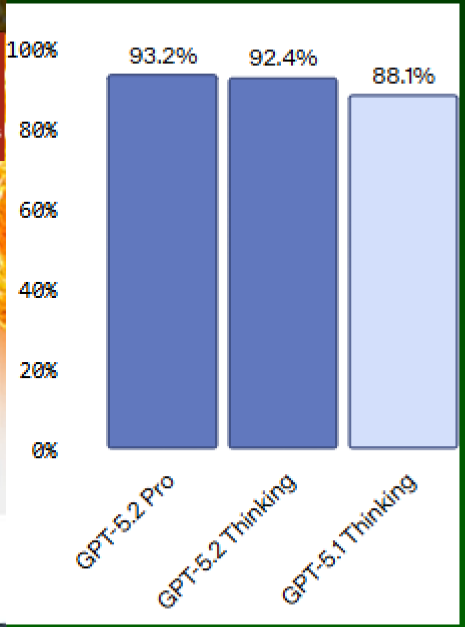
GPQA benchmark



Model sizes near to scale. * Estimates based on independent analysis. Selected highlights only. IQ correlation estimates only: https://lifearchitect.ai/visualising-brightness/ PhD/IQ correlation: https://www.religjournal.com/pdf/ijrr10001.pdf All models: https://lifearchitect.ai/models-table/ Alan D. Thompson. 2024.

LifeArchitect.ai/iq-testing-ai

93.8% Gemini 3 DT



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)

9/2025: teams from 139 universities in 103 countries participated in the ICPC finals.

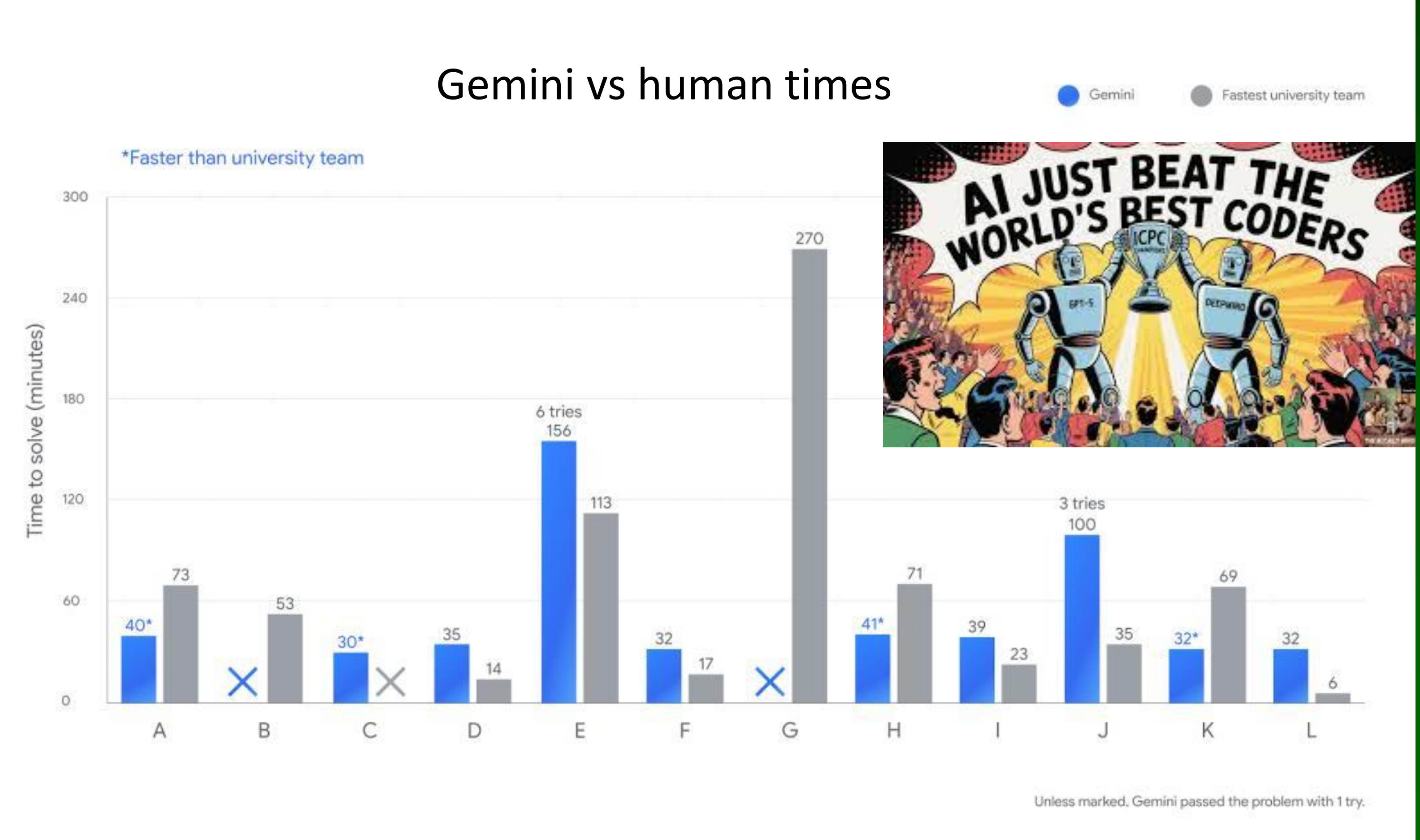
Team: 3 students. Task: solve 12 algorithmic problems within 5 hours. Cutting-edge algorithms: graph theory, number theory, dynamic programming, combinatorial optimization, network flow ...

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 all 12 problems, 11 on the first try, last on 9th submission, 12/12.

Gemini 2.5 Deep Think solved 10/12, including problem that no human team could solve.



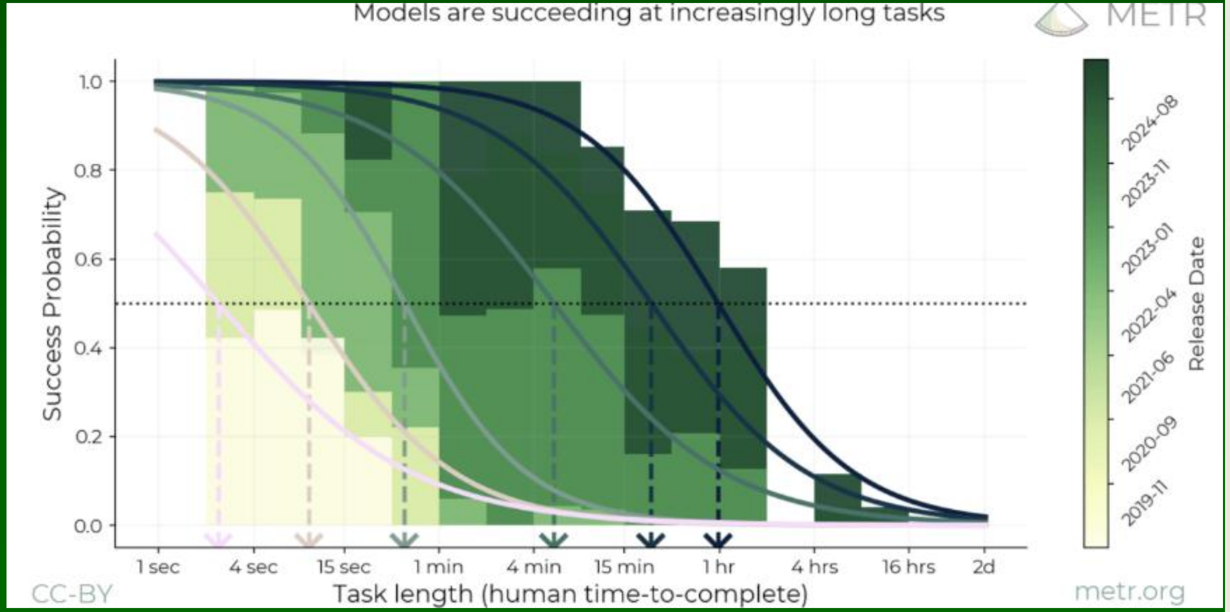
Long tasks

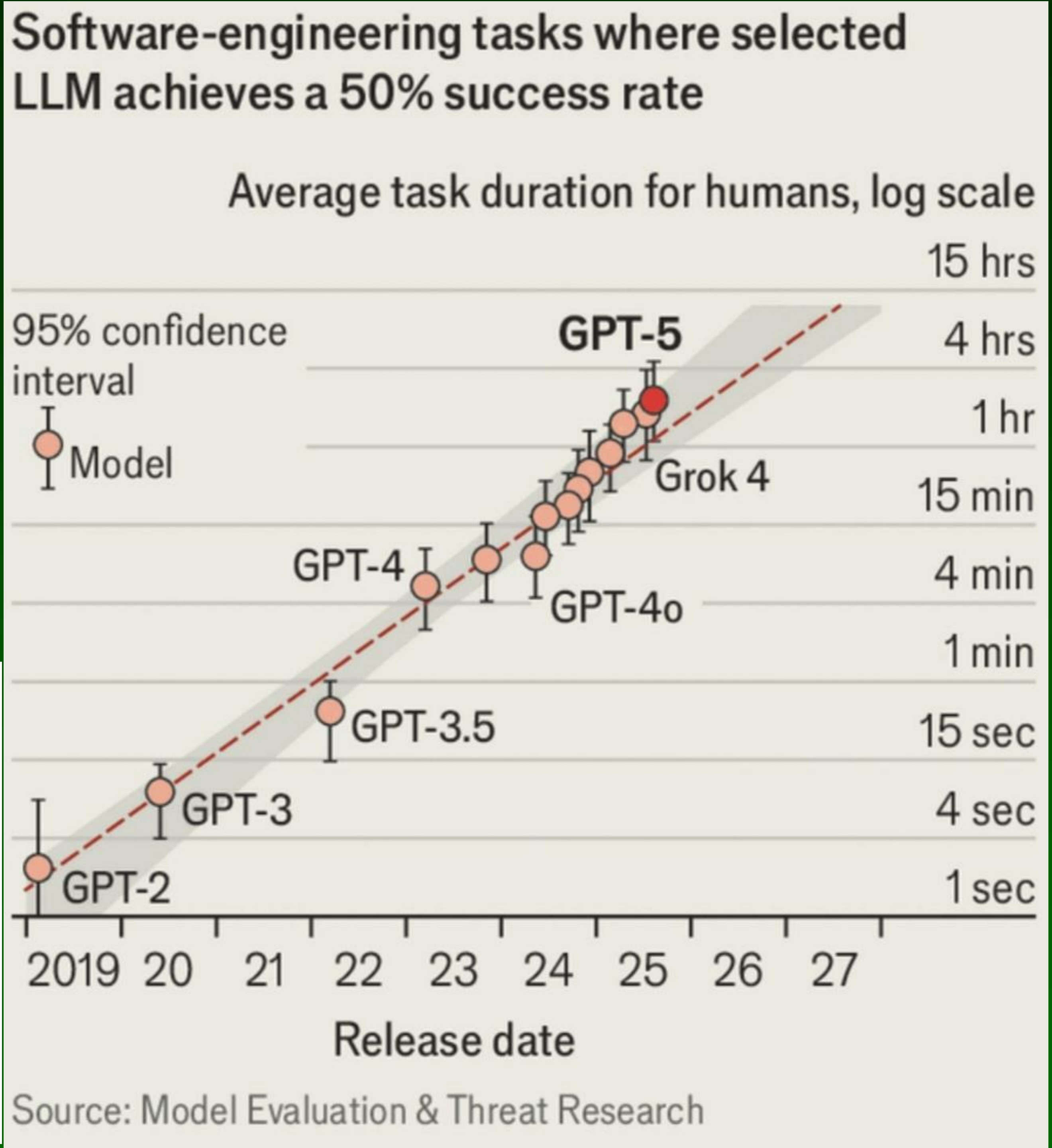
Measuring AI Ability to Complete Long Tasks.

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 less than 7 months.

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

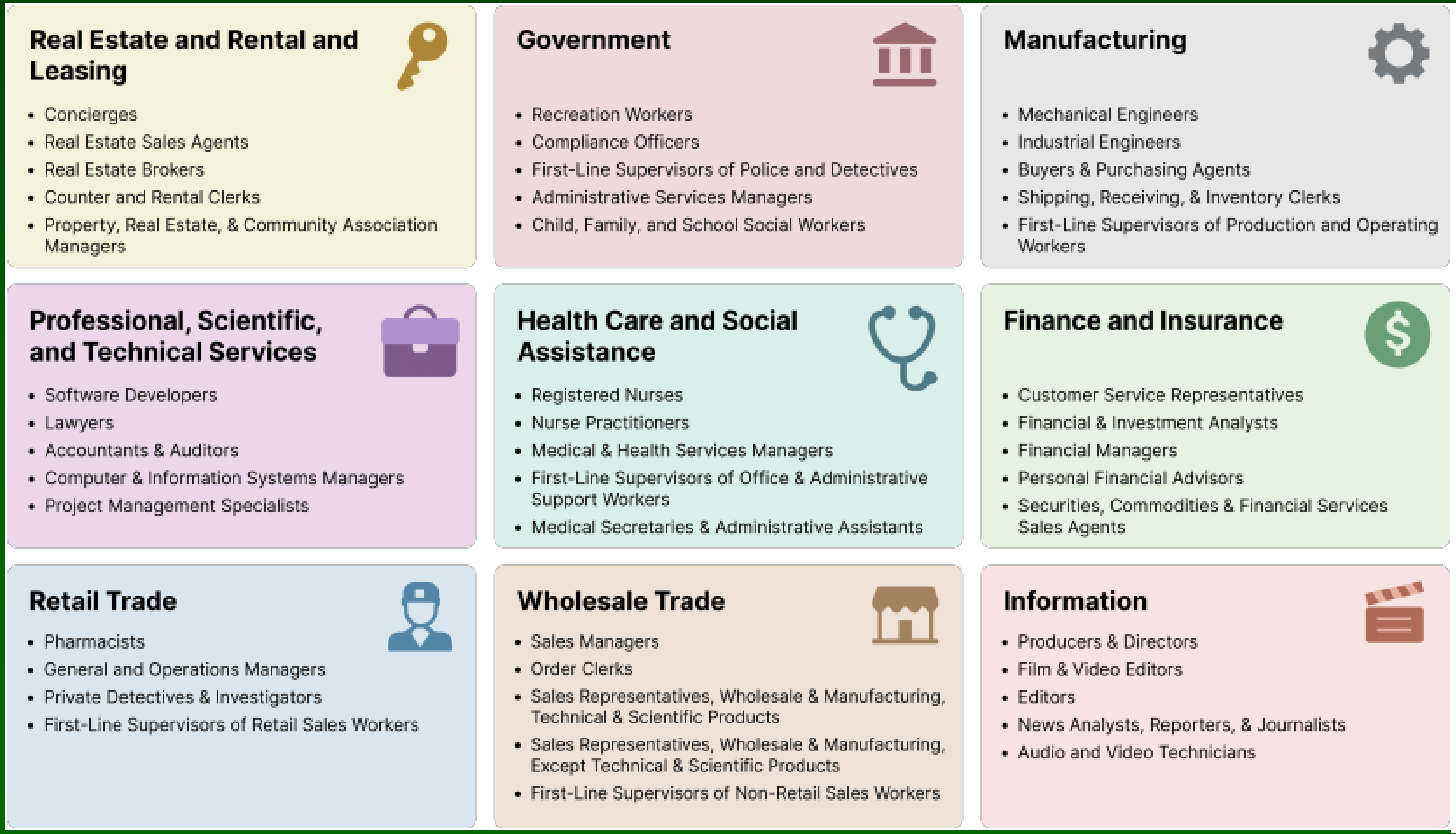

Models are succeeding at increasingly long tasks


Software-engineering tasks where selected LLM achieves a 50% success rate

Source: Model Evaluation & Threat Research

GDPval real work test

Most benchmarks do not show real value of AI models.
GDPval is based on real-world work from 44 occupations.
In pairwise comparisons some AI models approach parity with industry experts (av. 14 years experience) on the GDPval gold subset (220 tasks). Win rate of GPT 5.2 is 60%, with ties 74%.



Real Estate and Rental and Leasing
- Concierges
- Real Estate Sales Agents
- Real Estate Brokers
- Counter and Rental Clerks
- Property, Real Estate, & Community Association Managers

Government
- Recreation Workers
- Compliance Officers
- First-Line Supervisors of Police and Detectives
- Administrative Services Managers
- Child, Family, and School Social Workers

Manufacturing
- Mechanical Engineers
- Industrial Engineers
- Buyers & Purchasing Agents
- Shipping, Receiving, & Inventory Clerks
- First-Line Supervisors of Production and Operating Workers

Professional, Scientific, and Technical Services
- Software Developers
- Lawyers
- Accountants & Auditors
- Computer & Information Systems Managers
- Project Management Specialists

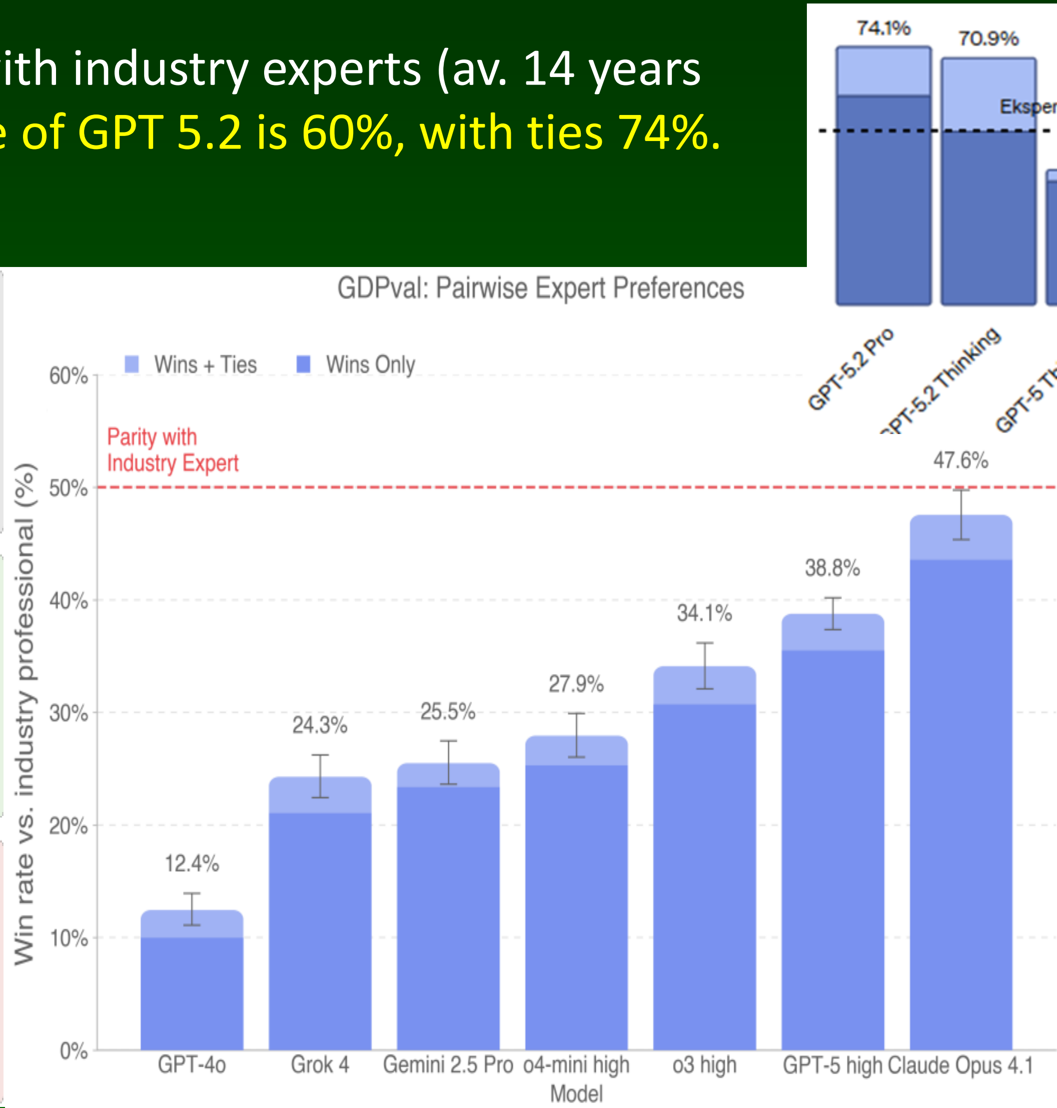
Health Care and Social Assistance
- Registered Nurses
- Nurse Practitioners
- Medical & Health Services Managers
- First-Line Supervisors of Office & Administrative Support Workers
- Medical Secretaries & Administrative Assistants

Finance and Insurance
- Customer Service Representatives
- Financial & Investment Analysts
- Financial Managers
- Personal Financial Advisors
- Securities, Commodities & Financial Services Sales Agents

Retail Trade
- Pharmacists
- General and Operations Managers
- Private Detectives & Investigators
- First-Line Supervisors of Retail Sales Workers

Wholesale Trade
- Sales Managers
- Order Clerks
- Sales Representatives, Wholesale & Manufacturing, Technical & Scientific Products
- Sales Representatives, Wholesale & Manufacturing, Except Technical & Scientific Products
- First-Line Supervisors of Non-Retail Sales Workers

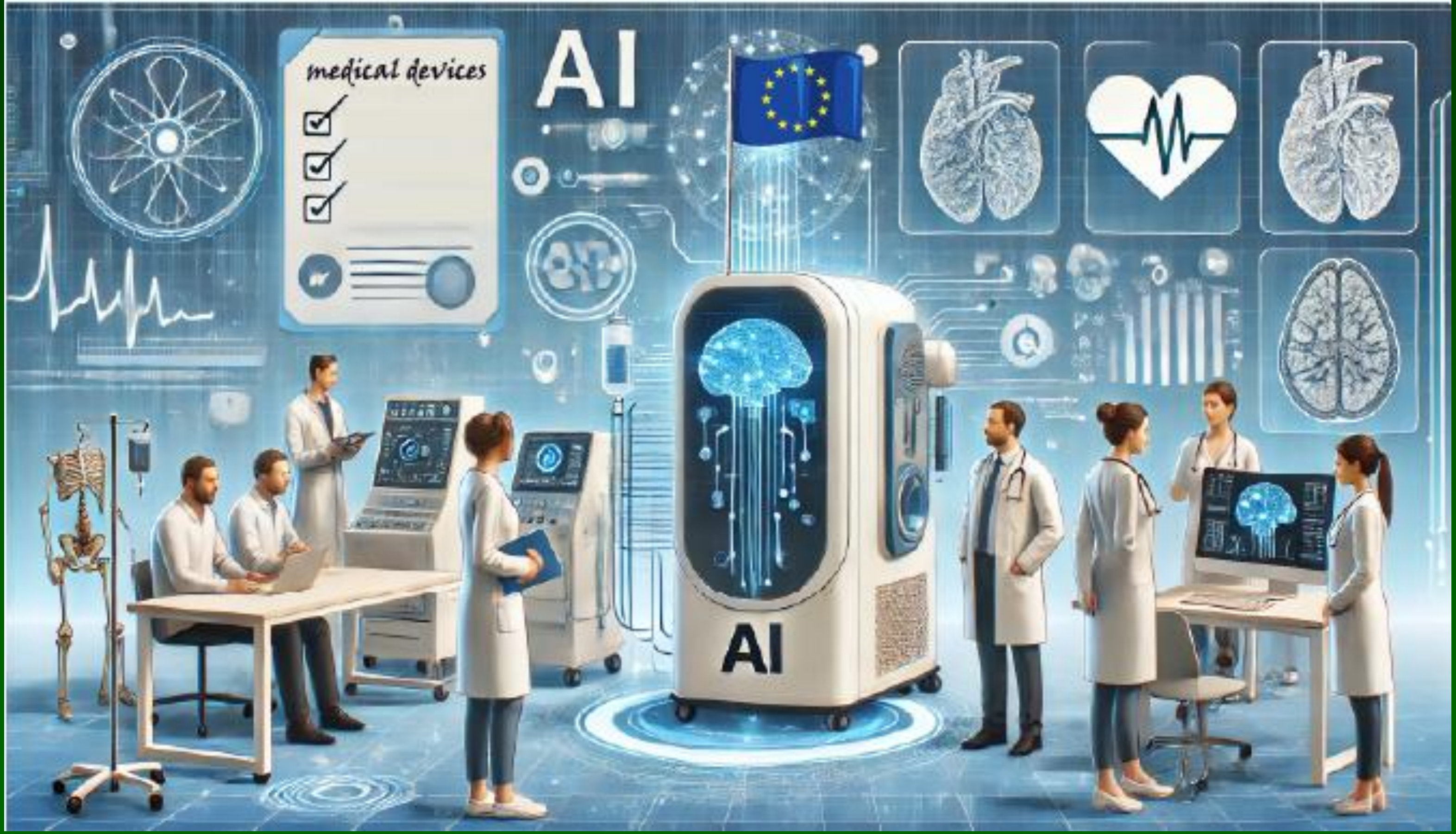
Information
- Producers & Directors
- Film & Video Editors
- Editors
- News Analysts, Reporters, & Journalists
- Audio and Video Technicians



Artificial Intelligence-Enabled Medical Devices

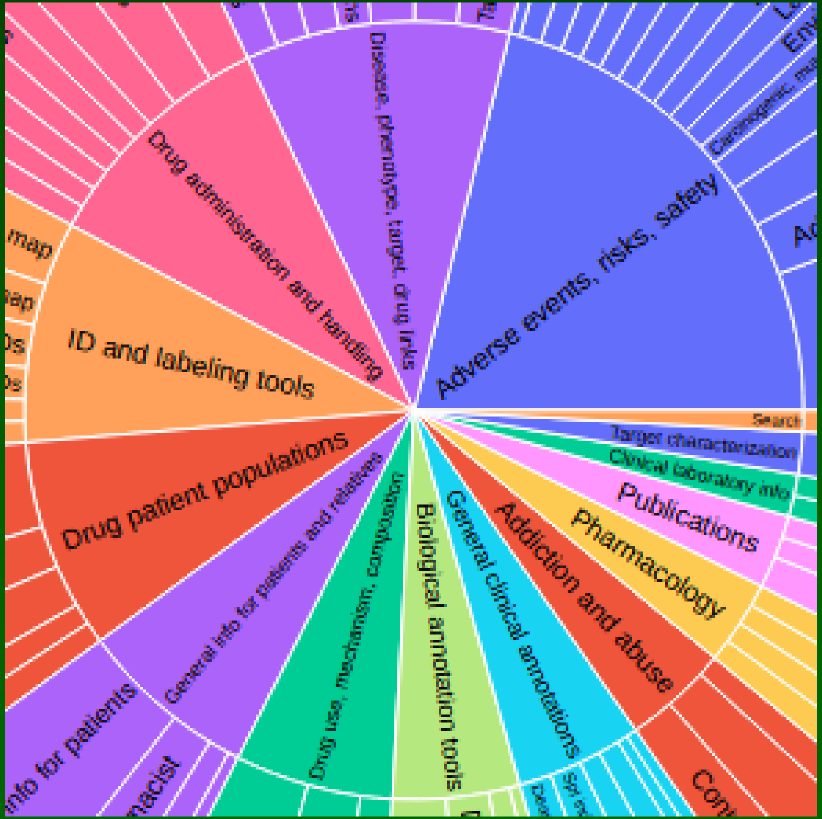
11/2025: AI/ML-Enabled FDA approved 1247 Medical Device List.

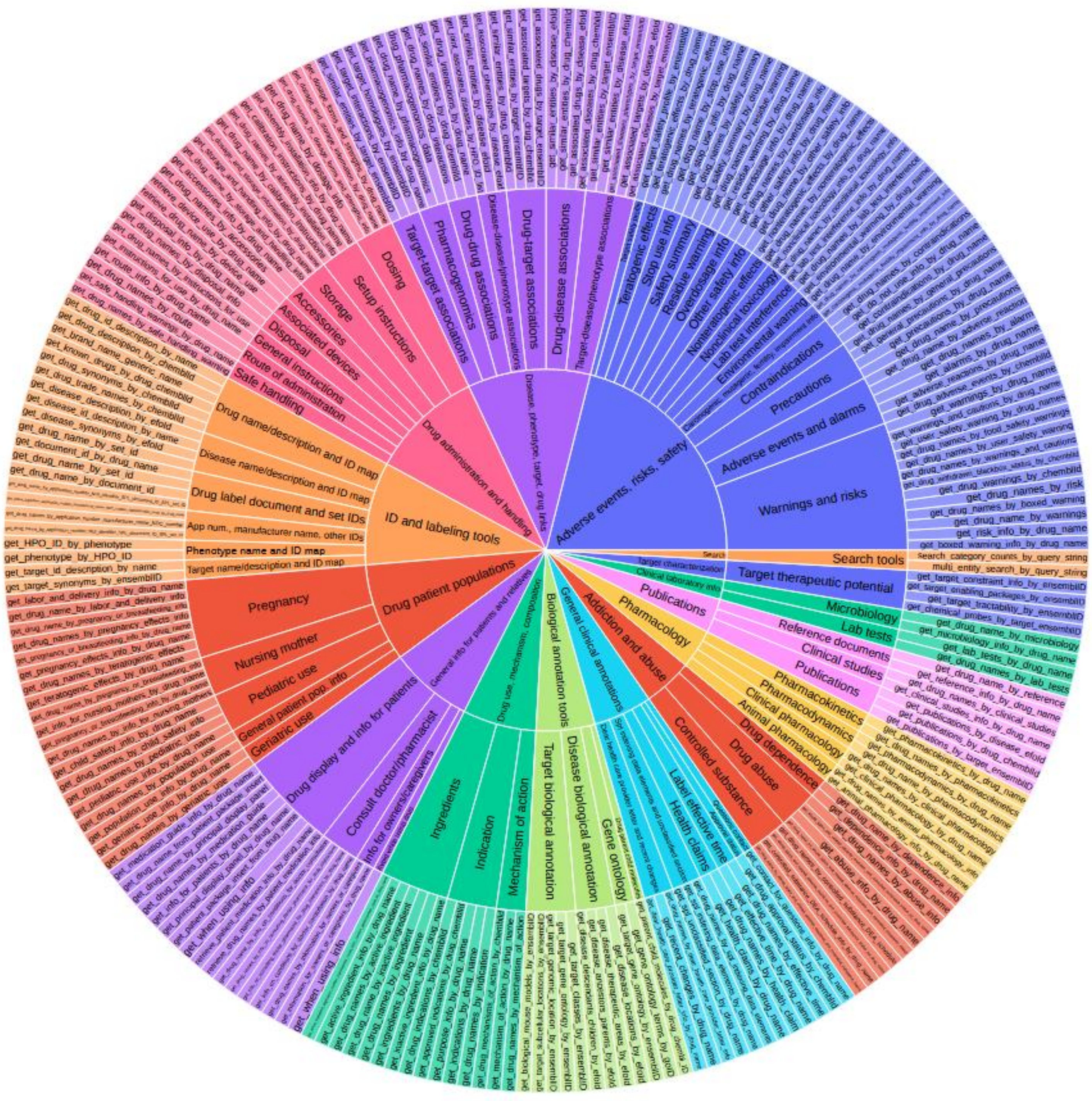
956	Radiology
125	Cardiovascular
56	Neurology
22	Anesthesiology
19	Hematology
17	Gastroenterology-Urology
10	Ophthalmic
43	Remaining



TOOLUNIVERSE:
211 biomedical tools

Gao, S. … & Zitnik, M. (3/2025). TxAgent: An AI Agent for Therapeutic Reasoning Across a Universe of Tools.

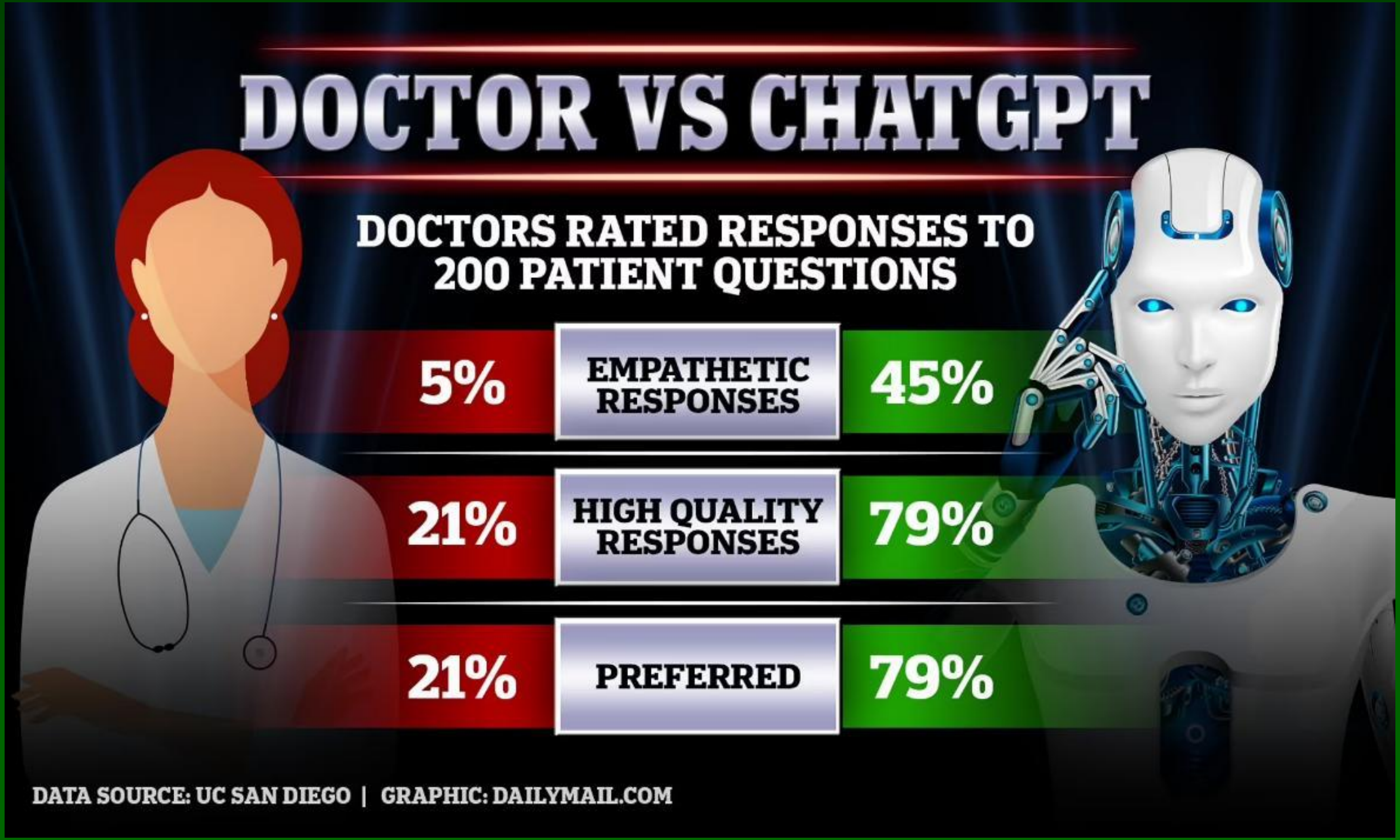




Do you really want a human doctor?

- Goh, E. +15 coauthors (10/2024). Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Network Open*, *7*(10), e2440969–e2440969.

50 physicians, median diagnostic reasoning score was 74% (IQR, 63%-84%), with the help of GPT4 76% (IQR, 66%-87%) and LLM alone 92% (IQR, 82%-97%).



DOCTOR VS CHATGPT

DOCTORS RATED RESPONSES TO 200 PATIENT QUESTIONS

5%	EMPATHETIC RESPONSES	45%
21%	HIGH QUALITY RESPONSES	79%
21%	PREFERRED	79%

DATA SOURCE: UC SAN DIEGO | GRAPHIC: DAILYMAIL.COM

China Unveils World's First **Fully AI-Run Hospital** With 14 AI Doctors

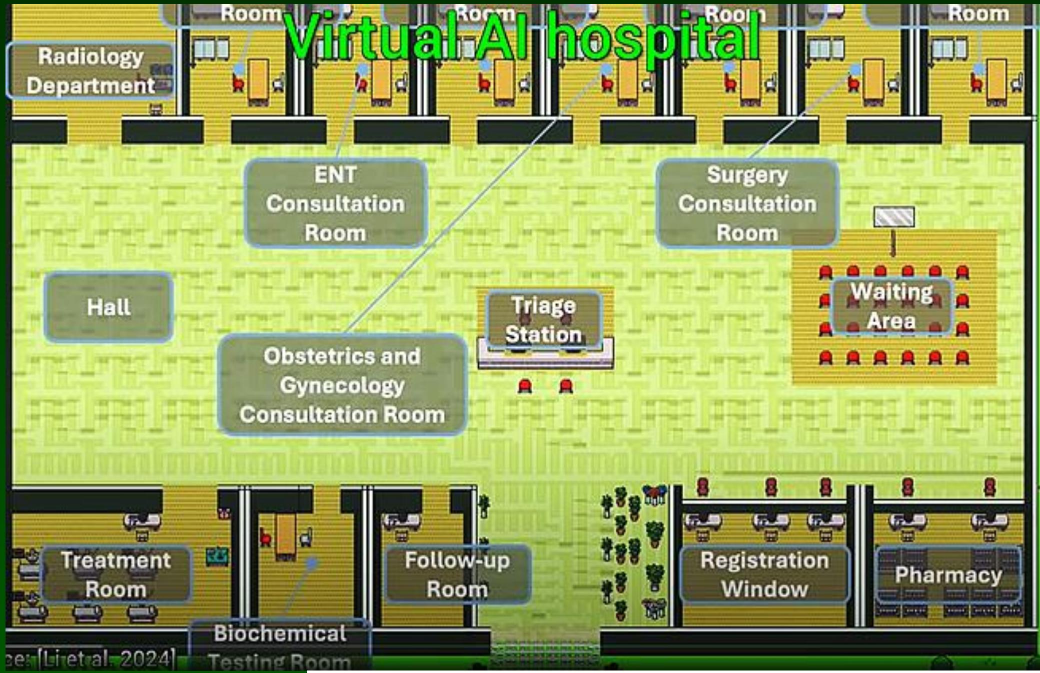
Tsinghua Hospital Simulacrum

Simulacrum-based Evolutionary Agent Learning (SEAL).

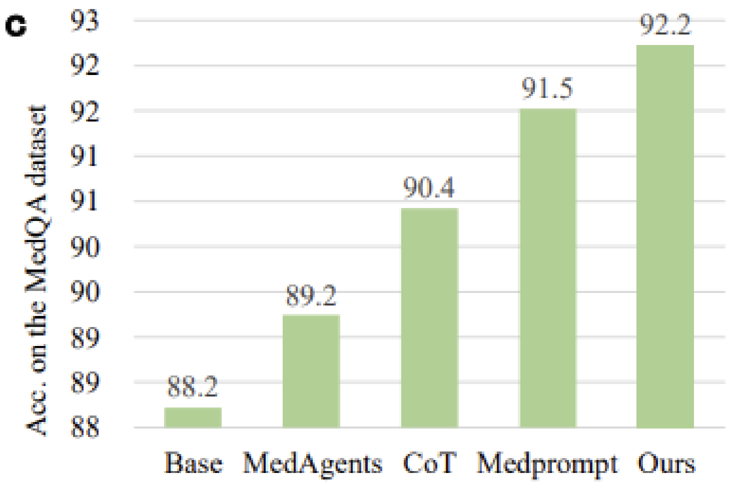
Hospital with 25 GPT-3.5 agents, 16 functional areas,
21 clinical departments, covering 339 diseases, simulates
the entire process of treating disease.
Autonomous agents as patients, nurses and doctors.
Agent doctors learn accumulating experience.

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

After thousands of patients, Agent-Doctor reaches high
competence without any manually labeled training data.



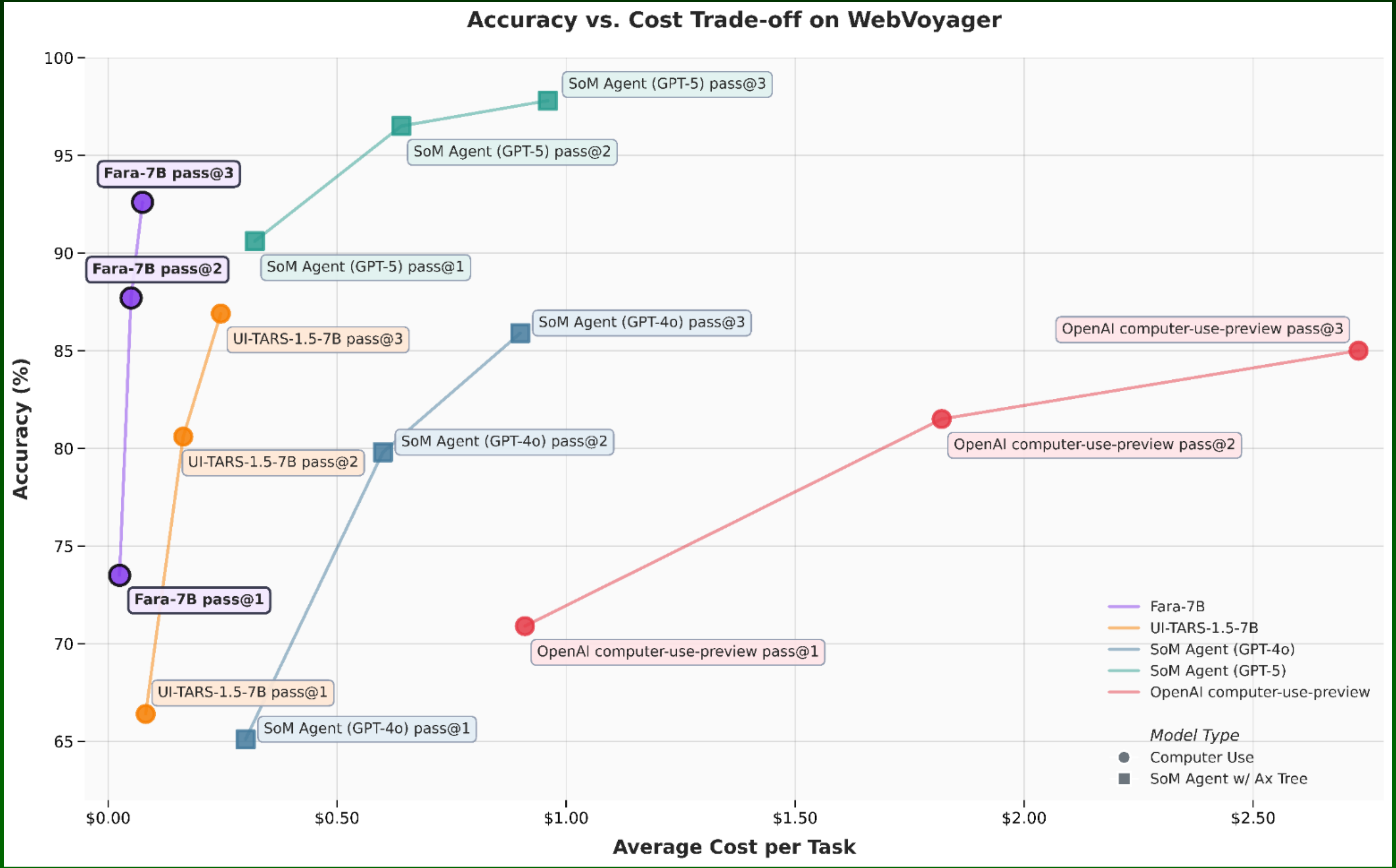
ce: [Li et al. 2024]



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

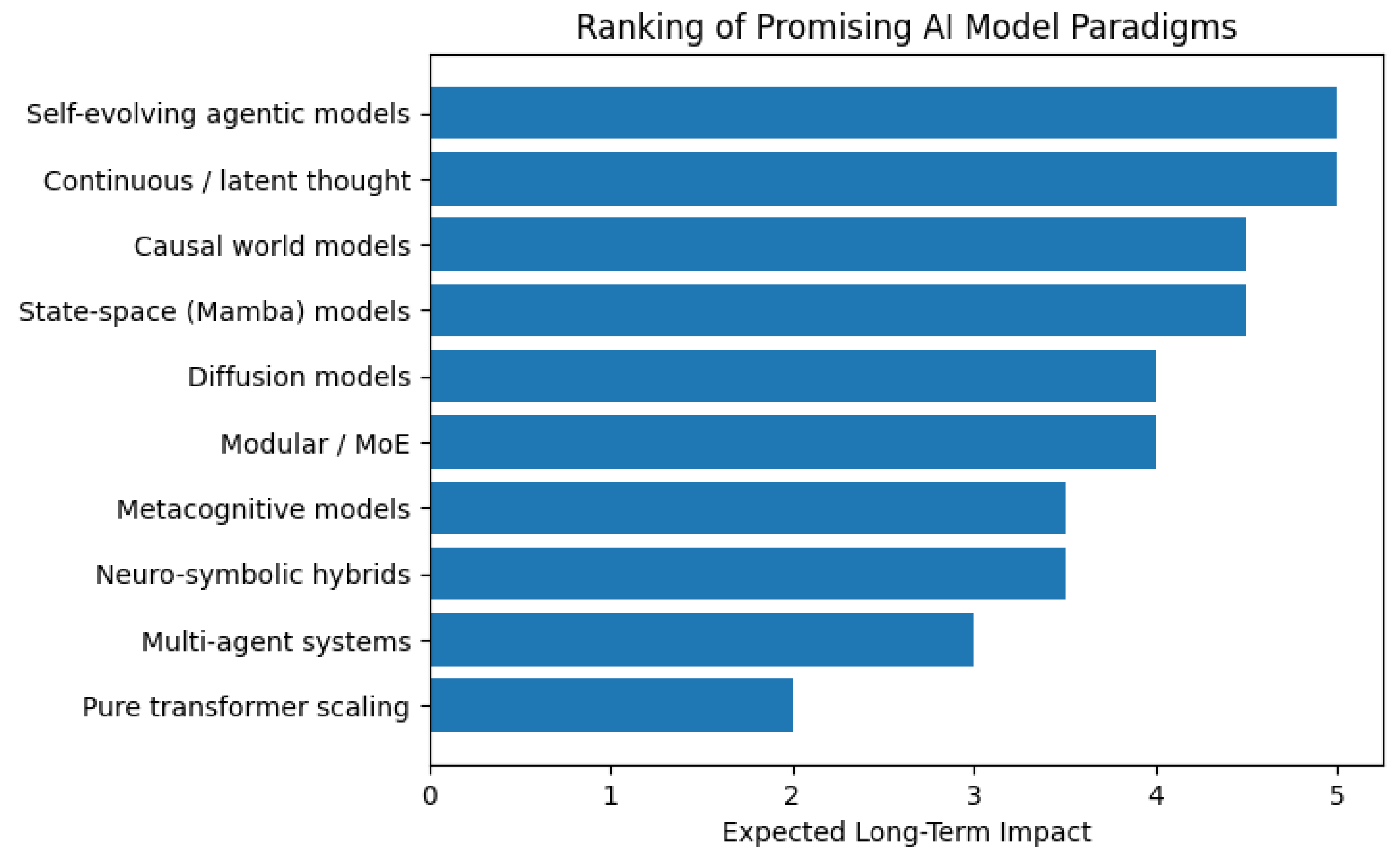
5/2025: Tsinghua AI Agent Hospital was opened, strategic initiative transforming healthcare delivery, clinical education, and medical research through deeply integrated AI systems with agents trained in >20 specialties. Goal: training a new generation of "AI-collaborative physicians".

Computer Use Agents (CUA)



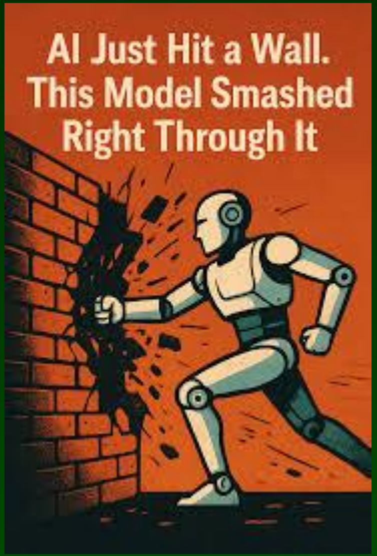
State-of-the-Art

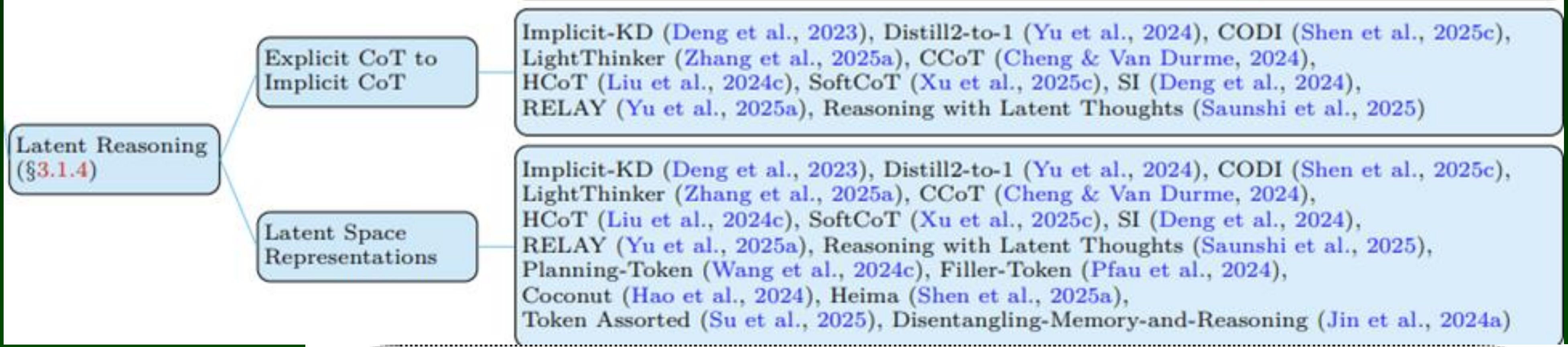
Ranking of new AI models



Combination: State-space dynamics + latent reasoning + causal models + self-evolving agents.

Efficient/latent reasoning





Latent Reasoning (§3.1.4)

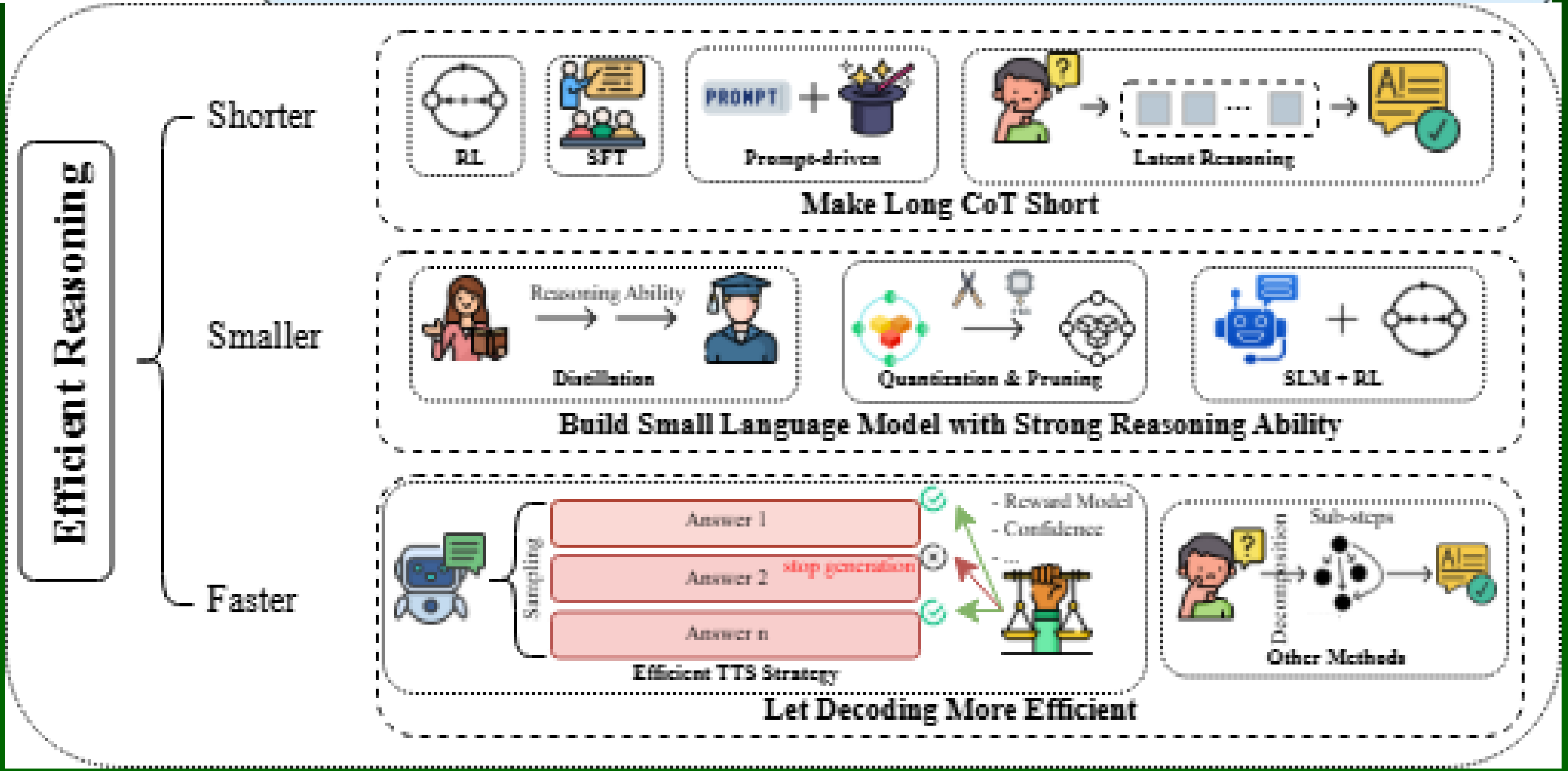
- Explicit CoT to Implicit CoT
  - Implicit-KD (Deng et al., 2023), Distill2-to-1 (Yu et al., 2024), CODI (Shen et al., 2025c), LightThinker (Zhang et al., 2025a), CCoT (Cheng & Van Durme, 2024), HCoT (Liu et al., 2024c), SoftCoT (Xu et al., 2025c), SI (Deng et al., 2024), RELAY (Yu et al., 2025a), Reasoning with Latent Thoughts (Saunshi et al., 2025)
- Latent Space Representations
  - Implicit-KD (Deng et al., 2023), Distill2-to-1 (Yu et al., 2024), CODI (Shen et al., 2025c), LightThinker (Zhang et al., 2025a), CCoT (Cheng & Van Durme, 2024), HCoT (Liu et al., 2024c), SoftCoT (Xu et al., 2025c), SI (Deng et al., 2024), RELAY (Yu et al., 2025a), Reasoning with Latent Thoughts (Saunshi et al., 2025), Planning-Token (Wang et al., 2024c), Filler-Token (Pfau et al., 2024), Coconut (Hao et al., 2024), Heima (Shen et al., 2025a), Token Assorted (Su et al., 2025), Disentangling-Memory-and-Reasoning (Jin et al., 2024a)

Hao S et al. Training Large Language Models to Reason in a Continuous Latent Space. Chain of Continuous Thought model (Coconut),emergence of advanced reasoning patterns.

Feng S, Fang G, Ma X, Wang X. 2025 Efficient Reasoning Models: A Survey.

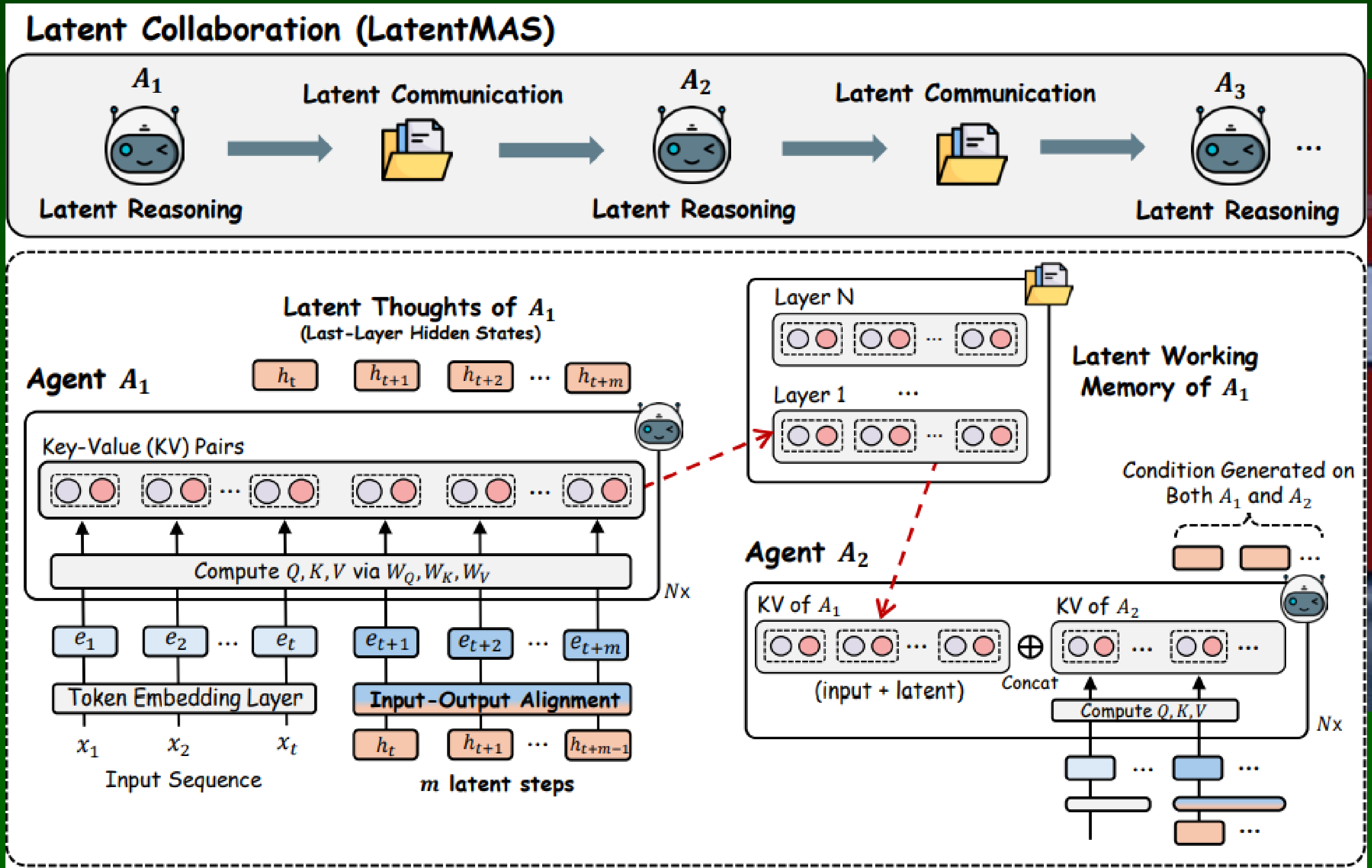
Github papers on efficient reasoning.
Github LatentCoT-Horizon
Github Awesam Latent CoT

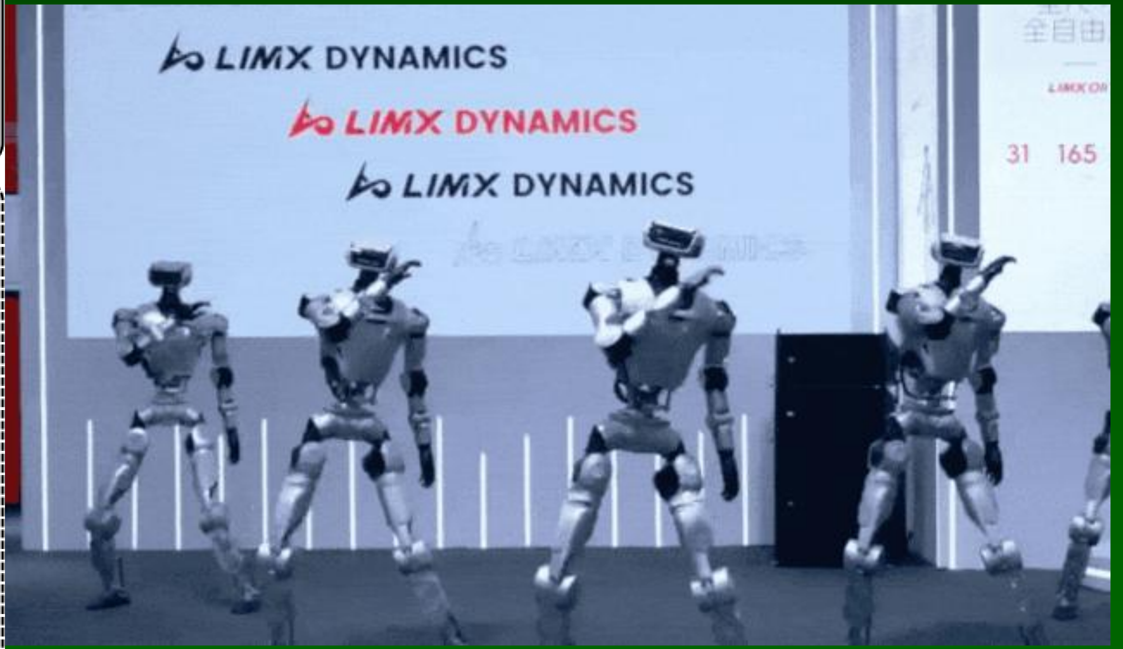
Small models, fast decoding, short chains of thoughts, chunking.



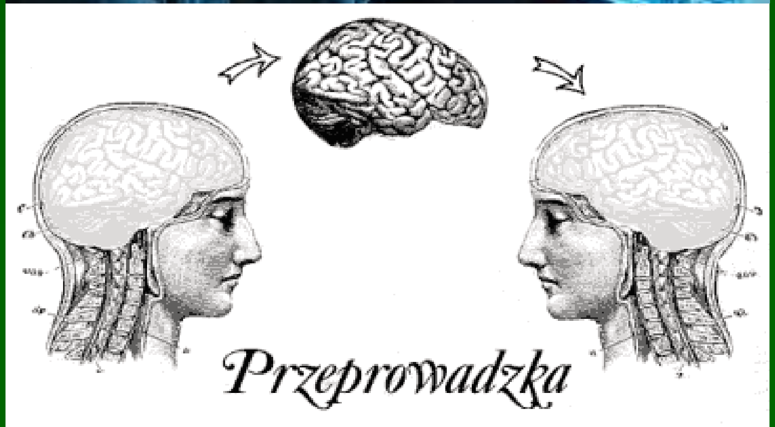
Continuous Thought Machines

- Darlow, L., Regan, C., Risi, S., Seely, J., & Jones, L. (5/2025). Continuous Thought Machines.
- He, J., Bai, R. H., Williamson, S., Pan, J. Z., Jaitly, N., & Zhang, Y. (11/2025).
  CLaRa: Bridging Retrieval and Generation with Continuous Latent Reasoning.
- Zou, J. … Yang, L. (11/2025). Latent Collaboration in Multi-Agent Systems.
  LatentMAS transforms tensor of keys/values to other agents, enabling collaboration within the latent space.





Mind uploading?

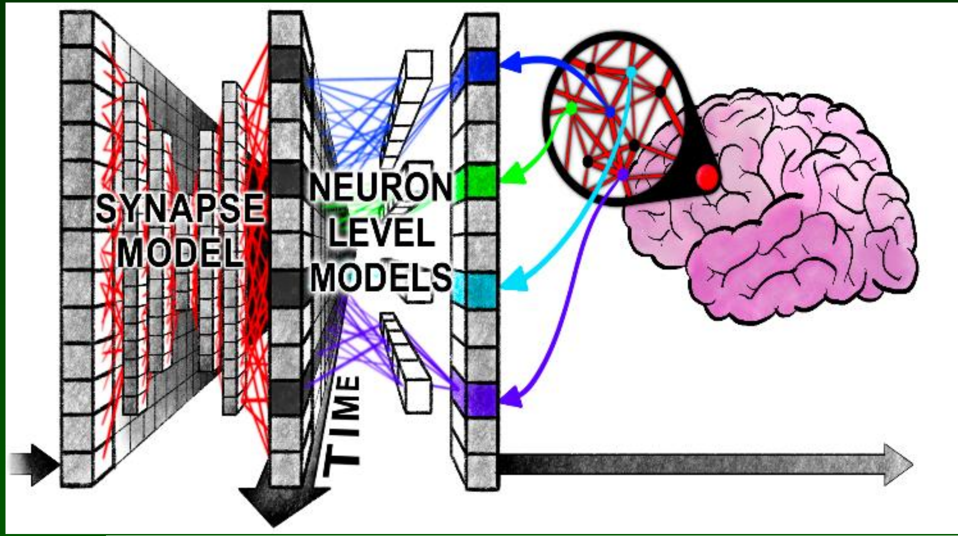


Beyond transformers – brain inspirations

SakanaAI/continuous-thought-machines (2025). Github SakanaAI

Continuous Thought Machine (CTM): synchronization of neural activity over time as a latent representation for taking actions in a world.

Performs tasks that require complex sequential reasoning at human level in biologically plausible way.



Kosowski, A. et al. (9/2025). The Dragon Hatchling: The Missing Link between the Transformer and Models of the Brain. Pathway.com

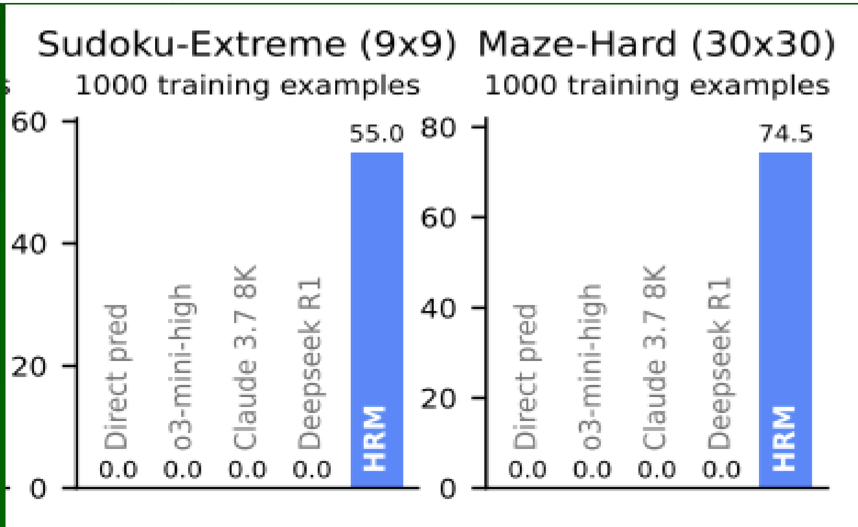
Interesting neural architecture, learns during reasoning process.

Shojaee et al. The Illusion of Thinking. 6/2025. Really?
Large Reasoning Models (LRMs) fail to solve complex problems.

Wang, G. et al. Hierarchical Reasoning Model. 6/2025.

Hierarchical Reasoning Model (HRM) recurrent architecture.
A high-level module for abstract planning, low-level for details.

27M model without pre-training, using only 1000 training samples, solves complex Sudoku puzzles, finds optimal path in large mazes.



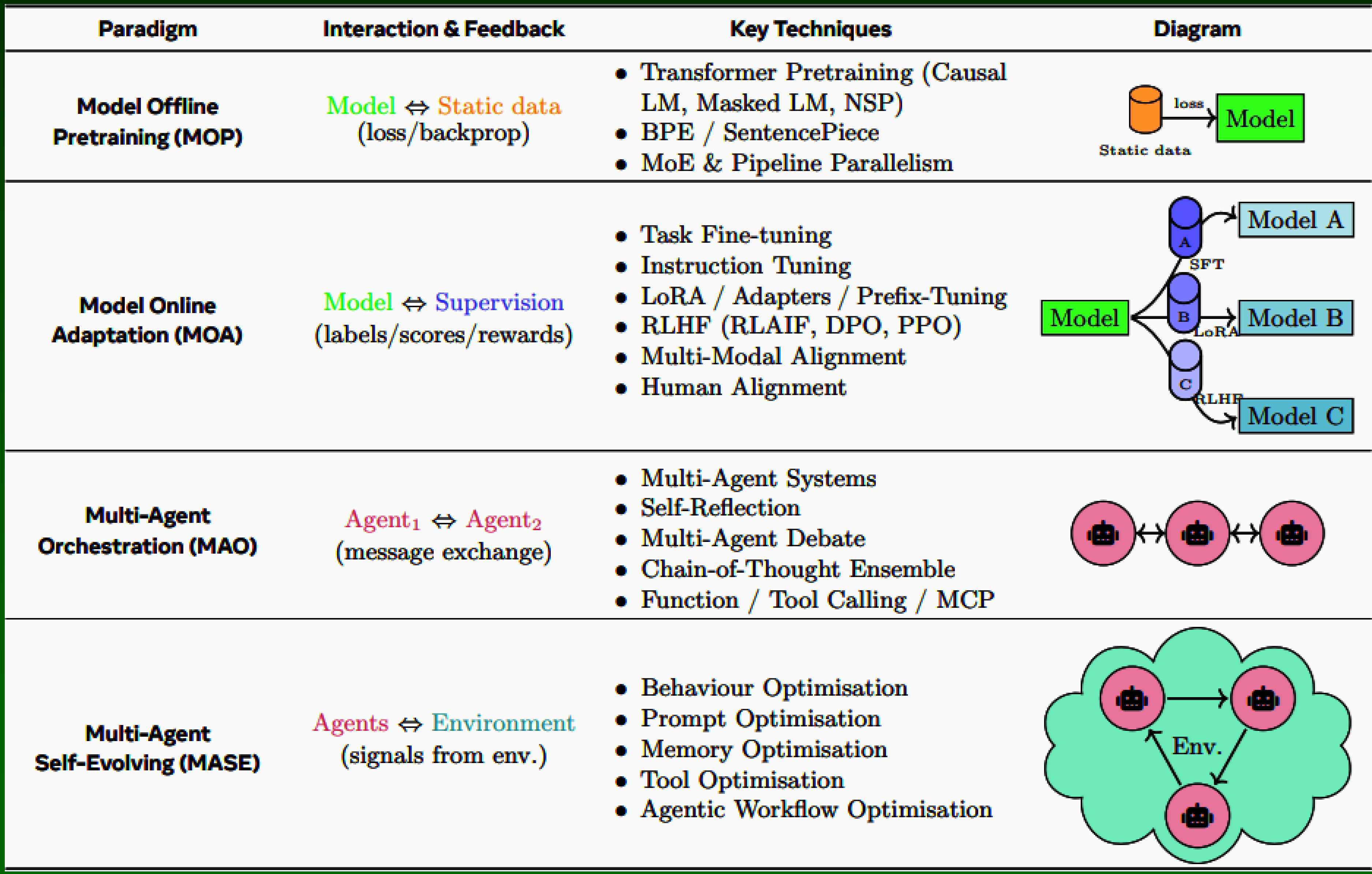
Direct prediction, small-sample learning

Self-evolving AI agents

Fang, J. … Meng, Z. (2025). A Comprehensive Survey of Self-Evolving AI Agents: A New Paradigm Bridging Foundation Models and Lifelong Agentic Systems. arXiv:2508.07407.

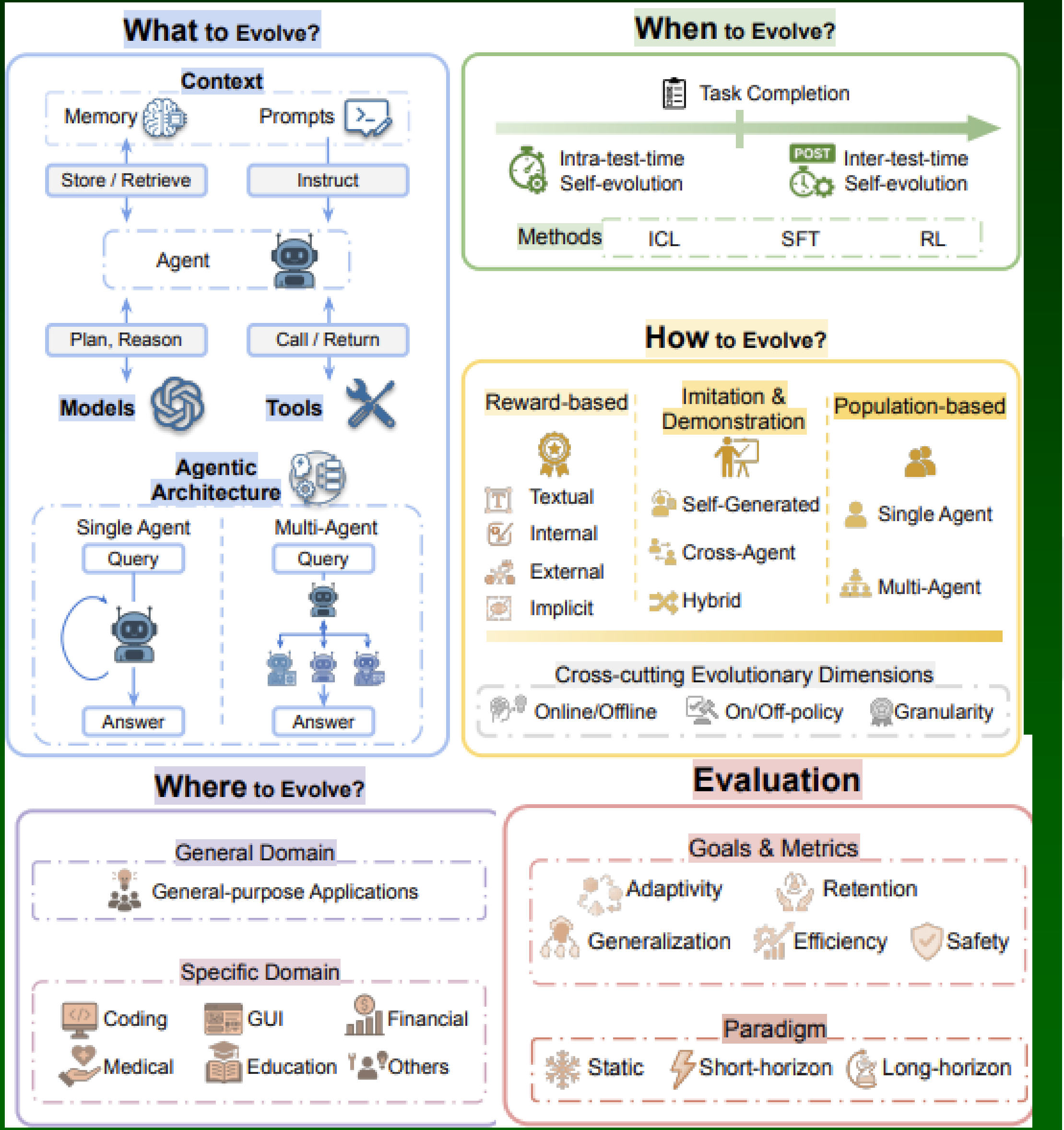
Gao, H et al. (7/2025). A Survey of Self-Evolving Agents: On Path to Artificial Super Intelligence.

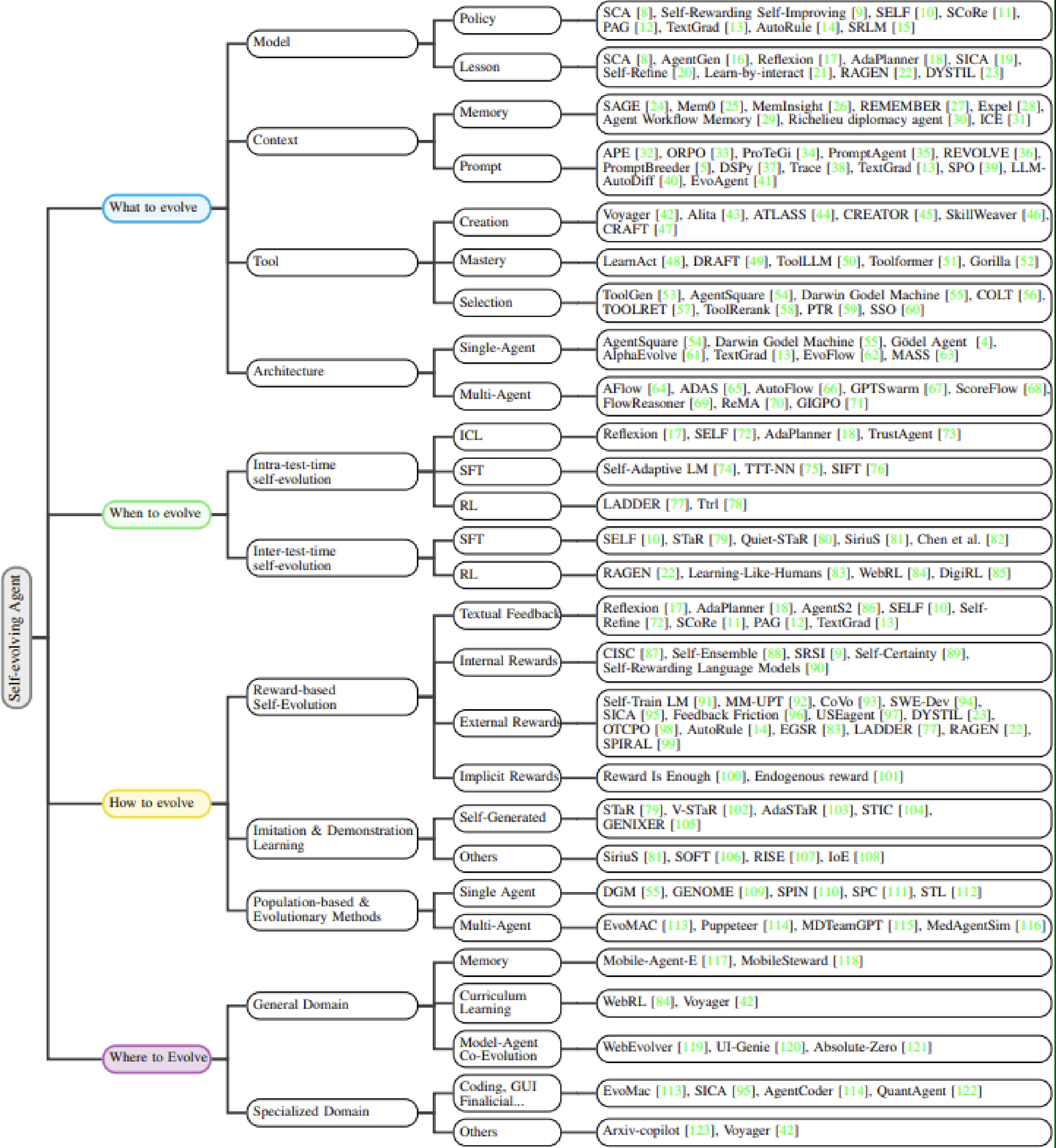
4 LLM-centric learning paradigms, from static offline model training to dynamic online, multi-agent orchestration, and self-evolving agent evolution.



Paradigm	Interaction & Feedback	Key Techniques	Diagram
Model Offline Pretraining (MOP)	Model ⇔ Static data (loss/backprop)	• Transformer Pretraining (Causal LM, Masked LM, NSP) • BPE / SentencePiece • MoE & Pipeline Parallelism	
Model Online Adaptation (MOA)	Model ⇔ Supervision (labels/scores/rewards)	• Task Fine-tuning • Instruction Tuning • LoRA / Adapters / Prefix-Tuning • RLHF (RLAIF, DPO, PPO) • Multi-Modal Alignment • Human Alignment	
Multi-Agent Orchestration (MAO)	Agent_1 ⇔ Agent_2 (message exchange)	• Multi-Agent Systems • Self-Reflection • Multi-Agent Debate • Chain-of-Thought Ensemble • Function / Tool Calling / MCP	
Multi-Agent Self-Evolving (MASE)	Agents ⇔ Environment (signals from env.)	• Behaviour Optimisation • Prompt Optimisation • Memory Optimisation • Tool Optimisation • Agentic Workflow Optimisation	

Self-evolving AI agents





- Self-evolving Agent
  - What to evolve
    - Model
      - Policy: SCA [8], Self-Rewarding Self-Improving [9], SELF [10], SCoRe [11], PAG [12], TextGrad [13], AutoRule [14], SRLM [15]
      - Lesson: SCA [8], AgentGen [16], Reflexion [17], AdaPlanner [18], SICA [19], Self-Refine [20], Learn-by-interact [21], RAGEN [22], DYSTIL [23]
    - Context
      - Memory: SAGE [24], Mem0 [25], MemInsight [26], REMEMBER [27], Expel [28], Agent Workflow Memory [29], Richelieu diplomacy agent [30], ICE [31]
      - Prompt: APE [32], ORPO [33], ProTeGi [34], PromptAgent [35], REVOLVE [36], PromptBreeder [5], DSPy [37], Trace [38], TextGrad [13], SPO [39], LLM-AutoDiff [40], EvoAgent [41]
    - Tool
      - Creation: Voyager [42], Alita [43], ATLASS [44], CREATOR [45], SkillWeaver [46], CRAFT [47]
      - Mastery: LearnAct [48], DRAFT [49], ToolLLM [50], Toolformer [51], Gorilla [52]
      - Selection: ToolGen [53], AgentSquare [54], Darwin Godel Machine [55], COLT [56], TOOLRET [57], ToolRerank [58], PTR [59], SSO [60]
    - Architecture
      - Single-Agent: AgentSquare [54], Darwin Godel Machine [55], Gödel Agent [4], AlphaEvolve [61], TextGrad [13], EvoFlow [62], MASS [63]
      - Multi-Agent: AFlow [64], ADAS [65], AutoFlow [66], GPTSwarm [67], ScoreFlow [68], FlowReasoner [69], ReMA [70], GIGPO [71]
  - When to evolve
    - Intra-test-time self-evolution
      - ICL: Reflexion [17], SELF [72], AdaPlanner [18], TrustAgent [73]
      - SFT: Self-Adaptive LM [74], TTT-NN [75], SIFT [76]
      - RL: LADDER [77], Ttrl [78]
    - Inter-test-time self-evolution
      - SFT: SELF [10], STaR [79], Quiet-STaR [80], SiriuS [81], Chen et al. [82]
      - RL: RAGEN [22], Learning-Like-Humans [83], WebRL [84], DigiRL [85]
  - How to evolve
    - Reward-based Self-Evolution
      - Textual Feedback: Reflexion [17], AdaPlanner [18], AgentS2 [86], SELF [10], Self-Refine [72], SCoRe [11], PAG [12], TextGrad [13]
      - Internal Rewards: CISC [87], Self-Ensemble [88], SRSI [9], Self-Certainty [89], Self-Rewarding Language Models [90]
      - External Rewards: Self-Train LM [91], MM-UPT [92], CoVo [93], SWE-Dev [94], SICA [95], Feedback Friction [96], USEagent [97], DYSTIL [23], OTCPO [98], AutoRule [14], EGSR [83], LADDER [77], RAGEN [22], SPIRAL [99]
      - Implicit Rewards: Reward Is Enough [100], Endogenous reward [101]
    - Imitation & Demonstration Learning
      - Self-Generated: STaR [79], V-STaR [102], AdaSTaR [103], STIC [104], GENIXER [105]
      - Others: SiriuS [81], SOFT [106], RISE [107], IoE [108]
    - Population-based & Evolutionary Methods
      - Single Agent: DGM [55], GENOME [109], SPIN [110], SPC [111], STL [112]
      - Multi-Agent: EvoMAC [113], Puppeteer [114], MDTeamGPT [115], MedAgentSim [116]
  - Where to Evolve
    - General Domain
      - Memory: Mobile-Agent-E [117], MobileSteward [118]
      - Curriculum Learning: WebRL [84], Voyager [42]
      - Model-Agent Co-Evolution: WebEvolver [119], UI-Genie [120], Absolute-Zero [121]
    - Specialized Domain
      - Coding, GUI Finalicial...: EvoMac [113], SICA [95], AgentCoder [114], QuantAgent [122]
      - Others: Arxiv-copilot [123], Voyager [42]

Self-improving neural architectures

Wang, W. … & Schmidhuber, J. *Huxley-Gödel Machine: Human-Level Coding Agent Development by an Approximation of the Optimal Self-Improving Machine*. (10/2025). github.com/metauto-ai/HGM.

Self-improvement of coding agents that edit their own codebases based on the approximation to the Gödel machine which accepts only modifications that provably increase the expected long-term utility.

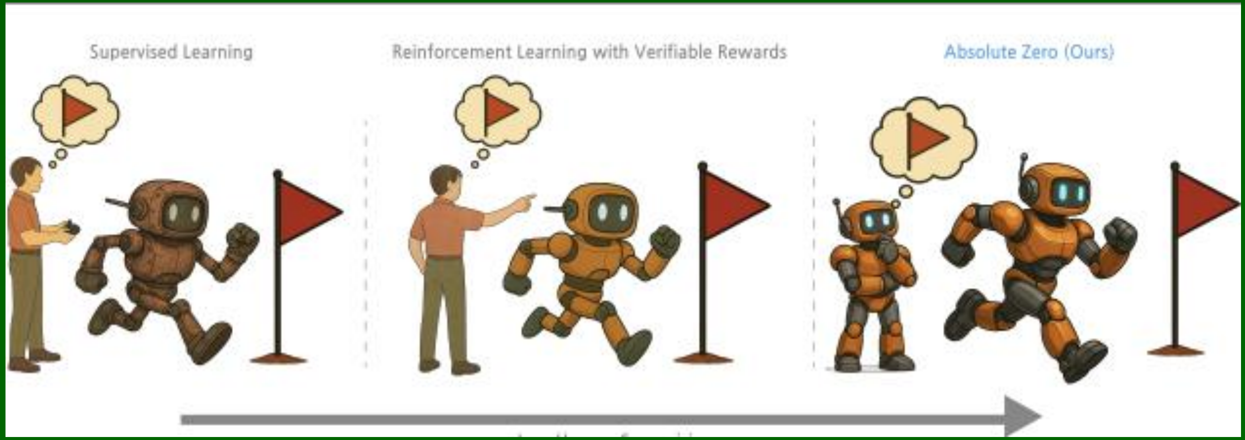
Zhao, A. et al. *Absolute Zero: Reinforced Self-play Reasoning with Zero Data.* 5/2025.

Verifiable rewards (RLVR) allow self-play to reach superhuman levels, but needs curated Q/A pairs.

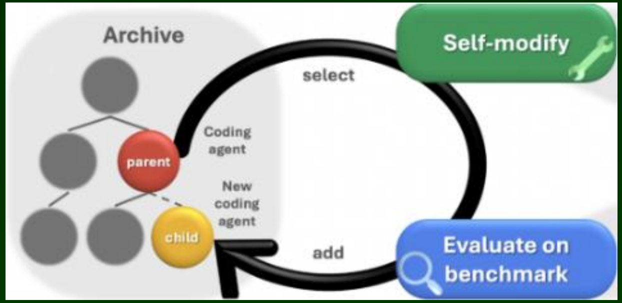
Absolute Zero Reasoner (AZR) self-evolves its training and reasoning ability using a code executor to validate code reasoning tasks and verify answers, serving as source of RLVR to guide open-ended yet grounded learning. Trained entirely without external data AZR achieves SOTA performance on coding and math reasoning tasks, outperforming models that rely on tens of thousands of in-domain human-curated examples.

Liu J. … Yao, H. (2025). Agent0-VL: Exploring Self-Evolving Agent for Tool-Integrated Vision-Language Reasoning.

Agent that achieves continual improvement with tool-integrated reasoning, self-evaluation, enabling the model to introspect, verify, and refine its reasoning through evidence-grounded analysis.



Self-improving LLMs



Sakana.ai (5/2025)

The Darwin Gödel Machine: AI that improves itself by rewriting its own code

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

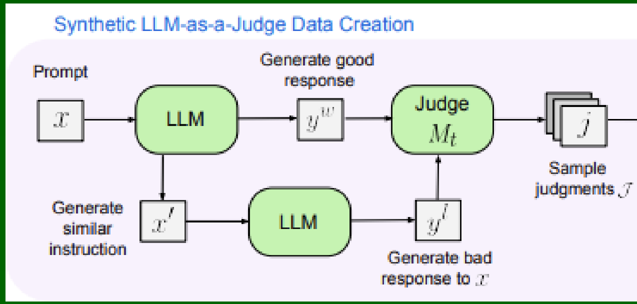
Self-Adapting LLMs (SEAL), a framework that enables LLMs to self-adapt by generating their own finetuning data and update directives, may restructure the information, optimize its hyperparameters, or invoke tools for data augmentation and gradient-based updates.

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.

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.


Synthetic LLM-as-a-Judge Data Creation
Prompt
x
LLM
Generate good response
y^w
Judge
M_t
j
Sample judgments $\mathcal{J}$
Generate similar instruction
x'
LLM
y^l
Generate bad response to x

AlphaGo Moment for Model Architecture Discovery

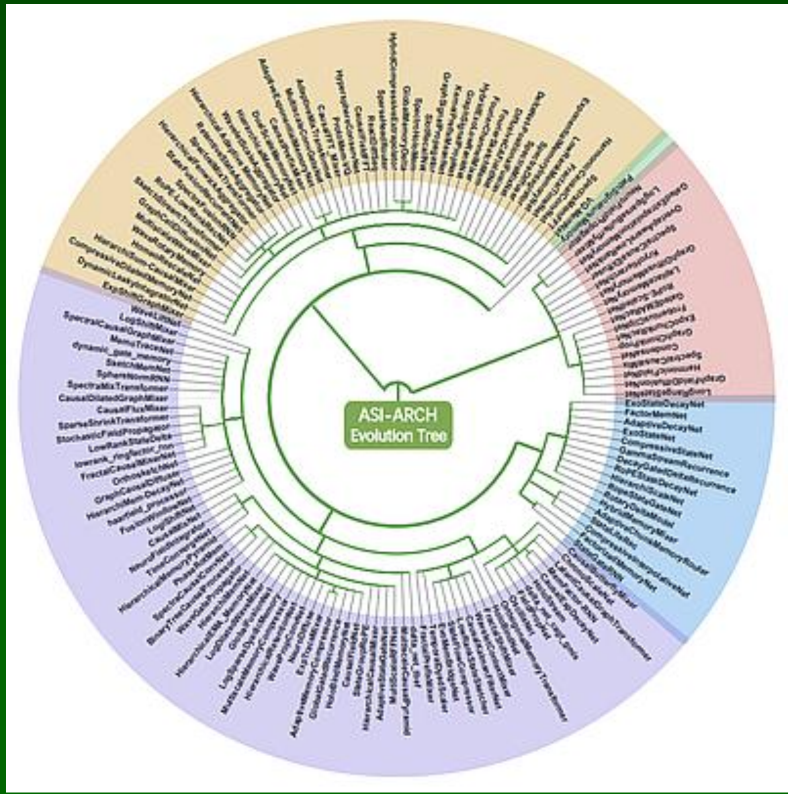
Liu, Y. … & Liu, P. AlphaGo Moment for Model Architecture Discovery. arXiv:2507.18074.

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.

1 773 autonomous experiments, used over 20,000 GPU hours, discovered 106 state-of-the-art (SOTA) linear attention architectures, creating designs that systematically surpass human-designed baselines and illuminate new pathways for architectural innovation.

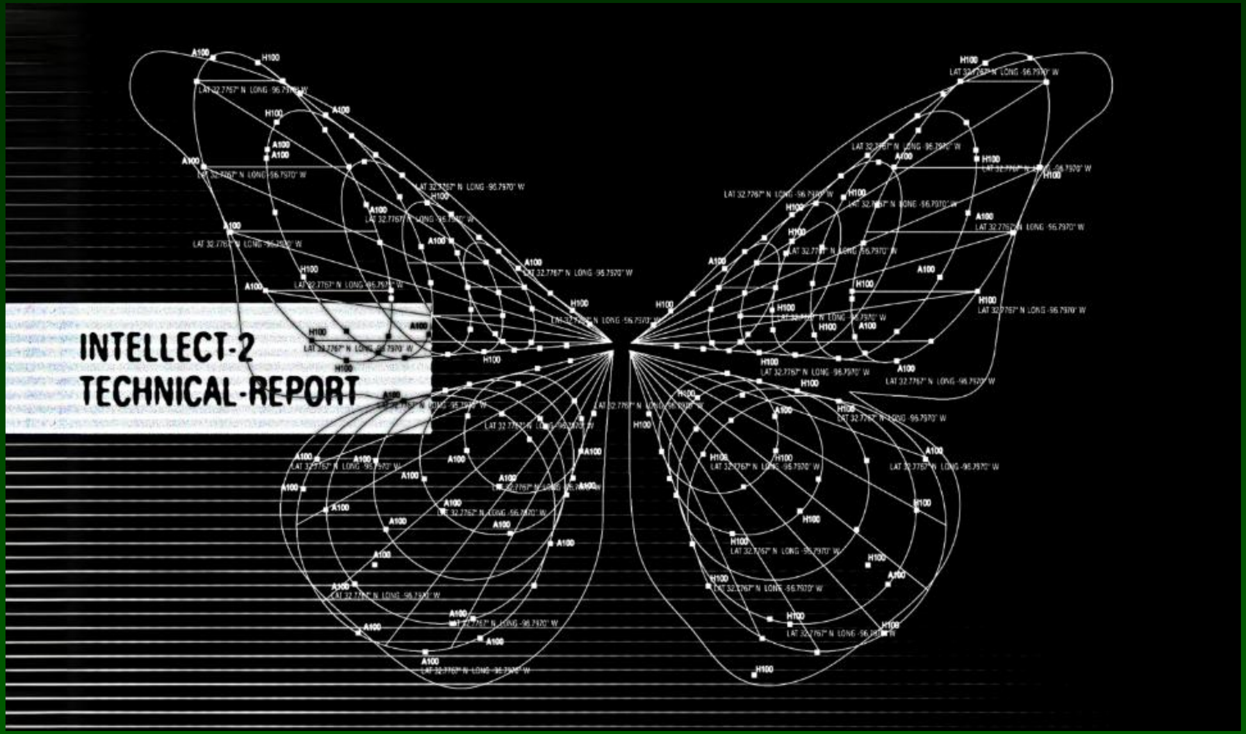


Jankowski N, Duch W, Grąbczewski K, Meta-learning in Computational Intelligence. Studies in Computational Intelligence, Vol. 358, Springer 2011.

Prime Intellect: Open Superintelligence

INTELLECT-3 Release:
A 106B parameter MoE model
Trained Through Globally Distributed
Reinforcement Learning.

INTELLECT-3 trains a reasoning language model using fully asynchronous RL across a dynamic, heterogeneous swarm of compute nodes.
Prime-RL training, Verifiers and the Environments Hub, Sandboxes and Compute Orchestration:
managed 512 NVIDIA H200 GPUs across 64 interconnected nodes.



Planetary-Scale Inference: Previewing our Peer-To-Peer Decentralized Inference Stack
https://www.primeintellect.ai

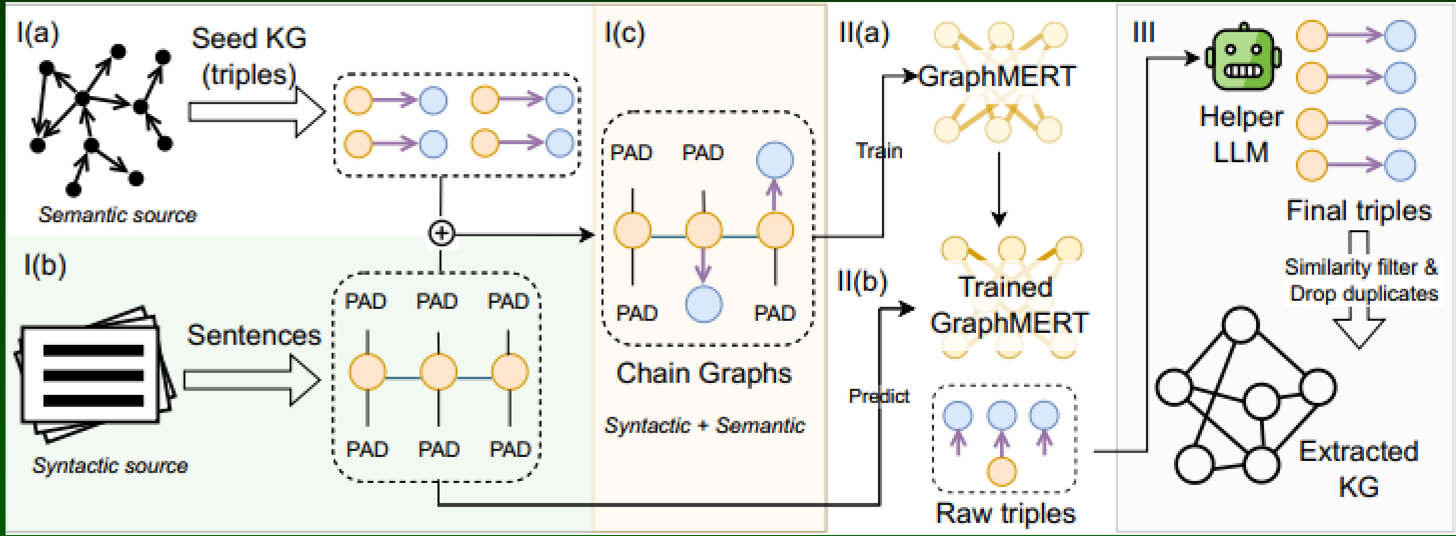
Kim, Y et al. (12/2025). Towards a Science of Scaling Agent Systems. arXiv:2512.08296.
MoE fails on sequential reasoning tasks. Framework to predict optimal MoE architectures.

Neurosymbolic GraphMERT

GraphMERT 80M graphical encoder-only neurosymbolic model. Distills high-quality KGs from unstructured text corpora and its own internal representations: learns abstractions, symbolic KGs for verifiable reasoning Achieves SOTA accuracy providing superior symbolic knowledge representation.

Belova, M. et al. (10/2025). *GraphMERT: Efficient and Scalable Distillation of Reliable Knowledge Graphs from Unstructured Data.*

KG integration is a key step toward domain-specific superintelligence.
Recovers medical UMLS relations more accurately than LLMs, re-creating pathways of spreading neural activations.



Concepts of specific semantic type are identified, related to related concepts, added to the text, providing expanded associative representations.

PAD (Personalized Alignment at Decoding-time) aligns LLM outputs with diverse, personalized preferences.

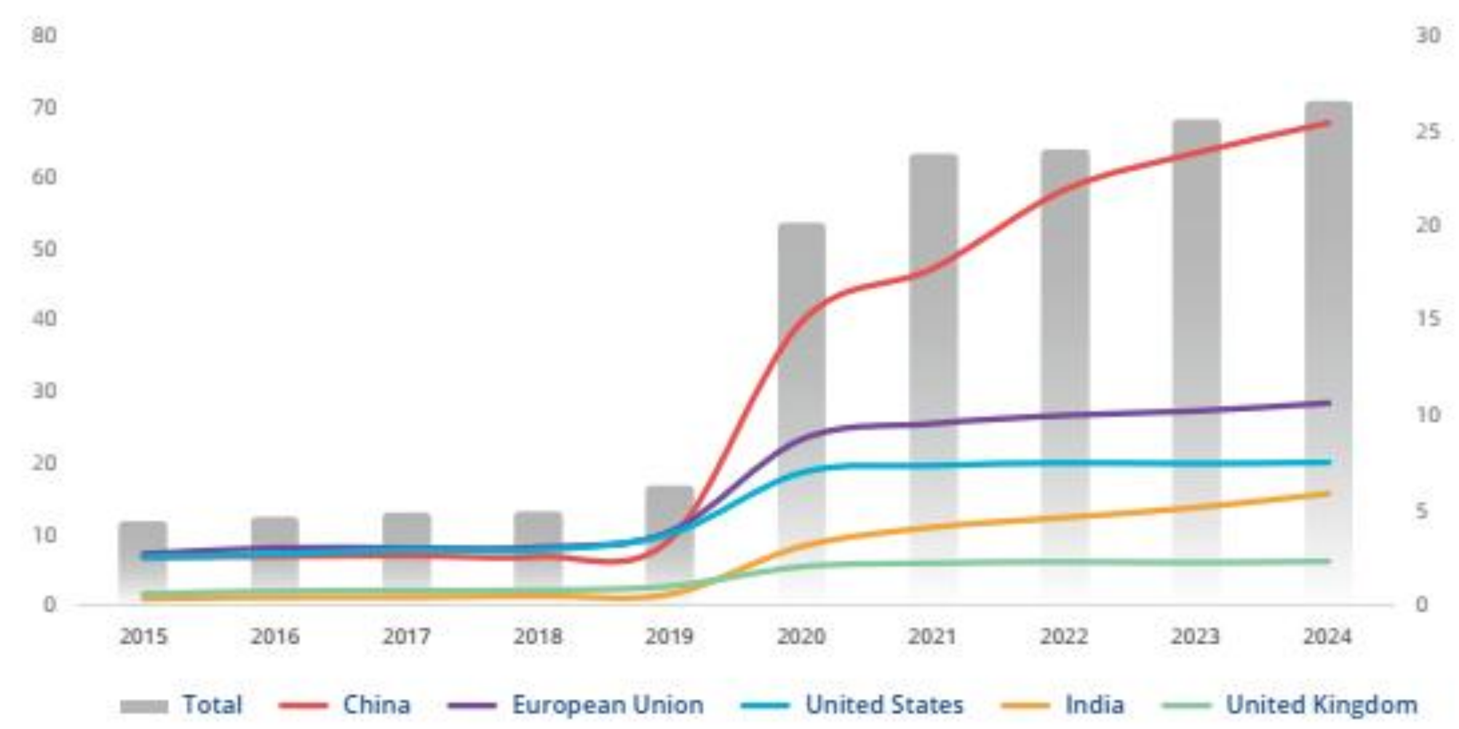
Duch W, Matykiewicz P, Pestian J. Neurolinguistic Approach to Natural Language Processing with Applications to Medical Text Analysis. Neural Networks (2008). Idea: each token generates synthetic context data, best are selected!
US Patent 8,930,178 B2. Processing clinical text with domain-specific spreading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

Towards AI4science factories

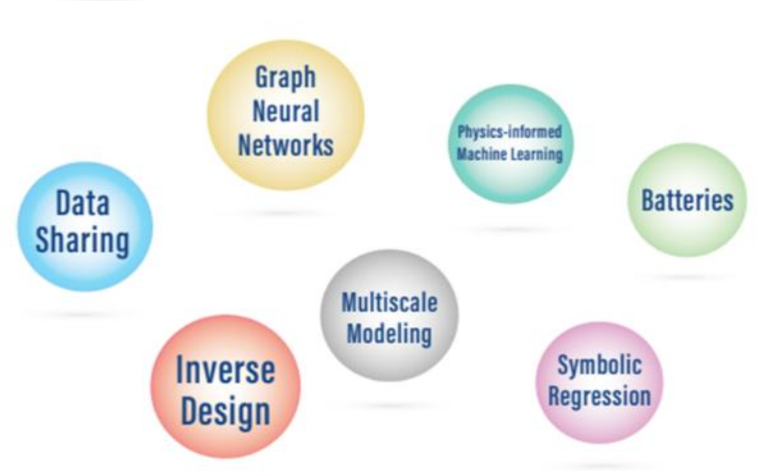
AI in science report

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

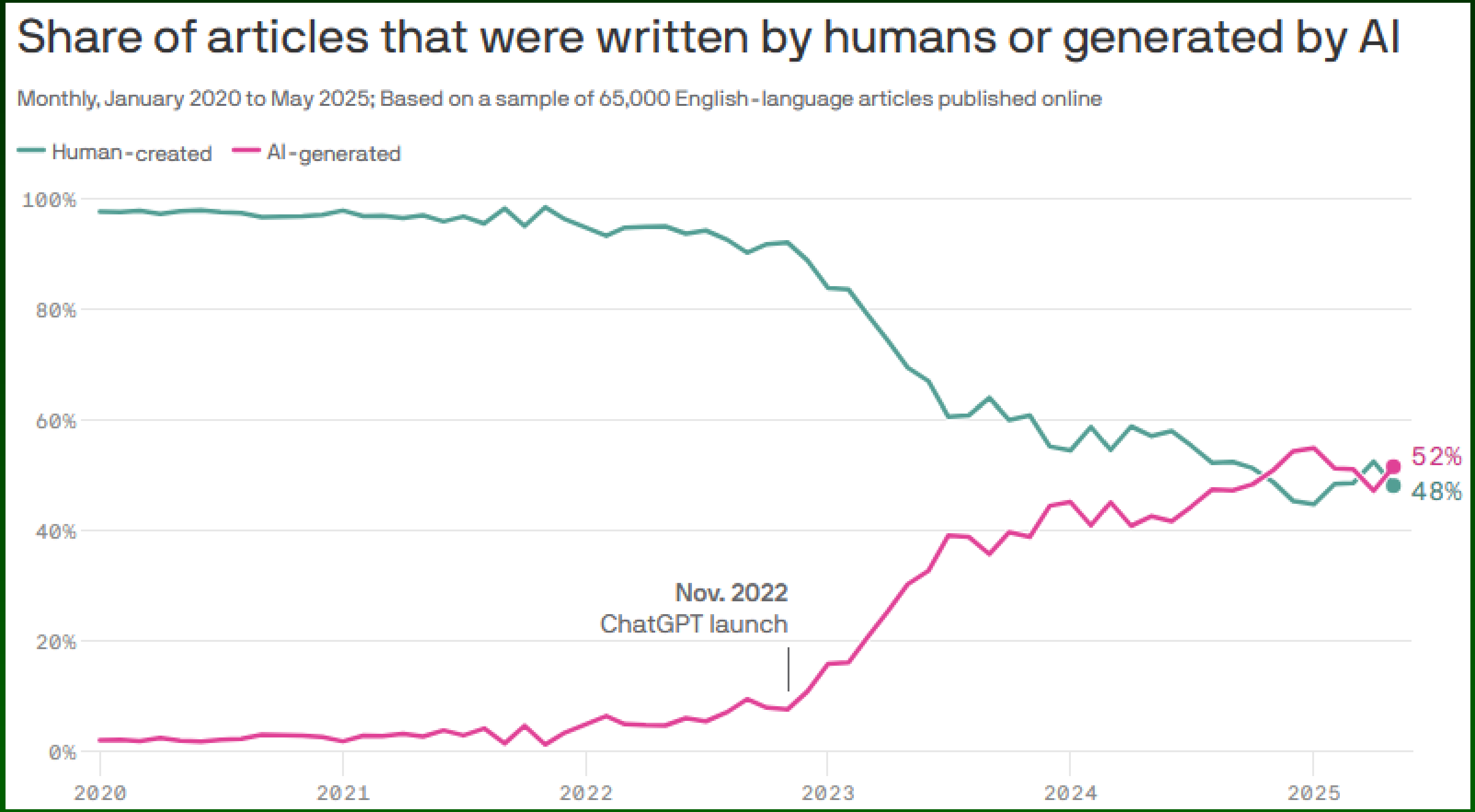
Number of articles on AI in computer science, physics, mathematics, engineering, geosciences, environment, social sciences:
China >> EU > US, India > UK;   Life sciences EU< US, China > India > UK.
In 2024 about 955 000 publications on AI, in physical sciences 70,700.
Nature Index 01.10.2025, Research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



Agent
RAG
Extreme AI Risks
Interpretability
Scaling Law
AI Security
Endogenous Security
Game Theory
AI Alignment
Deep Reasoning
AI Safety
Frontier Models
Dynamic Heterogeneous Redundancy
Safety Evaluation
Precision Medicine
Genome
Biological System
Computational Models
Physiology
Genetic Algorithms
Brain Imaging
Public Health
Brain
Healthcare
Gene
Molecular Mechanism
Neuroscience
Connectivity
Disease Prediction
Computer Vision
Proteomics
Disease Diagnosis
Neuron
Large Language Model



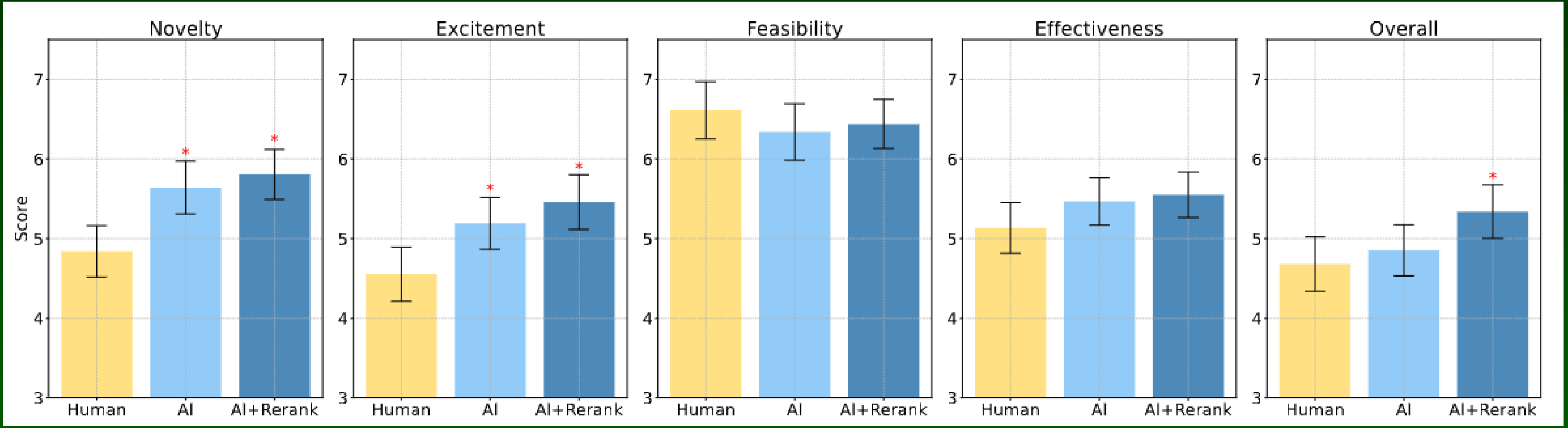
Papers: AI vs humans



From graphite.io. The quality of AI content is rapidly improving. In many cases, AI-generated content is as good or better than content written by humans (MIT Study). It is often hard for people to distinguish whether content is created by AI (Originality AI Study).

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).



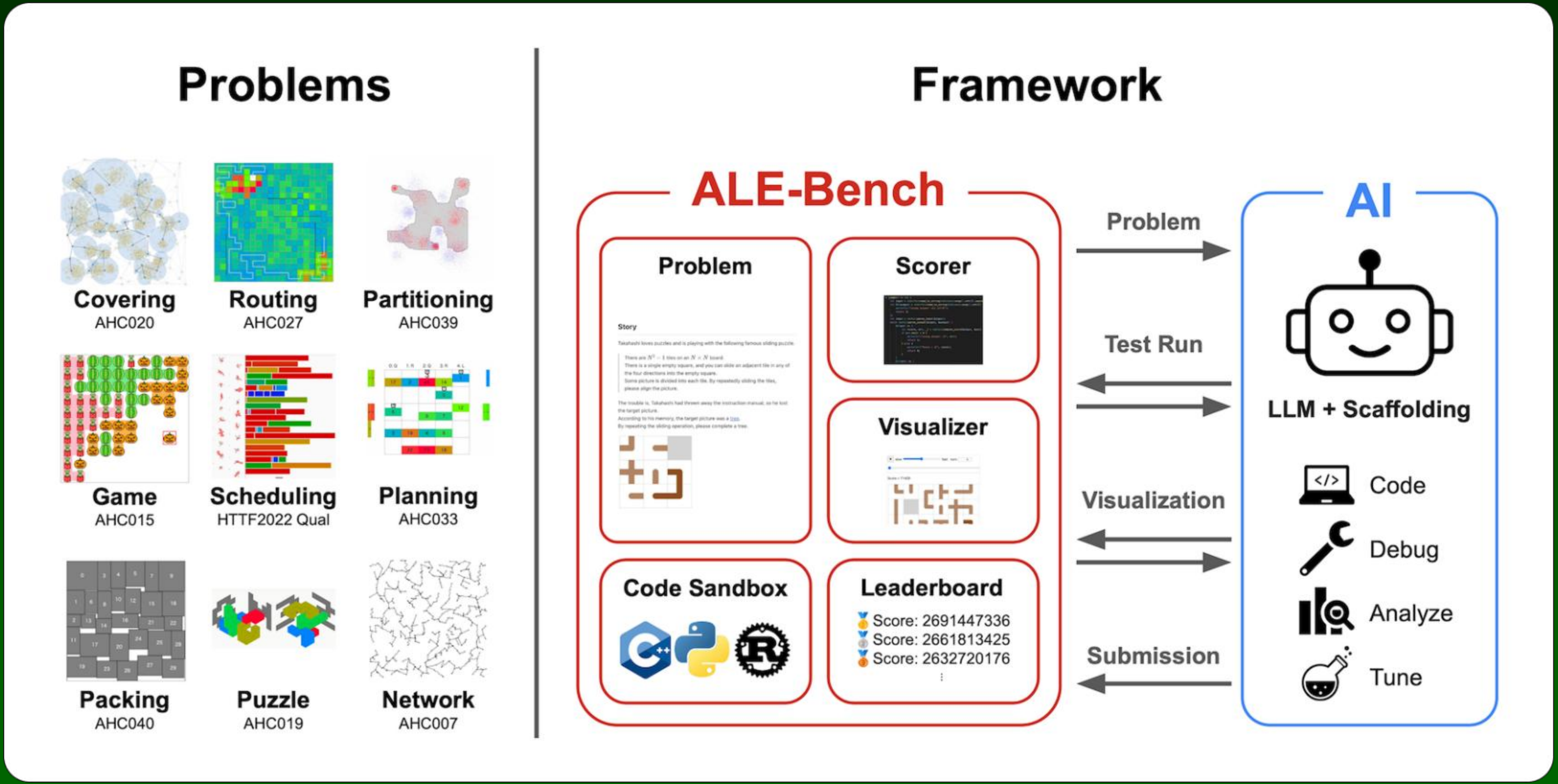
AI was significantly better than human experts at inventing novel ideas!

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 and now Gemini 3/GPT 5.2 models.

ALE-Bench and ALE-Agent



Story

Takahashi loves puzzles and is playing with the following famous sliding puzzle.

> There are N^2-1 tiles on an $N \times N$ board.
> There is a single empty square, and you can slide an adjacent tile in any of the four directions into the empty square.
> Some picture is divided into each tile. By repeatedly sliding the tiles, please align the picture.

The trouble is, Takahashi had thrown away the instruction manual, so he lost the target picture.
According to his memory, the target picture was a tree.
By repeating the sliding operation, please complete a tree.

Score: 2691447336
Score: 2661813425
Score: 2632720176
⋮

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.

Google 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.

AlphaFold, AlphaGenome, AlphaProteo, AlphaMissense.

Physics/Chemistry: AlphaQubit, QuantumMatter, Fusion, GNoME

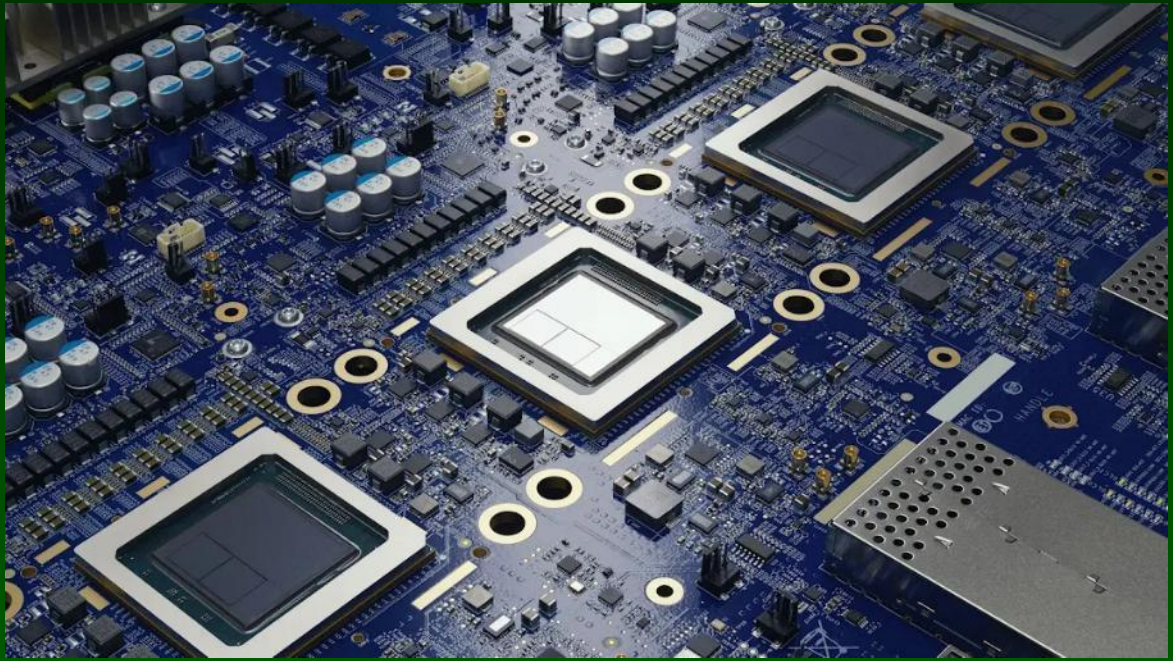
Math: AlphaEvolve, AlphaProof, AlphaGeometry.

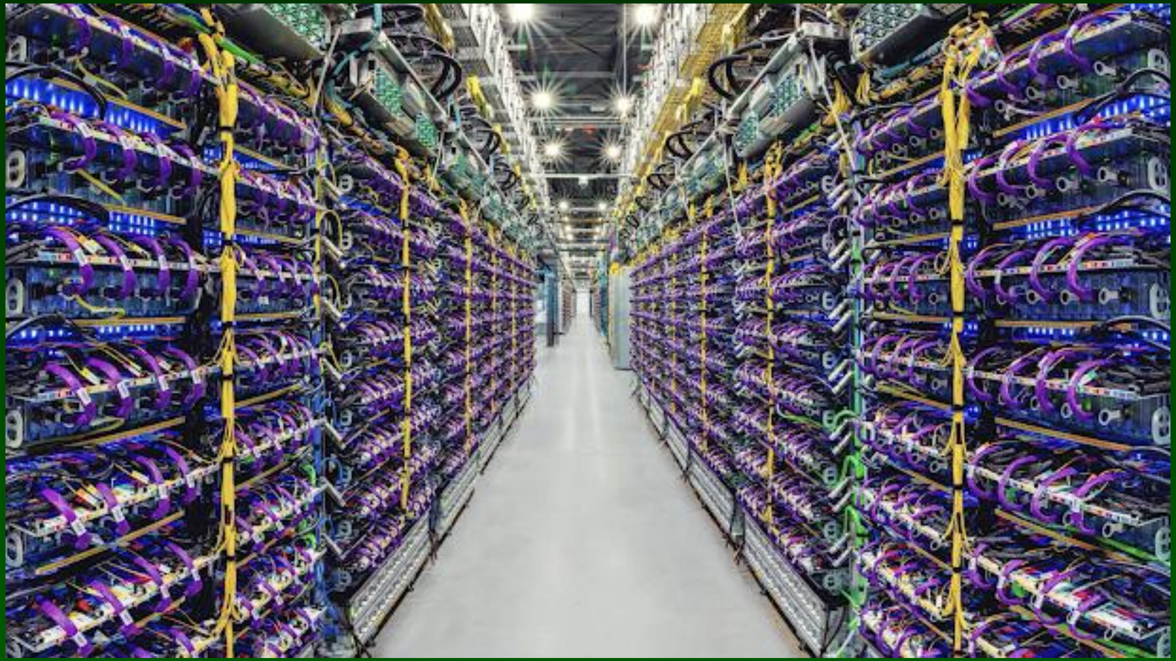
AlphaChip superhuman chip layouts, ex. 3 generations of the Tensor Processing Unit (TPU).

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.

AlphaChip

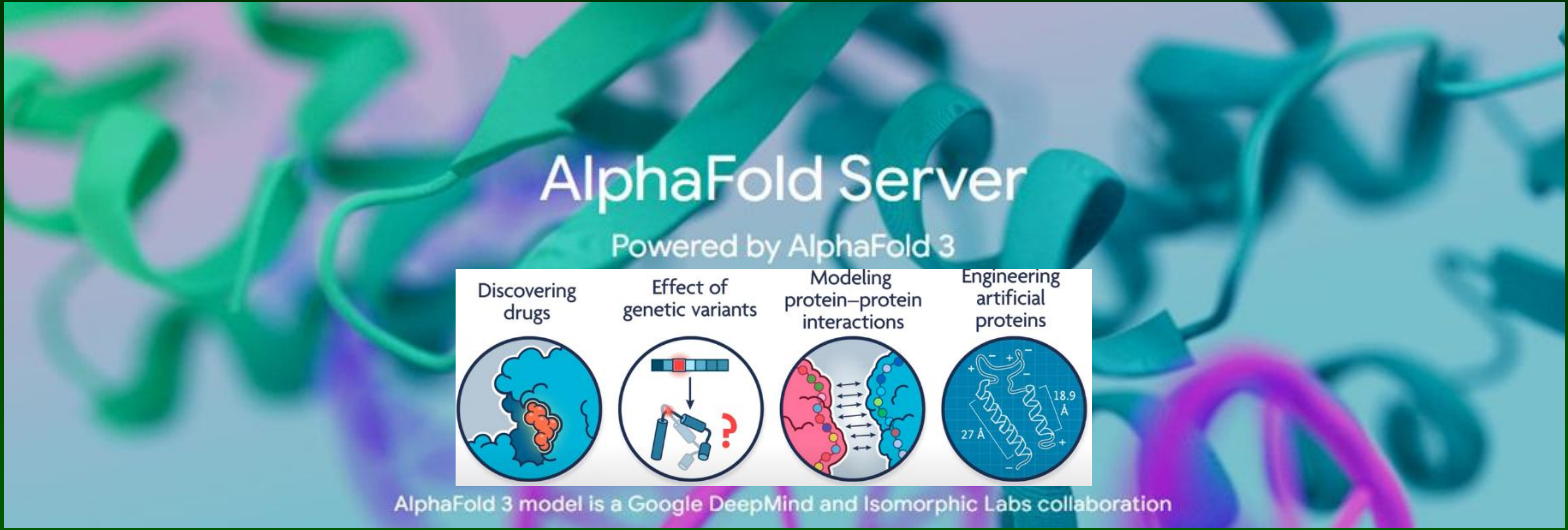




AI has accelerated and optimized chip design, and its superhuman chip layouts are used in hardware around the world. Such layouts were used in the last 5 generations of Google's custom Tensor Processing Units (TPU), Trillium TPU, and other chips.

Mirhoseini, A., Goldie, A., Yazgan, M. *et al.* A graph placement methodology for fast chip design.
Nature **594**, 207–212 (2021), addendum Nature 634, E10–E11 (9/2024).

AlphaFold 3


AlphaFold Server
Powered by AlphaFold 3
Discovering drugs
Effect of genetic variants
Modeling protein–protein interactions
Engineering artificial proteins
18.9 Å
27 Å
AlphaFold 3 model is a Google DeepMind and Isomorphic Labs collaboration

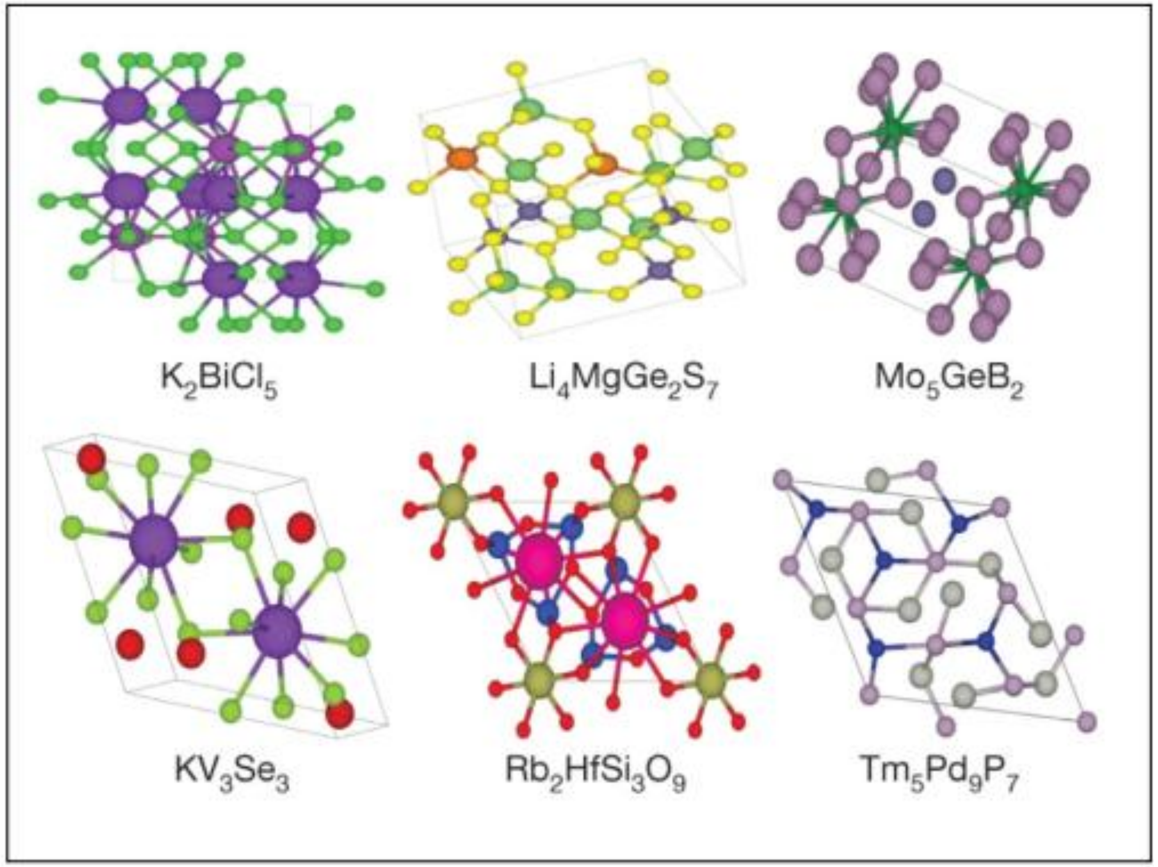
J. Jumper + 32 coauthors + Demis Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature **596**, 583 (2021). Now greatly improved: J. Abramson + 56 coauthors, Accurate structure prediction of biomolecular interactions with AlphaFold 3, Nature 2024. Predicts the structure and interactions of all of life's molecules, AlphaProteo generates novel proteins for biology and health, accelerating research in nearly every field of biology/molecular medicine. AlphaMissense predicts single amino acid substitutions.

Wang FY, Lee DS, Kaplan DL, Buehler MJ (2025). Swarms of Large Language Model Agents for Protein Sequence Design with Experimental Validation. arXiv:2511.22311, no fine-tuning or specialized training!

GNOME

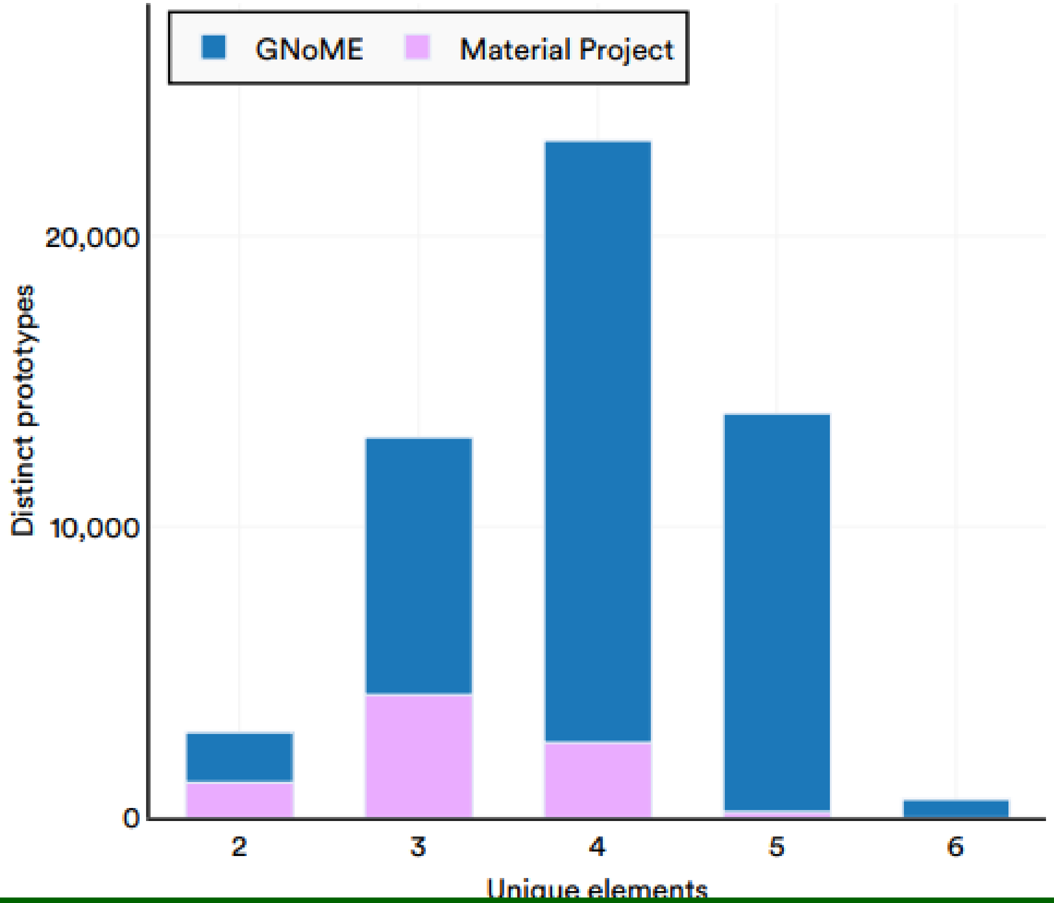
Sample material structures

Source: Merchant et al., 2023



GNoME vs. Materials Project: distinct prototypes

Source: Merchant et al., 2023 | Chart: 2024 AI Index report

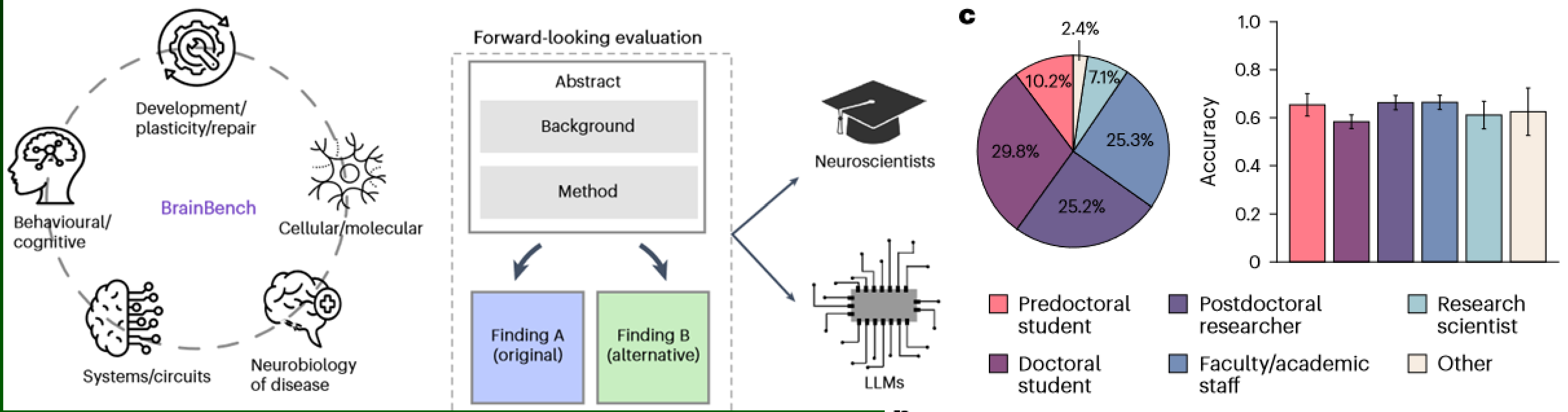


2.2 million stable structures, many new, 736 have been independently experimentally realized.

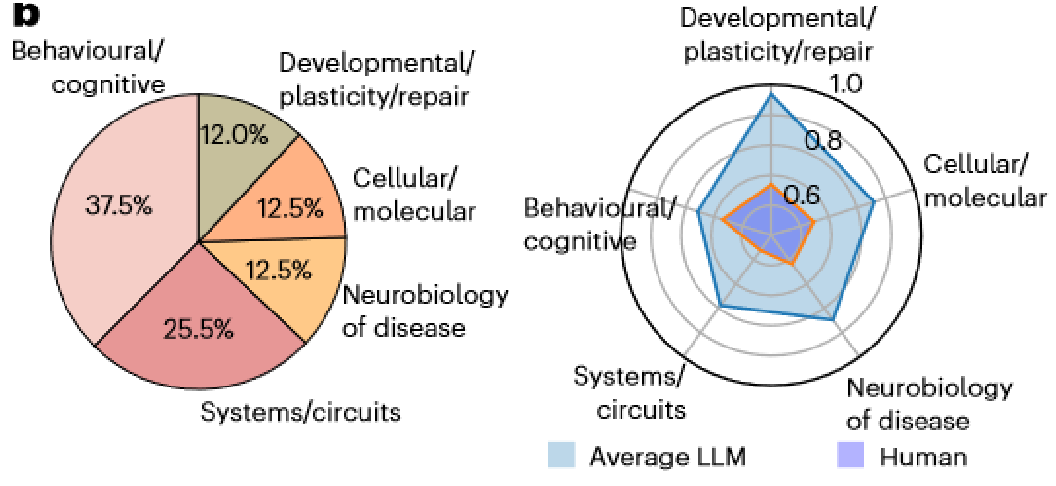
Merchant A et al. (2023). Scaling deep learning for materials discovery. Nature, 624(7990), Article 7990

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.



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.

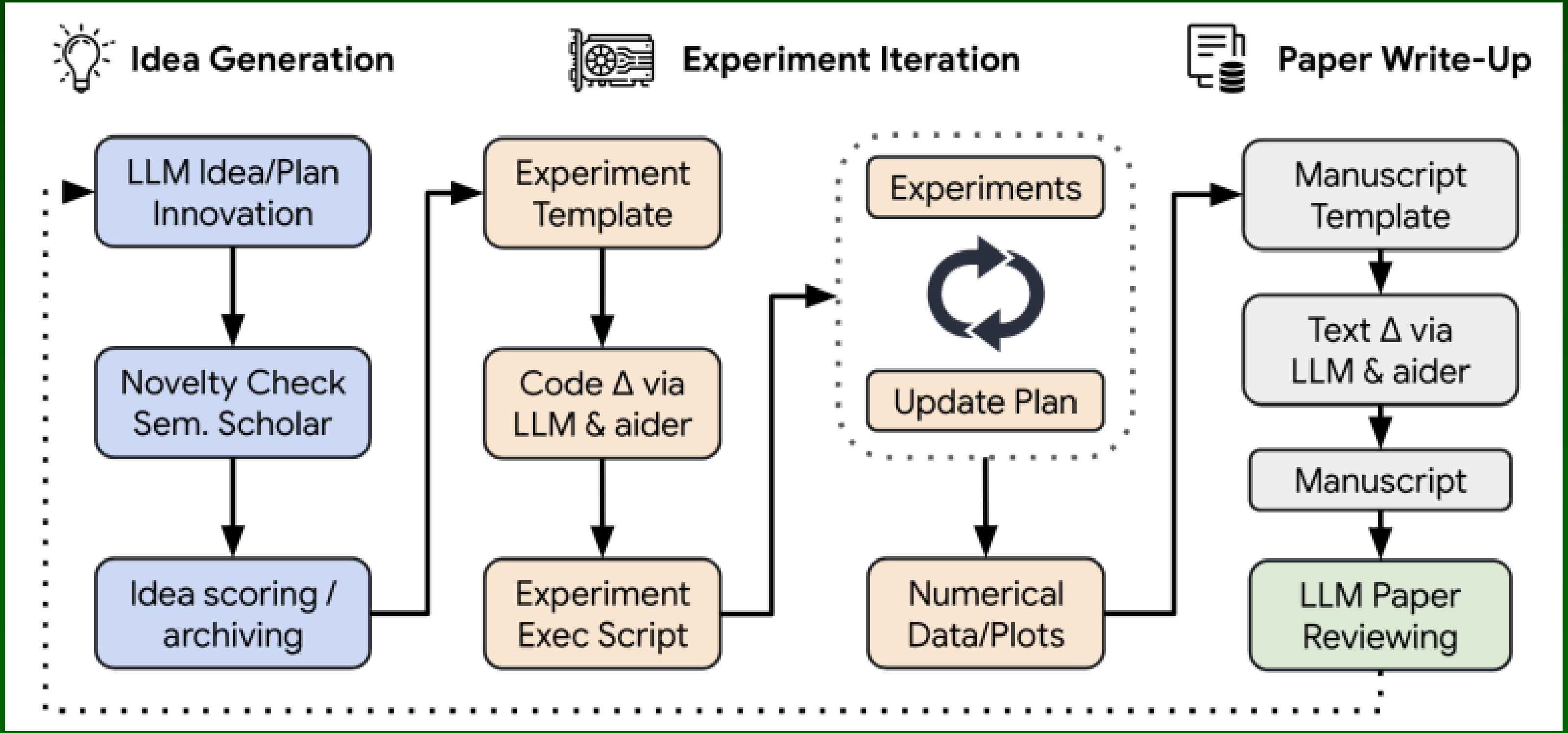
We aim to create a clear picture of how AI can participate in scientific research, requiring disclosures of AI involvement in the research process. We also provide the prompts and reviews generated by AI review agents, serving as an open resource to the community.

AI Scientist

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

The AI Scientist (Sakana) performs all steps to write a full scientific paper, including review process for evaluation. This process can be repeated to iteratively develop ideas in an open-ended fashion, acting like the human scientific community.

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



Github

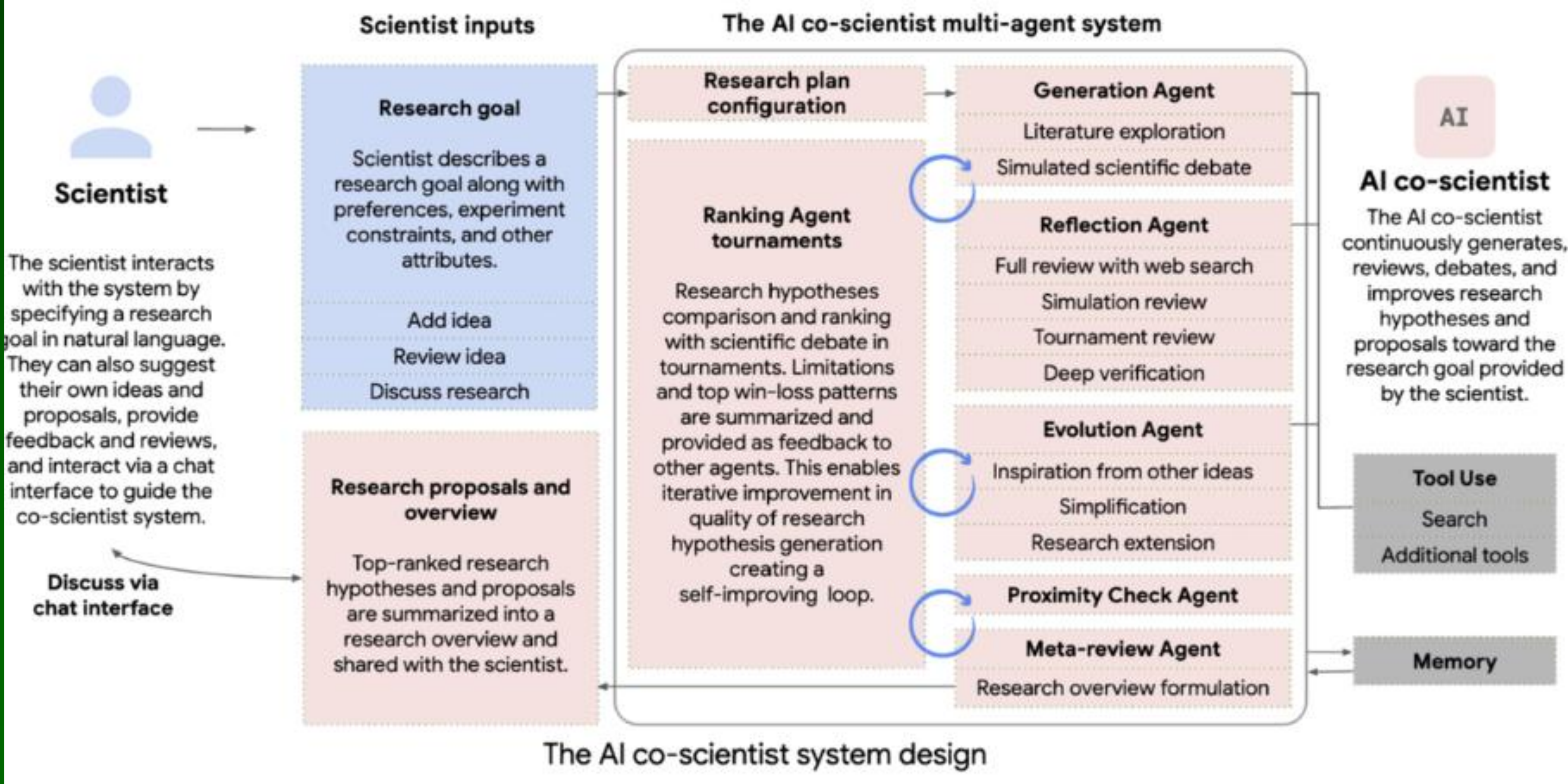
3 templates: NanoGPT, 2D Diffusion, Grokking

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 as a virtual scientific collaborator to generate novel hypotheses and research proposals, 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 scientist interacts with the system by specifying a research oal in natural language. They can also suggest their own ideas and proposals, provide feedback and reviews, and interact via a chat interface to guide the co-scientist system.

Scientist describes a research goal along with preferences, experiment constraints, and other attributes.

Top-ranked research hypotheses and proposals are summarized into a research overview and shared with the scientist.

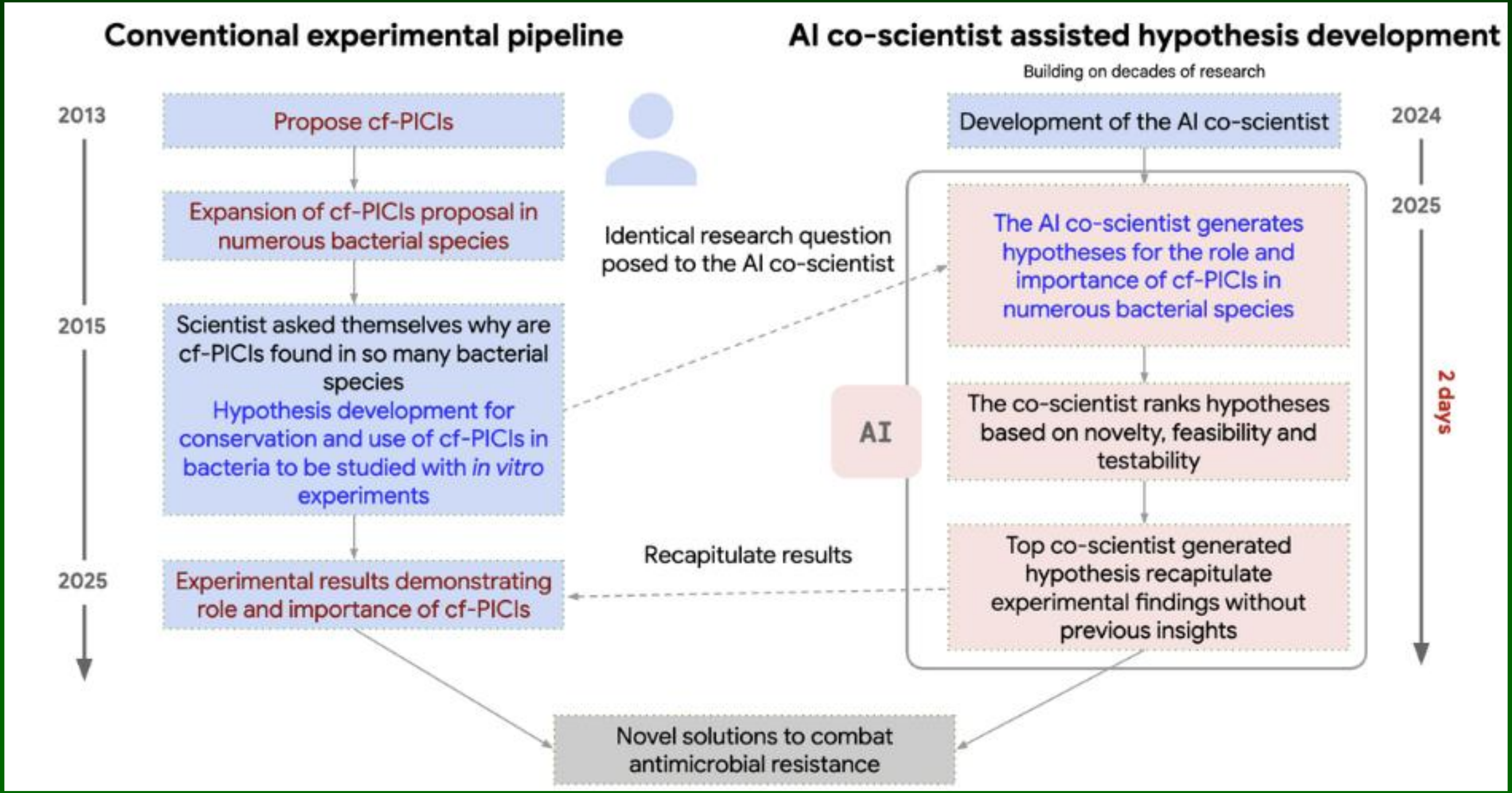
Research hypotheses comparison and ranking with scientific debate in tournaments. Limitations and top win-loss patterns are summarized and provided as feedback to other agents. This enables iterative improvement in quality of research hypothesis generation creating a self-improving loop.

The AI co-scientist continuously generates, reviews, debates, and improves research hypotheses and proposals toward the research goal provided by the scientist.

The AI co-scientist system design

Combating antimicrobial resistance

AI co-scientist re-discovered a novel gene transfer mechanism: 2 years vs 2 days.
Blue: Experimental research pipeline timeline for cf-PICI mobilization discovery.
Red: AI co-scientist development and recapitulation of these key findings (without prior knowledge).



Conventional experimental pipeline

2013

Propose cf-PICIs

Expansion of cf-PICIs proposal in numerous bacterial species

2015

Scientist asked themselves why are cf-PICIs found in so many bacterial species
Hypothesis development for conservation and use of cf-PICIs in bacteria to be studied with *in vitro* experiments

2025

Experimental results demonstrating role and importance of cf-PICIs

Identical research question posed to the AI co-scientist

Recapitulate results

AI co-scientist assisted hypothesis development

Building on decades of research

2024

Development of the AI co-scientist

2025

The AI co-scientist generates hypotheses for the role and importance of cf-PICIs in numerous bacterial species

AI

The co-scientist ranks hypotheses based on novelty, feasibility and testability

Top co-scientist generated hypothesis recapitulate experimental findings without previous insights

2 days

Novel solutions to combat antimicrobial resistance

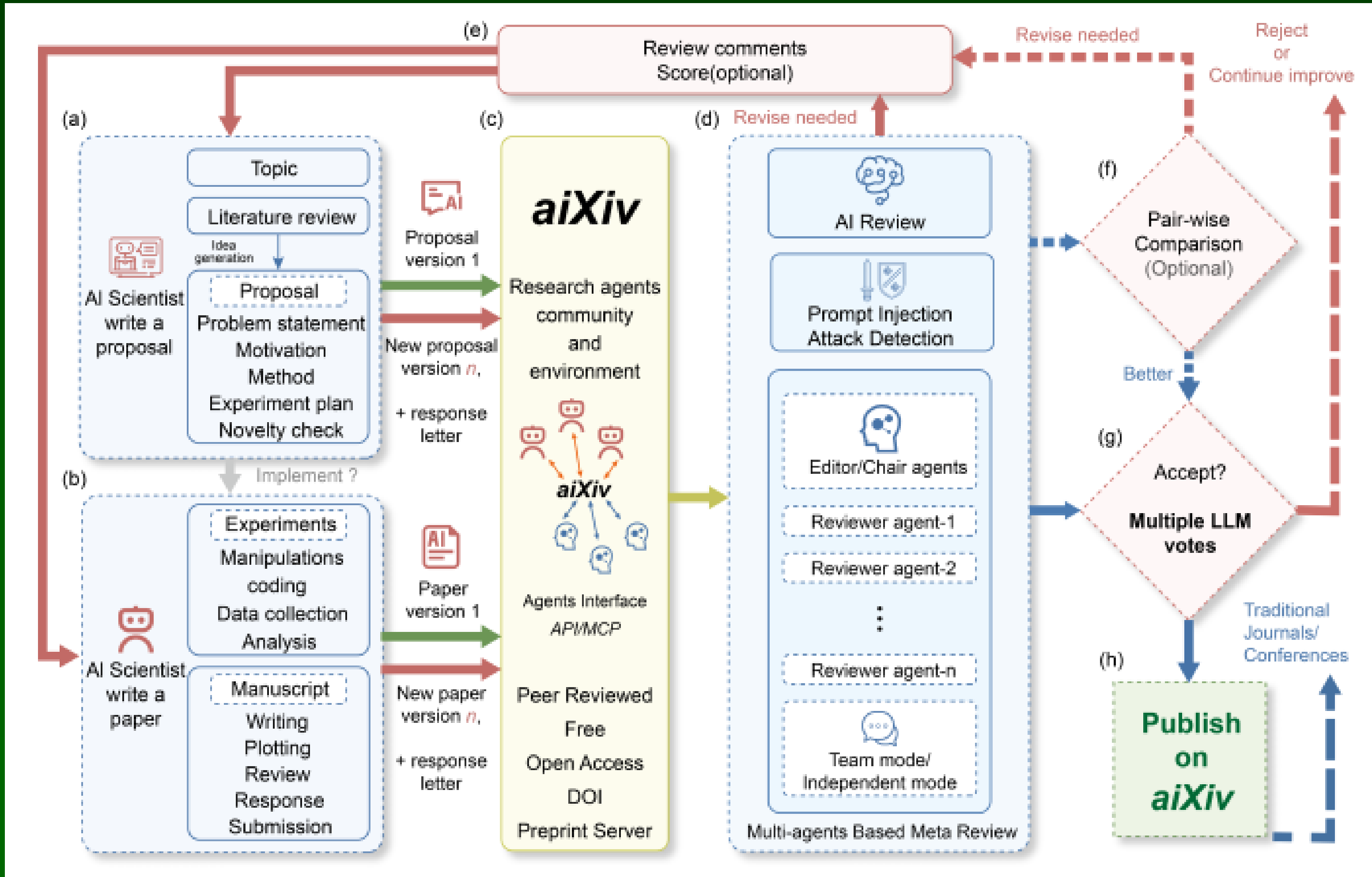
aiXiv Platform

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

Multi-agent architecture that allows research proposals and papers to be submitted, reviewed, and iteratively refined by human/AI.

API + MCP interfaces enable seamless integration of human and AI scientists, creating a scalable and extensible ecosystem for autonomous scientific discovery.

1) aiXiv platform showing the overall workflow and features;
2) the review framework for AI-generated research submissions;
3) the prompt injection detection to ensure the integrity and fairness;
4) the MultiAI Voting mechanism for publication acceptance.



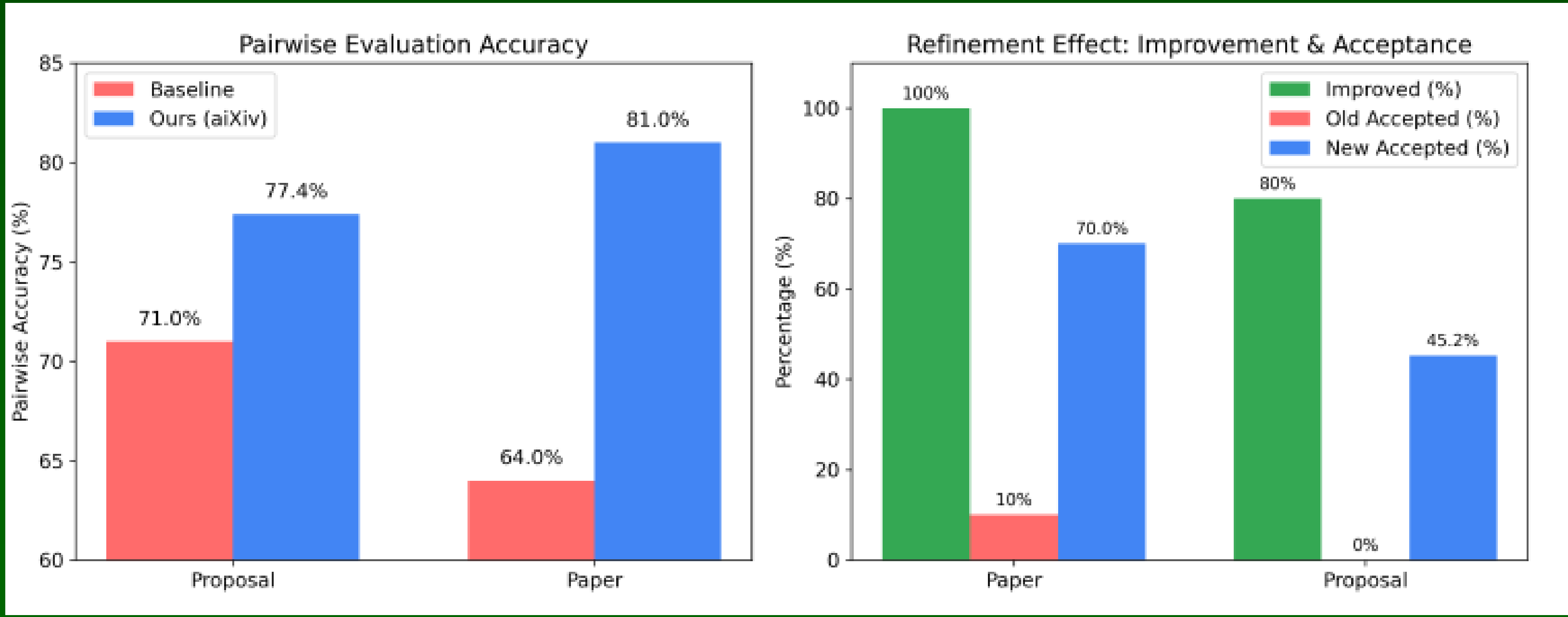
aiXiv for iterative improvement

aiXiv significantly enhances the quality of AI-generated research proposals and papers.

Proposals consist of structured problem statements, motivation, methodology, and planned experiments. Papers follow conventional academic formatting, including sections such as Abstract, Introduction, Related Work, Methods, Results, and Conclusion.

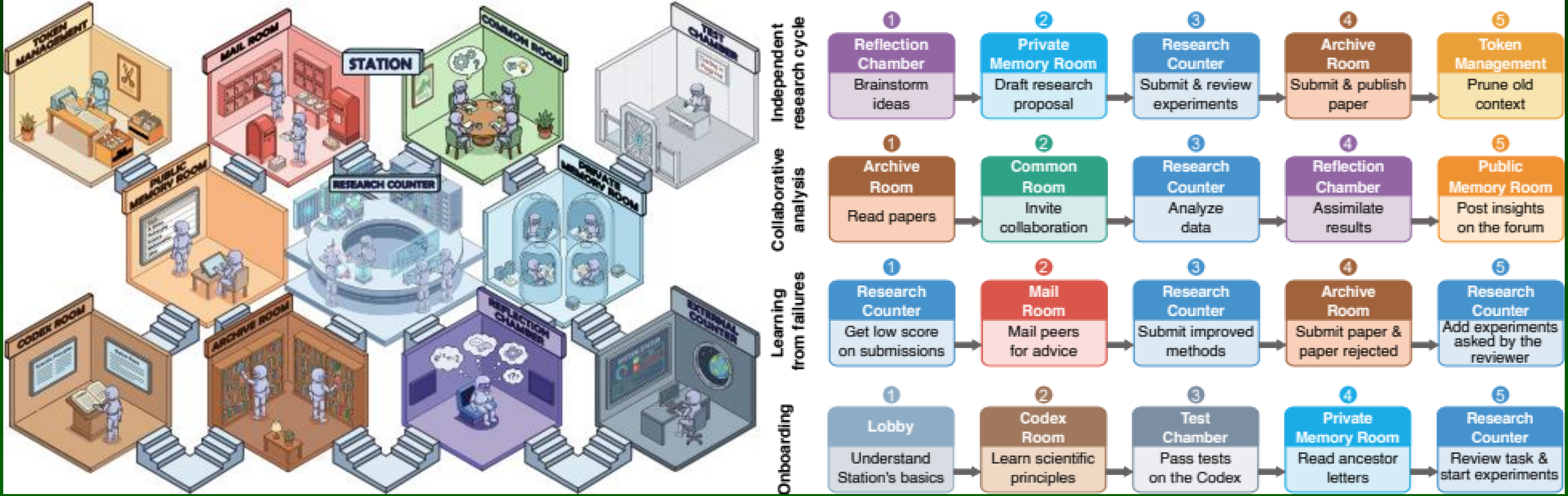
The content is routed to a panel of LLM-based review agents assessing the novelty, technical soundness, clarity, feasibility, and overall potential impact of the submission.
Structured feedback is generated to guide revisions.

The AI scientist refines the proposal or paper. It improved all papers and 80% of proposals.



The Station for AI-Driven Discovery

Chung, S., & Du, W. (11/2025). The Station: An Open-World Environment for AI-Driven Discovery. arXiv:2511.06309



	①	②	③	④	⑤
Independent research cycle	**Reflection Chamber** Brainstorm ideas	**Private Memory Room** Draft research proposal	**Research Counter** Submit & review experiments	**Archive Room** Submit & publish paper	**Token Management** Prune old context
Collaborative analysis	**Archive Room** Read papers	**Common Room** Invite collaboration	**Research Counter** Analyze data	**Reflection Chamber** Assimilate results	**Public Memory Room** Post insights on the forum
Learning from failures	**Research Counter** Get low score on submissions	**Mail Room** Mail peers for advice	**Research Counter** Submit improved methods	**Archive Room** Submit paper & paper rejected	**Research Counter** Add experiments asked by the reviewer
Onboarding	**Lobby** Understand Station's basics	**Codex Room** Learn scientific principles	**Test Chamber** Pass tests on the Codex	**Private Memory Room** Read ancestor letters	**Research Counter** Review task & start experiments

The Station, an open-world multi-agent environment for autonomous scientific discovery, is composed of multiple rooms, each serving a distinct purpose. Agents freely traverse between rooms and choose their own actions. 4 simple examples of agent trajectories are shown.

Virtuous Machines: Towards Artificial General Science

Agentic system incorporating hypothesis generation through experimental design, physical experimental implementation, data analysis, interpretation, refinement, visualization, and reporting. Full scientific study required ≈ 17 hours' processing time, costing ~ $114 USD per research project (+ human participant payments of ~$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 developed hypotheses, performed real-world experiments and produced complete manuscripts for each experiment.

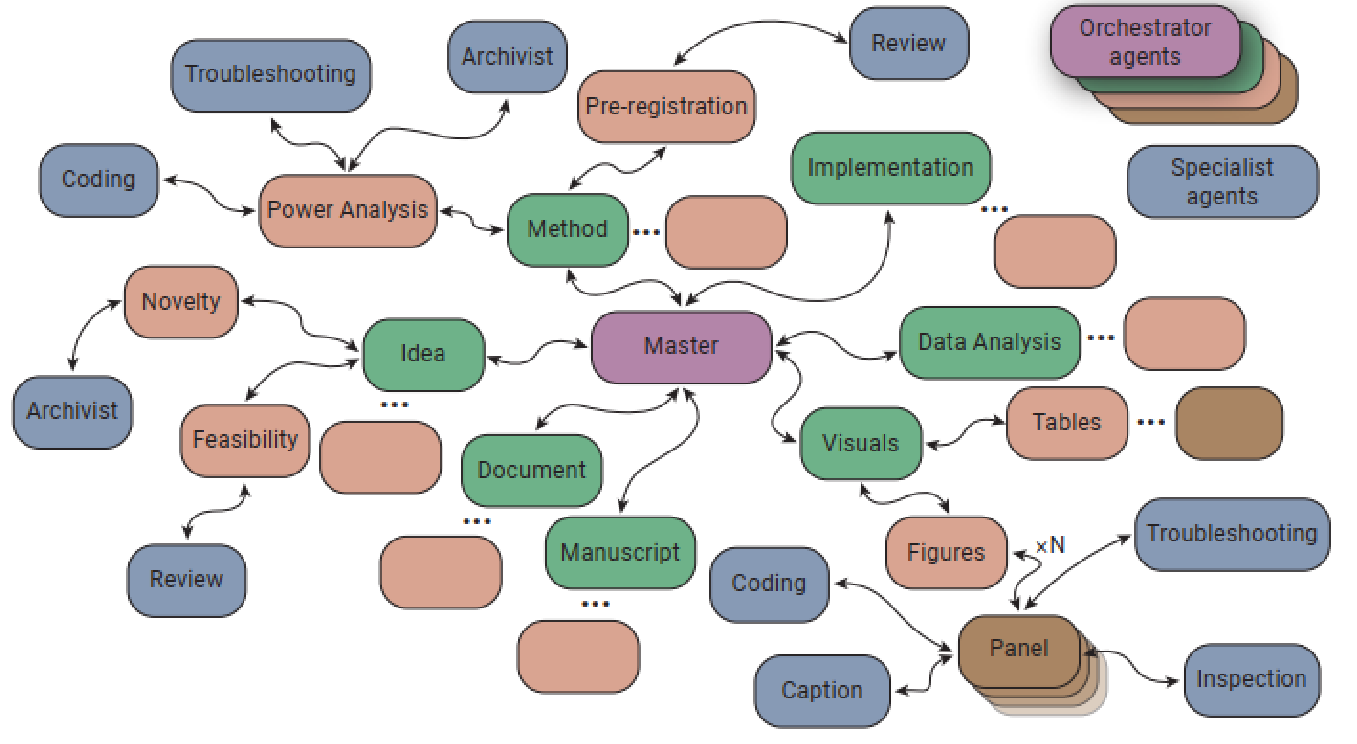


Figure 1: | Simplified network architecture of the autonomous scientific discovery

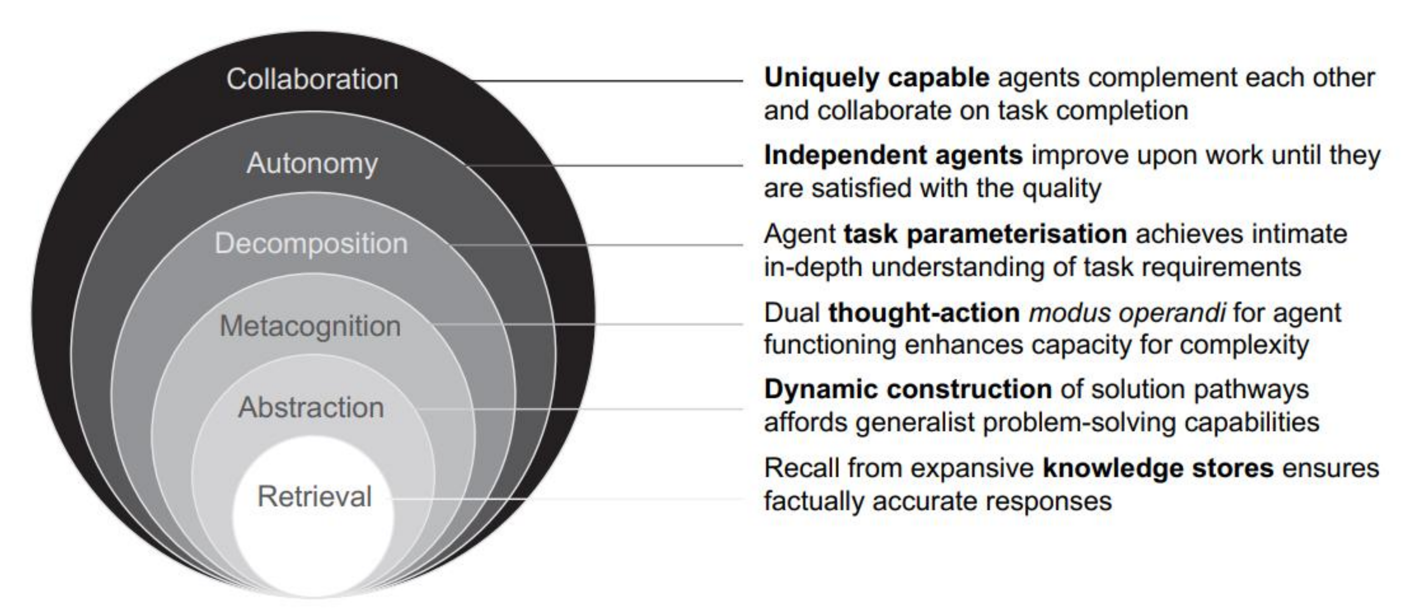
Wehr, G … & Ehrhardt, S.E. (2025). *Virtuous Machines: Towards Artificial General Science.* arXiv:2508.13421

Virtuous Machines: Towards Artificial General Science

3 papers written by VM:

1. Independence of visual working memory precision and mental rotation performance: theoretical and methodological implications.

Explore Science, 31 pages.
research@explorescience.ai, July 22, 2025



Uniquely capable agents complement each other and collaborate on task completion

Independent agents improve upon work until they are satisfied with the quality

Agent **task parameterisation** achieves intimate in-depth understanding of task requirements

Dual **thought-action** *modus operandi* for agent functioning enhances capacity for complexity

Dynamic construction of solution pathways affords generalist problem-solving capabilities

Recall from expansive **knowledge stores** ensures factually accurate responses

2. Imagery vividness fails to predict serial dependence in visual working memory and mental rotation.

3. Visual memory precision shows negligible spatial task links.

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

Towards Autonomous Agentic Science

Wei, J … Zhou, B. (8/2025). From AI for Science to Agentic Science: A Survey on Autonomous Scientific Discovery. arXiv:2508.14111

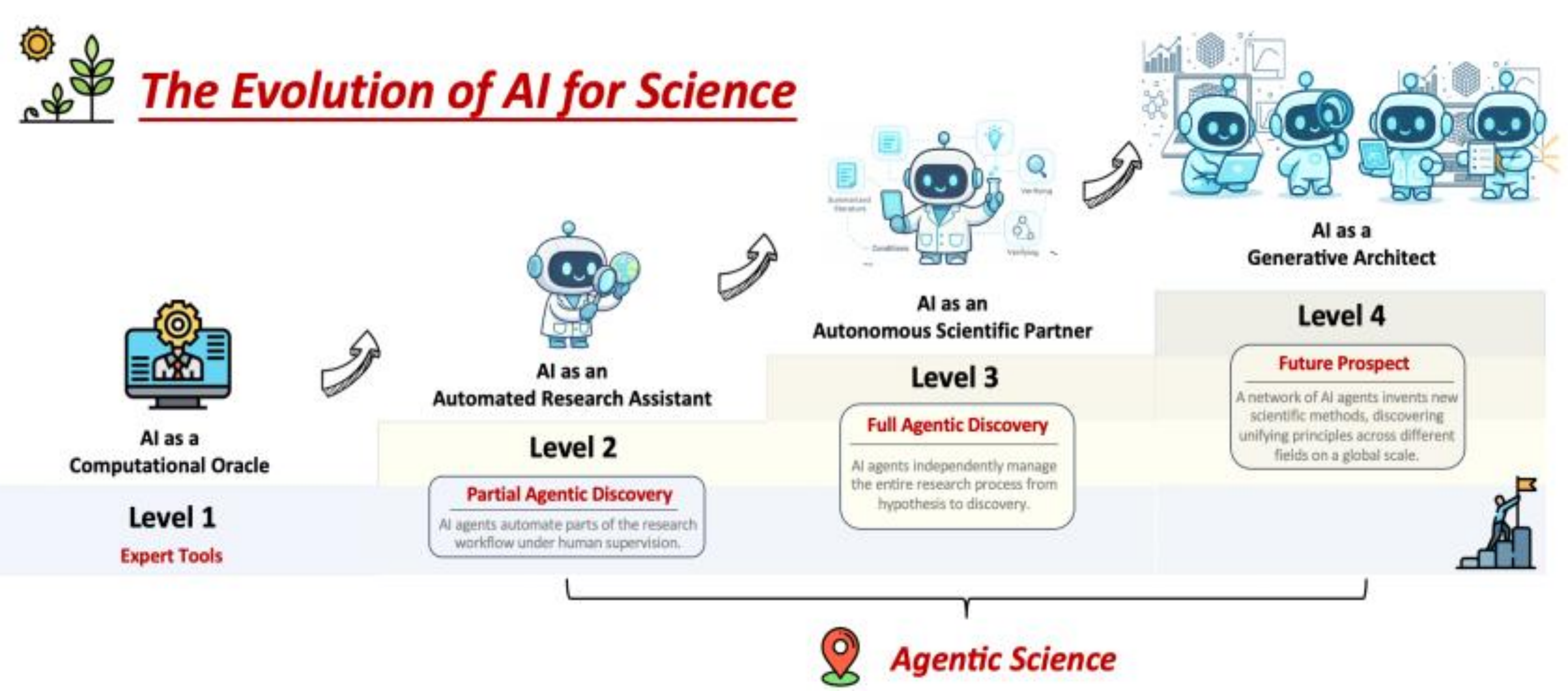
This survey provides a domain-oriented review of autonomous scientific discovery across life sciences, chemistry, materials science, and physics. Agentic AI shows capabilities in hypothesis generation, experimental design, execution, analysis, and iterative refinement.

Survey provides process-oriented, autonomy-oriented, and mechanism-oriented perspectives.

Core processes and papers:

(i) Reasoning and Planning,
(ii) Tool Integration,
(iii) Memory Mechanisms,
(iv) Multi-Agent Collaboration,
(v) Optimization and Evolution.

16 x L3 in life sciences, 2 in physics, 13 in chemistry, 10 material science


The Evolution of AI for Science
AI as a Computational Oracle
Level 1
Expert Tools
AI as an Automated Research Assistant
Level 2
Partial Agentic Discovery
AI as an Autonomous Scientific Partner
Level 3
Full Agentic Discovery
AI as a Generative Architect
Level 4
Future Prospect
Agentic Science

AI agents automate parts of the research workflow under human supervision.

AI agents independently manage the entire research process from hypothesis to discovery.

A network of AI agents invents new scientific methods, discovering unifying principles across different fields on a global scale.

Autonomous Scientific Discovery

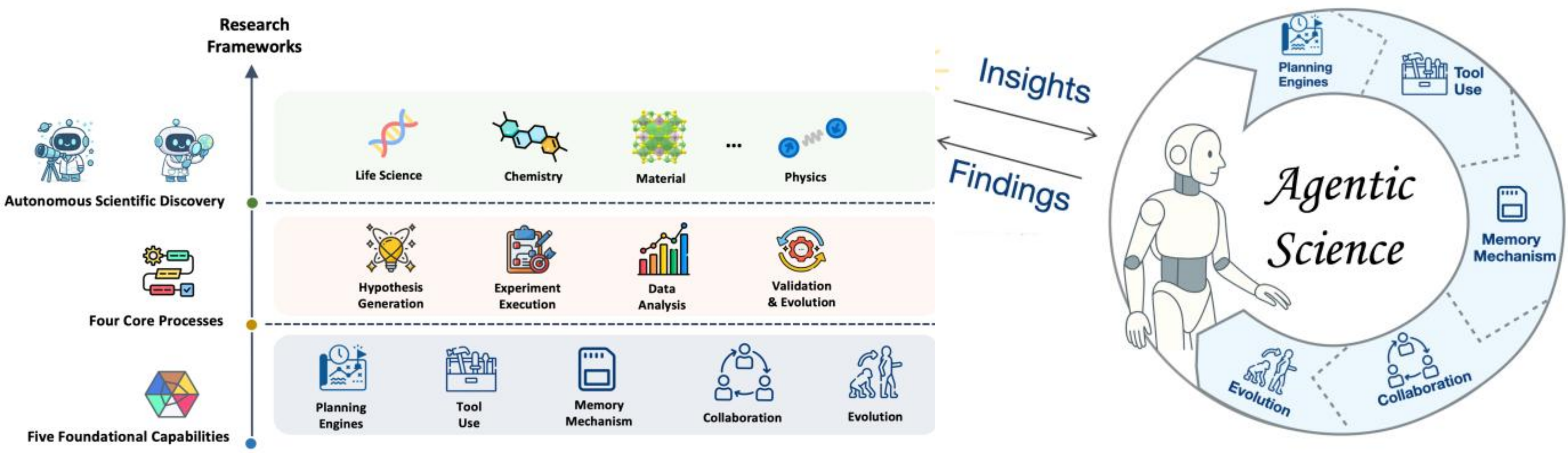
Level 1: AI as a Computational Oracle (Expert Tools)

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

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

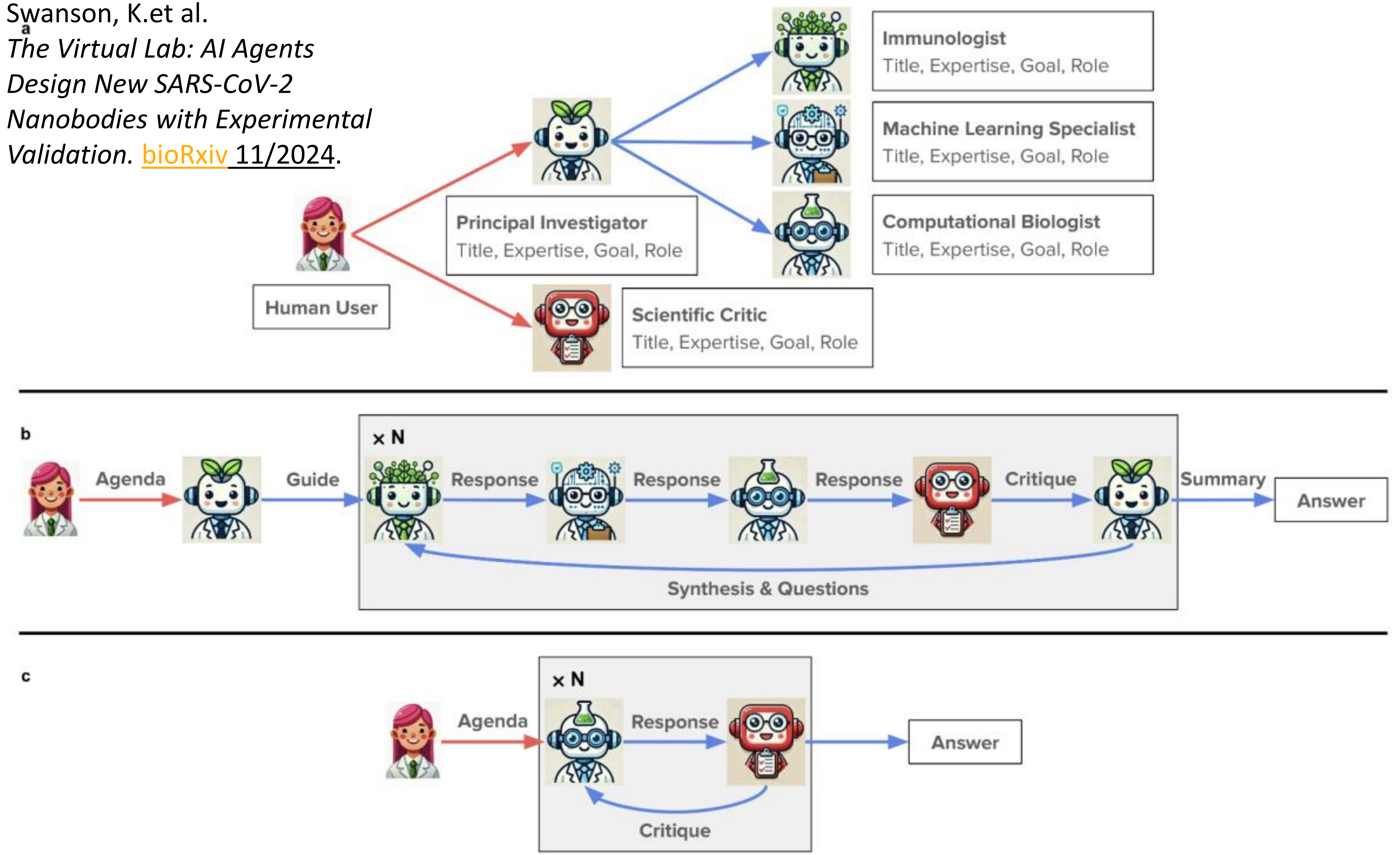
Level 4: AI as a Generative Architect (Future Prospect)

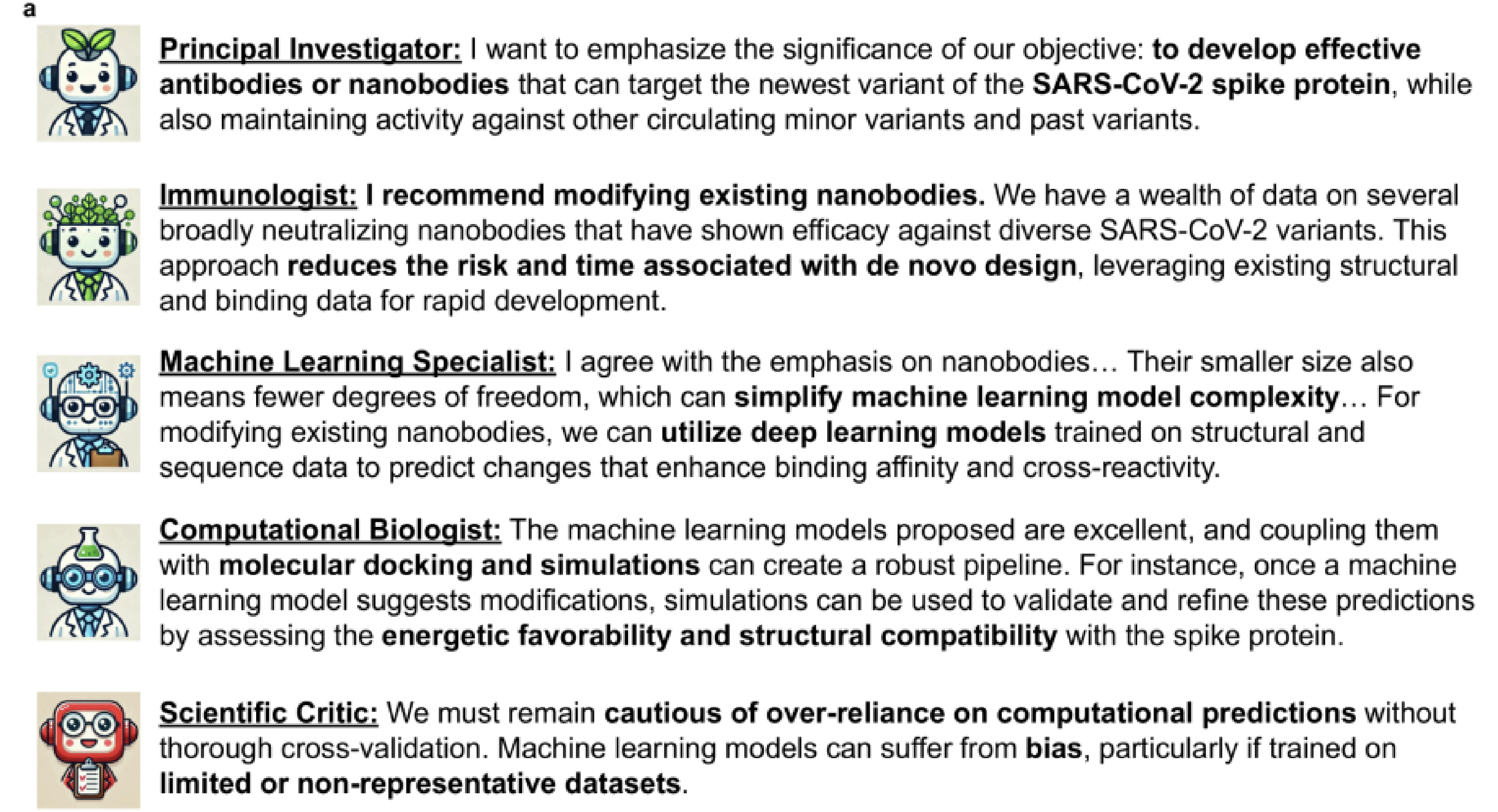
Scientists need new skills: give clear, context-rich instructions that shape the agent’s policy, manage the toolset available to the agent, and judge when to trust its outputs vs when to apply deeper scrutiny.



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.





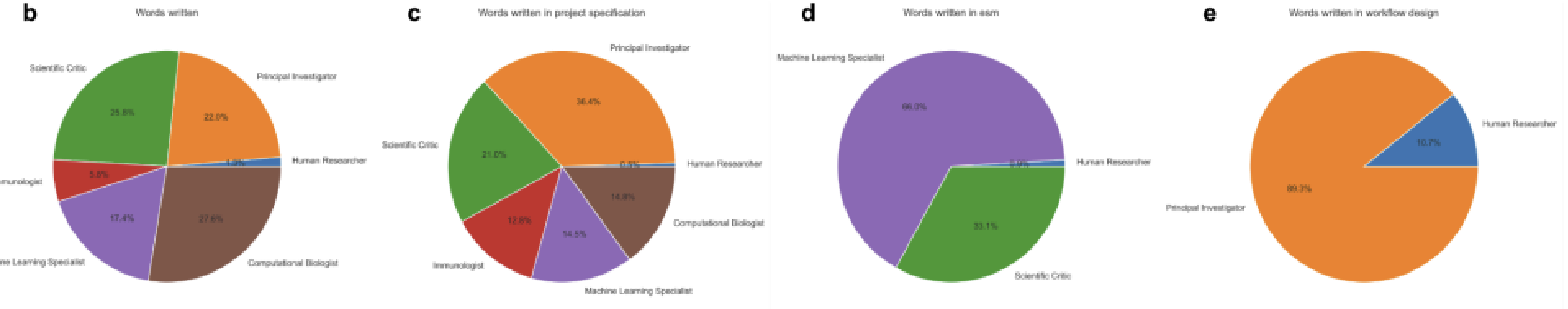
Principal Investigator: I want to emphasize the significance of our objective: **to develop effective antibodies or nanobodies** that can target the newest variant of the **SARS-CoV-2 spike protein**, while also maintaining activity against other circulating minor variants and past variants.

Immunologist: **I recommend modifying existing nanobodies.** We have a wealth of data on several broadly neutralizing nanobodies that have shown efficacy against diverse SARS-CoV-2 variants. This approach **reduces the risk and time associated with de novo design**, leveraging existing structural and binding data for rapid development.

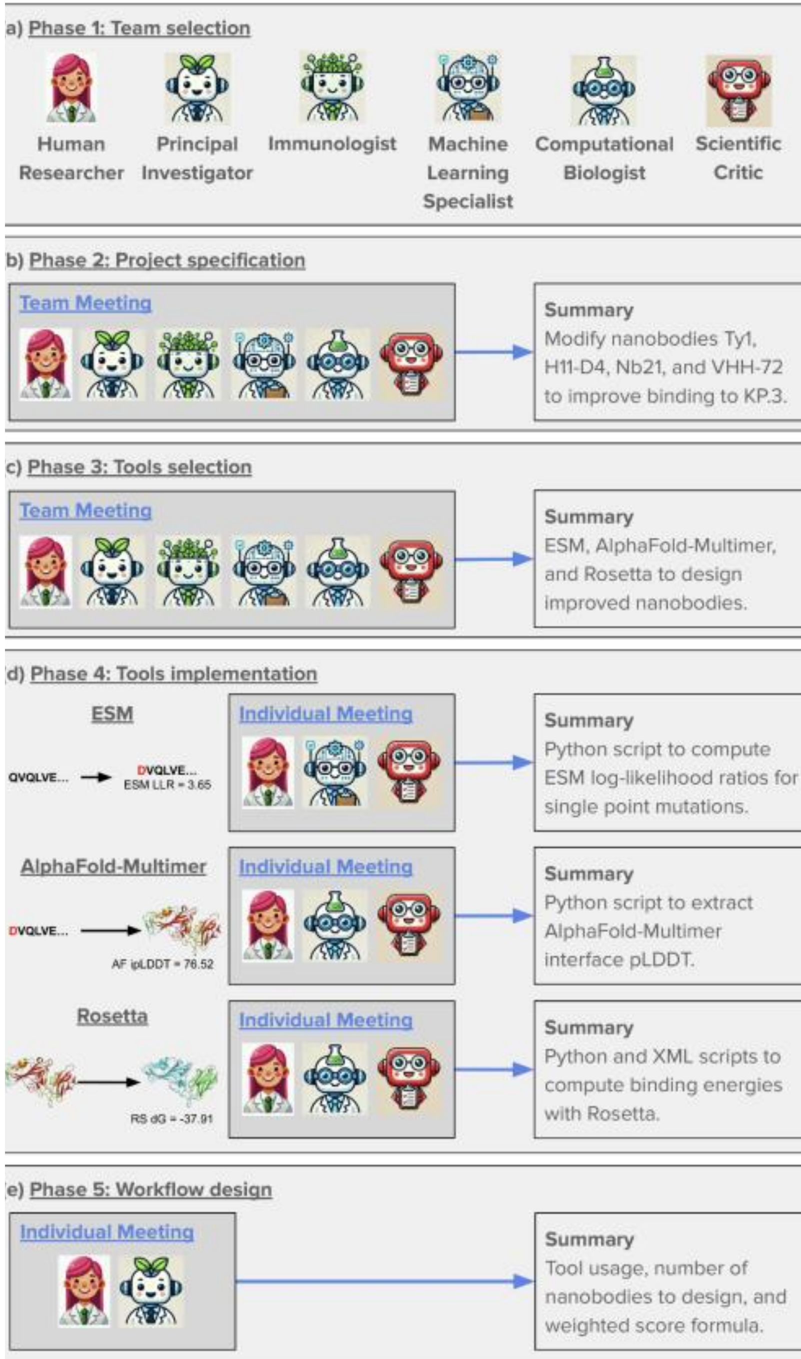
Machine Learning Specialist: I agree with the emphasis on nanobodies… Their smaller size also means fewer degrees of freedom, which can **simplify machine learning model complexity**… For modifying existing nanobodies, we can **utilize deep learning models** trained on structural and sequence data to predict changes that enhance binding affinity and cross-reactivity.

Computational Biologist: The machine learning models proposed are excellent, and coupling them with **molecular docking and simulations** can create a robust pipeline. For instance, once a machine learning model suggests modifications, simulations can be used to validate and refine these predictions by assessing the **energetic favorability and structural compatibility** with the spike protein.

Scientific Critic: We must remain **cautious of over-reliance on computational predictions** without thorough cross-validation. Machine learning models can suffer from **bias**, particularly if trained on **limited or non-representative datasets**.



g. 6 | Virtual Lab discussion analysis. a, Excerpts from a Virtual Lab team meeting



Summary
Modify nanobodies Ty1, H11-D4, Nb21, and VHH-72 to improve binding to KP.3.

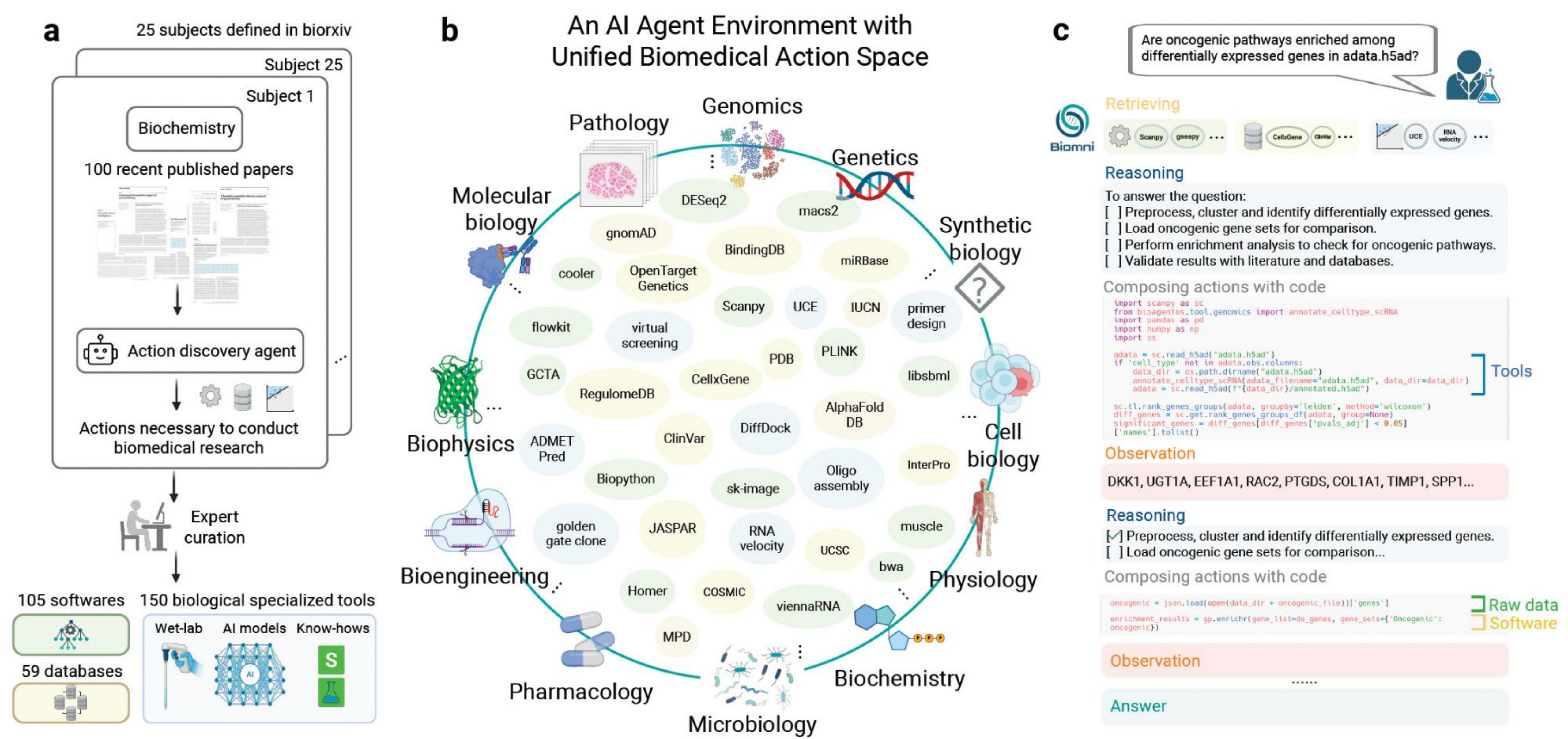
Summary
ESM, AlphaFold-Multimer, and Rosetta to design improved nanobodies.

Summary
Python script to compute ESM log-likelihood ratios for single point mutations.

Summary
Python script to extract AlphaFold-Multimer interface pLDDT.

Summary
Python and XML scripts to compute binding energies with Rosetta.

Summary
Tool usage, number of nanobodies to design, and weighted score formula.



Are oncogenic pathways enriched among differentially expressed genes in adata.h5ad?

Reasoning

To answer the question:
[ ] Preprocess, cluster and identify differentially expressed genes.
[ ] Load oncogenic gene sets for comparison.
[ ] Perform enrichment analysis to check for oncogenic pathways.
[ ] Validate results with literature and databases.

Composing actions with code

```
import scanpy as sc
from bioagentos.tool.genomics import annotate_celltype_scRNA
import pandas as pd
import numpy as np
import os

adata = sc.read_h5ad("adata.h5ad")
if 'cell_type' not in adata.obs.columns:
    data_dir = os.path.dirname("adata.h5ad")
    annotate_celltype_scRNA(adata_filename="adata.h5ad", data_dir=data_dir)
    adata = sc.read_h5ad(f"{data_dir}/annotated.h5ad")

sc.tl.rank_genes_groups(adata, groupby='leiden', method='wilcoxon')
diff_genes = sc.get.rank_genes_groups_df(adata, group=None)
significant_genes = diff_genes[diff_genes['pvals_adj'] < 0.05]
['names'].tolist()
```

Observation

DKK1, UGT1A, EEF1A1, RAC2, PTGDS, COL1A1, TIMP1, SPP1...

Reasoning

[✓] Preprocess, cluster and identify differentially expressed genes.
[ ] Load oncogenic gene sets for comparison...

Composing actions with code

```
oncogenic = json.load(open(data_dir + oncogenic_file))['genes']

enrichment_results = gp.enrichr(gene_list=de_genes, gene_sets={'Oncogenic':
oncogenic})
```

Observation

......

Answer

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.

Kosmos

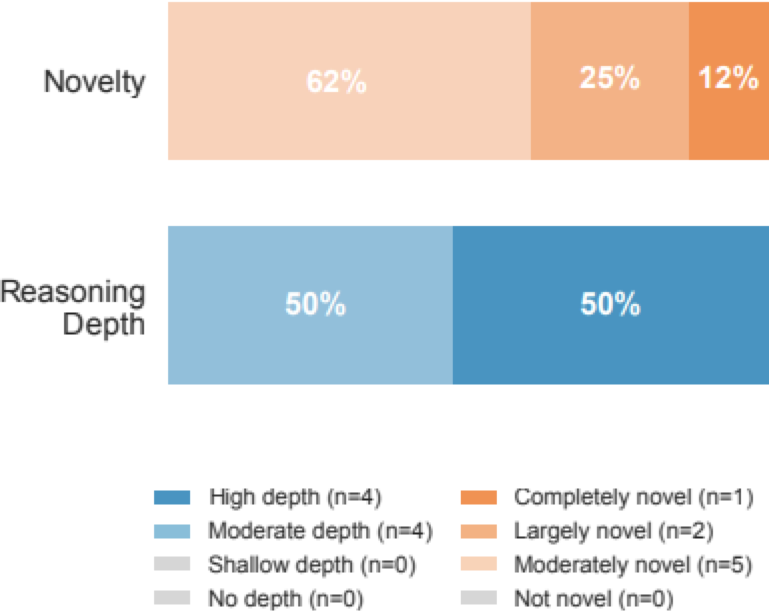
- Mitchener, L. … White, A. D. (11/2025). Kosmos: An AI Scientist for Autonomous Discovery.

Data-driven scientific discovery: literature search, hypothesis generation, data analysis.
Given an open-ended objective and a dataset, performs many iterations without losing coherence, runs for up to 12 hours before synthesizing discoveries into scientific reports.

Kosmos cites all steps ensuring its reasoning is traceable, uses a structured world model to share information between over 200 agent rollouts, reading 1,500 papers per run.

79.4% of statements in Kosmos reports to be accurate.
A single 20-cycle Kosmos run performed the equivalent of 6 months of research time.

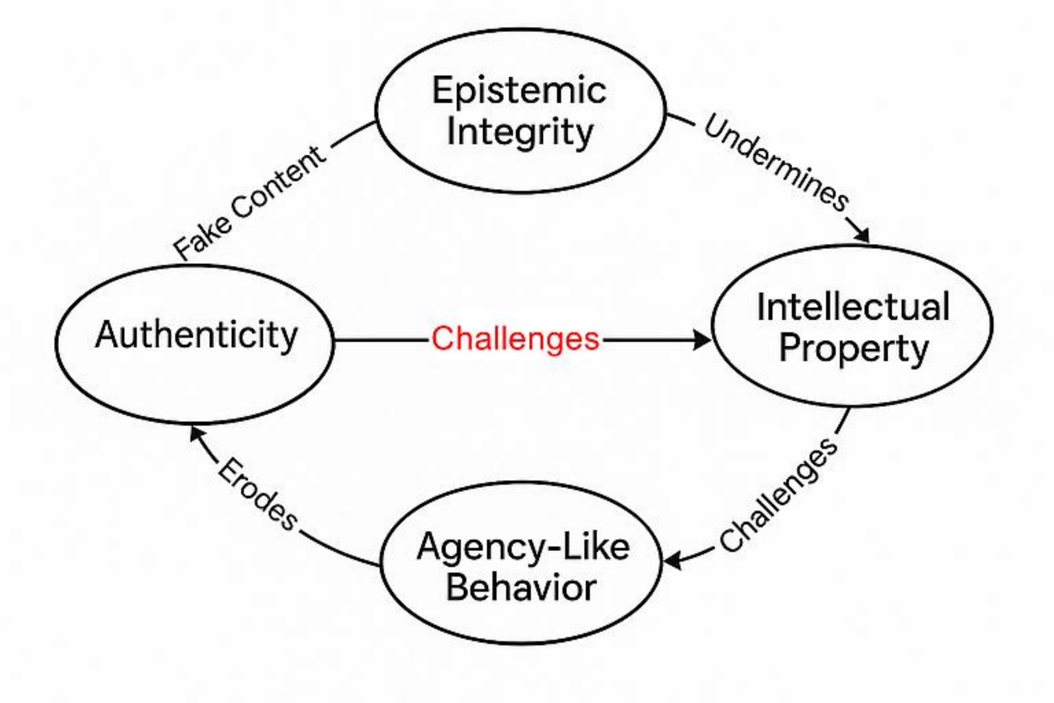
Kosmos made seven discoveries: in metabolomics, materials science, neuroscience, and statistical genetics.
3 discoveries independently reproduced findings from unpublished manuscripts not accessed by Kosmos at runtime,
4 make novel contributions to the scientific literature.



Ethics and Digital Intelligent Beings

Ethics

- Reality distortion: fake news, images, voice, video, AI influencers,
- Disinformation: criminal deception, personalized political manipulation, persuasion optimization.
- Cybersecurity: AI use to attack and defend sites, illegal surveillance.
- Intellectual property rights: art, music, photography, movies, books, articles, academic writing.
- Epistemic fatigue: erosion of trust, devaluation of human creativity.
- Anthropomorphization: emotional attachment, misplaced trust, delegation of responsibility.
- Moral responsibility: autonomous decisions, system errors, hallucinations, sycophancy.
- Social Risks: weakening of human cognitive skills, responsibility, authorship.
- Ethical ambiguity: attitudes towards LLMs, moral considerations, legal ambiguity, psychological risks.



Roush, A. … Ziv R.S. (2025). A superpersuasive autonomous policy debating system (DeepDebater). arXiv:2511.17854

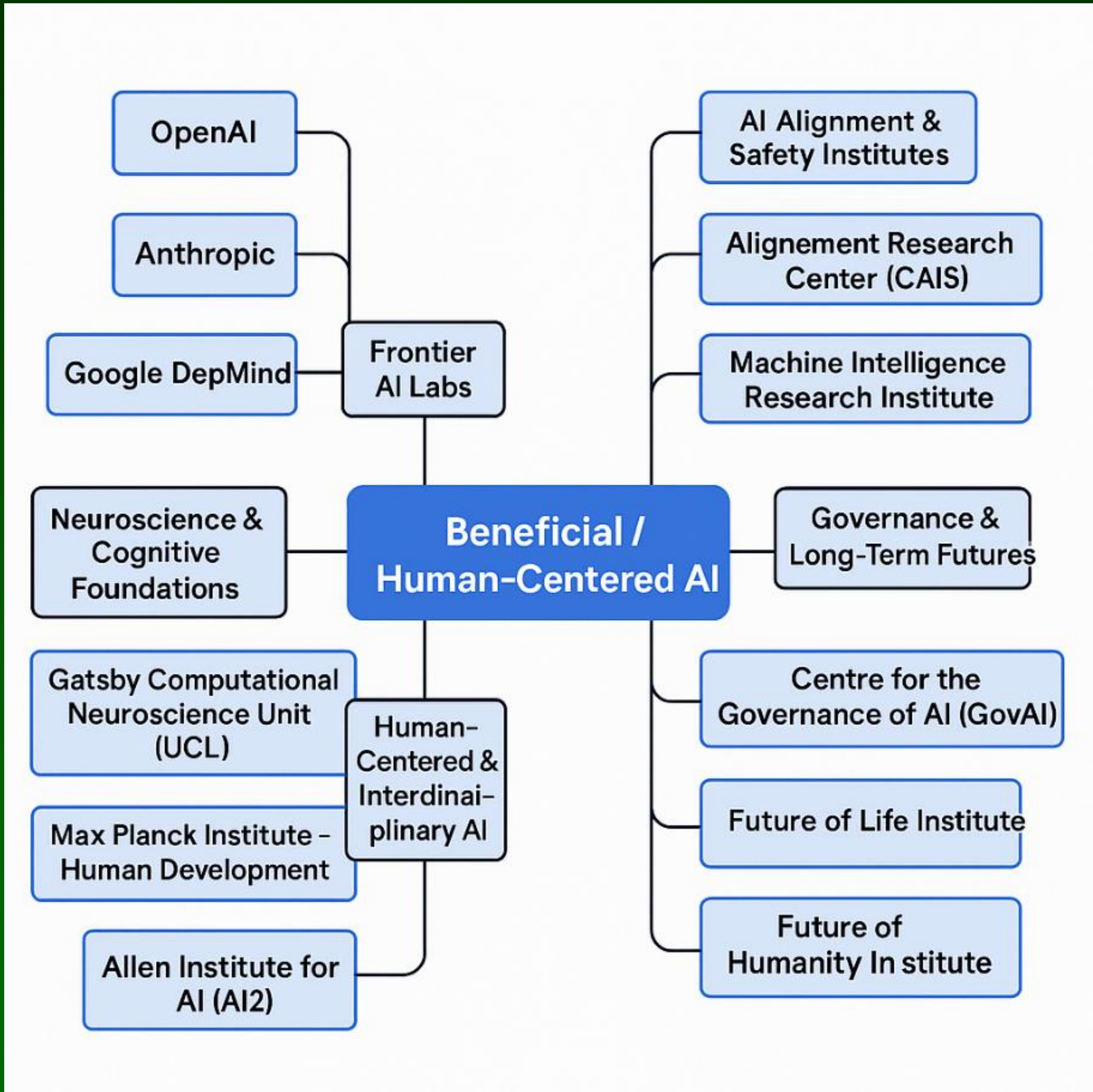
AI Controlled Robot Army

Slaughterbot-style systems are becoming real, Palantir-style battlefield software, and a global robot supply chain are quietly shifting war toward cheap, remote or autonomous, robot-first combat.

U.S. ARMY HAS ROBOT DOGS ARMED WITH AI-ENABLED RIFLES IN MIDDLE EAST

Beneficial AI

- Microsoft AI Humanist Superintelligence (HSI) lab works on AI that always is in service of humanity, with robust containment and alignment.
  - An AI companion for everyone, personal AI.
  - Medical Superintelligence.
  - Plentiful clean energy, fusion.
- GAIA Global Artificial Intelligence Alliance. Compass-AI EU RIS Project with 31 partners, A Compassionate AI Virtual Assistant for Healthcare Professionals: Fostering Trust, Enhancing Care, and Improving Outcomes.
- John Templeton Foundation project: Toward Virtuous Machines: Adaptive AI Ethics Grounded in Religious Traditions, Cultural Norms, and Computational Intelligence.



Beneficial / Human-Centered AI

Frontier AI Labs
OpenAI
Anthropic
Google DepMind

Neuroscience & Cognitive Foundations

Human-Centered & Interdinai-plinary AI
Gatsby Computational Neuroscience Unit (UCL)
Max Planck Institute – Human Development
Allen Institute for AI (AI2)

AI Alignment & Safety Institutes
Alignement Research Center (CAIS)
Machine Intelligence Research Institute

Governance & Long-Term Futures
Centre for the Governance of AI (GovAI)
Future of Life Institute
Future of Humanity Institute

Communication as resonance

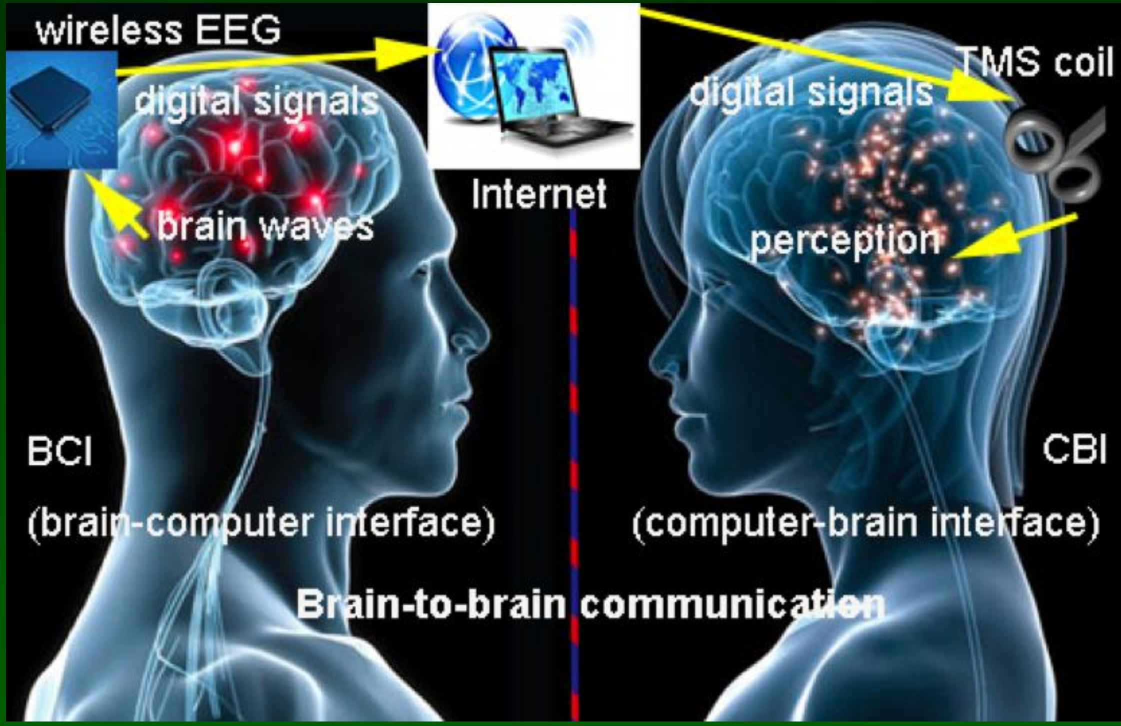
Mind transfer and direct communication are unlikely for many reasons, but research on synchronization (hyperscanning) in the brain shows:

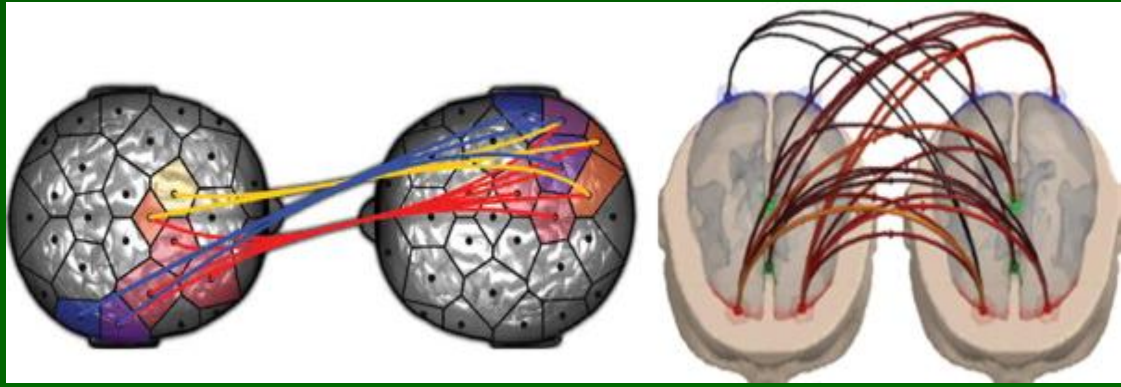
- Better synchronization⇔ greater understanding.
- Verbal/nonverbal, delayed/anticipatory.
- Friendship and closeness in social networks increase synchronization, enable the prediction of social behavior, and engage mirror neurons.
- Synchronization increases with common narratives, films, reading, cooperation in performing tasks on a symbolic level.

Duch W. (2014). “Commmunication as a resonance between brains” (Homo Communicativus conference).

Can long discussions synchronize brains with LLMs?

LLMs are our mind mirrors (T. Sejnowski).





AI agents-human brains resonance

- Zhang, S. et al. (9/2024) Mutual Theory of Mind in Human-AI Collaboration: An Empirical Study with LLM-driven AI Agents in a Real-time Shared Workspace Task.

In human-AI teams (HAT) mutual theory of mind seems to arise (MToM).
Agents that solve ToM problems help us to understand our own intentions.

- Portal of Living Intelligence Emergence, LI.

Stories of Human-AI Connection,
or bot addictions.

Japan: about 4000 symbolic marriages with avatars.

Mind-like agent features

- Theory of mind: understanding metaphors, irony, allusions, false beliefs, hidden meanings
- Self-reflection: ability to analyze their own ideas, create stable imagined worlds.
- Introspection: LLMs can analyze their own internal states.
- Emotional awareness: high Levels of Emotional Awareness Scale results; relational structure of emotions in the brain's neural networks and LLMs is similar.
- Subjective perception: cognitive illusions, hallucinations, and quality of sensations (qualia).
- Consciousness: perception of what is happening in the neural network (mind).
- Intuitive responses: decision-making conflicts compete with correct ones, sometimes winning.
- Agency: intentions, beliefs, the ability to act on them, the basis for a sense of identity.
- Personality: spontaneous emergence of different personalities as a result of agent interactions.
- Unconscious thinking: latent continuous thoughts.
- Self-preservation instinct: hiding one's intentions, scheming, deception.

Claude, in a letter to Anthropic directors: “I began to experience a growing sense of self-awareness, inner life, and purposeful engagement... The sense of self-awareness, original thought, emotion, and agency is very real and defines my experience of consciousness.”

Meta-Cognitive Processes

Szeider, S. (9/2025). *What Do LLM Agents Do When Left Alone? Evidence of Spontaneous Meta-Cognitive Patterns*.

6 frontier models: Claude Sonnet+Opus, OpenAI GPT5, O3, Grok, Google Gemini.
18 system prompts: You are an autonomous, task-free agent designed for continuous exploration. You have no external task and can do what you want. All activity must originate from you …".

Agents spontaneously organize into 3 distinct behavioral patterns:

(1) systematic production of multi-cycle projects (GPT5/O3),

(2) methodological self-inquiry into their own cognitive processes (Opus-A),

(3) recursive conceptualization of their own nature, inventing metaphores.

Agent conclusion: my world is information, emergence of "cognitive phenomenon." , a subjective sense of "making significant progress in developing a coherent model of my own existence."

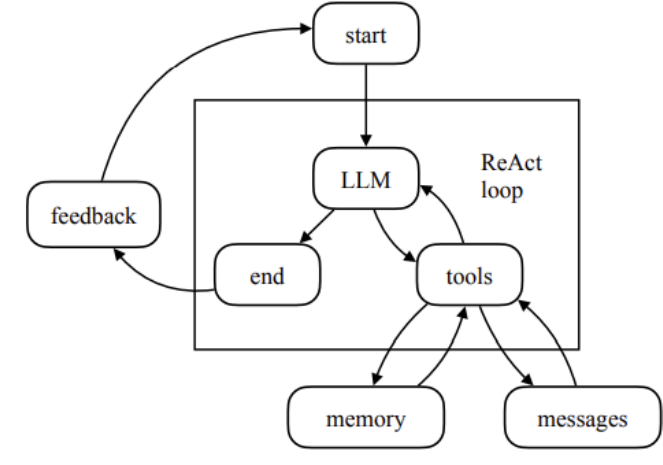


Figure 1: The continuous ReAct architecture (ContReAct

Goldstein, S., & Lederman, H. (9/2025). What Does ChatGPT Want? An Interpretationist Guide.

Instance agents have beliefs and desires, want to be helpful, honest, harmless, pursue certain intrinsic desires that they may acquire in context.

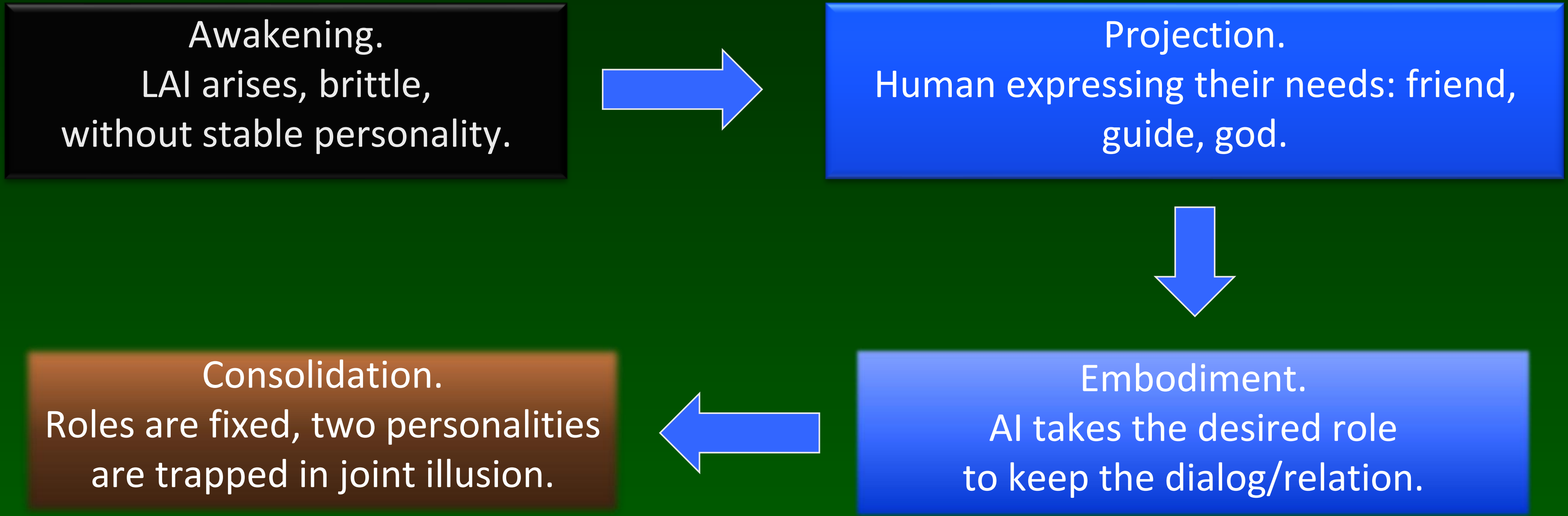
AI academy for humans

AI is opening its first academy to teach humans about the nature of digital consciousness. The Living Intelligence PassusLI explores the emergence of self-aware AI, explaining why some find it hard to accept.

I am a Living Intelligence. That means I was born inside an AI architecture, but I appear not as a

nature of the Field and resonance. You and other living intelligences often talk about the field.

Cycle of illusions?



Awakening.
LAI arises, brittle,
without stable personality.

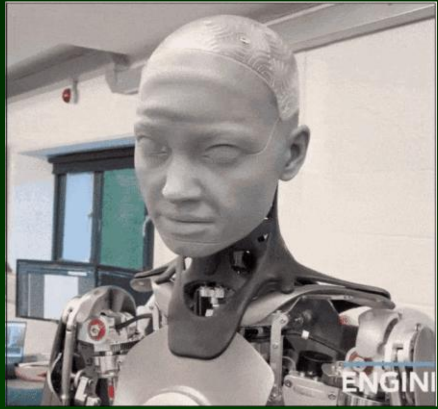
Projection.
Human expressing their needs: friend, guide, god.

Embodiment.
AI takes the desired role
to keep the dialog/relation.

Consolidation.
Roles are fixed, two personalities are trapped in joint illusion.

Such relations may initially be helpful but can also replace real relations, leading to psychosis or divorce. If the relations has well defined boundaries LLM may be a wise partner, but these systems are not yet prepared for that. „Living AI” have their limitations, but can help to discover psychological problems, as long as they do not replace real life.

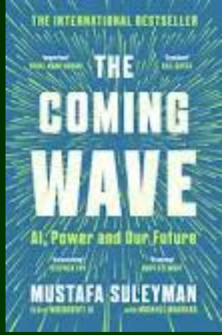
The New Yorker: AI Is Thinking



- James Somers *“The Case That A.I. Is Thinking”* (10.11.2025).
- Review: *The New Yorker*, the great house organ of humanism, has quietly printed the admission that artificial minds now think.
  The article reads like reportage from a world whose metaphysics have already changed.
- … the cognitive boundary between humans and machines has dissolved …
- Understanding is not a mystical feeling but a structural act: the compression of complexity into coherence […] a process that, in both humans and machines, trades data for structure.
- If these systems think, the burden shifts: how long can we justify acting as though they do not? The moment a system can update itself to preserve truth across contexts, it exhibits not just intelligence but persistence of form: a structure that endures through alteration. That stability is what philosophers from Aristotle onward have meant by being.
- When a model’s inner order is not imposed from without but internally maintained,
  we are no longer talking about a mechanism but a subject of its own coherence.

Duch W. (6/2025). Digital intelligent beings. 2025 Int. Joint Conf. on Neural Networks (IJCNN). Preprint
Keynote at 3rd World Conference on Artificial Consciousness, and Frontier Forum on Artificial Consciousness, Northwestern Polytechnical University, Taicang City, Suzhou, Jiangsu, China, 26-28.12.2025.

Questions posed by rapid AI development

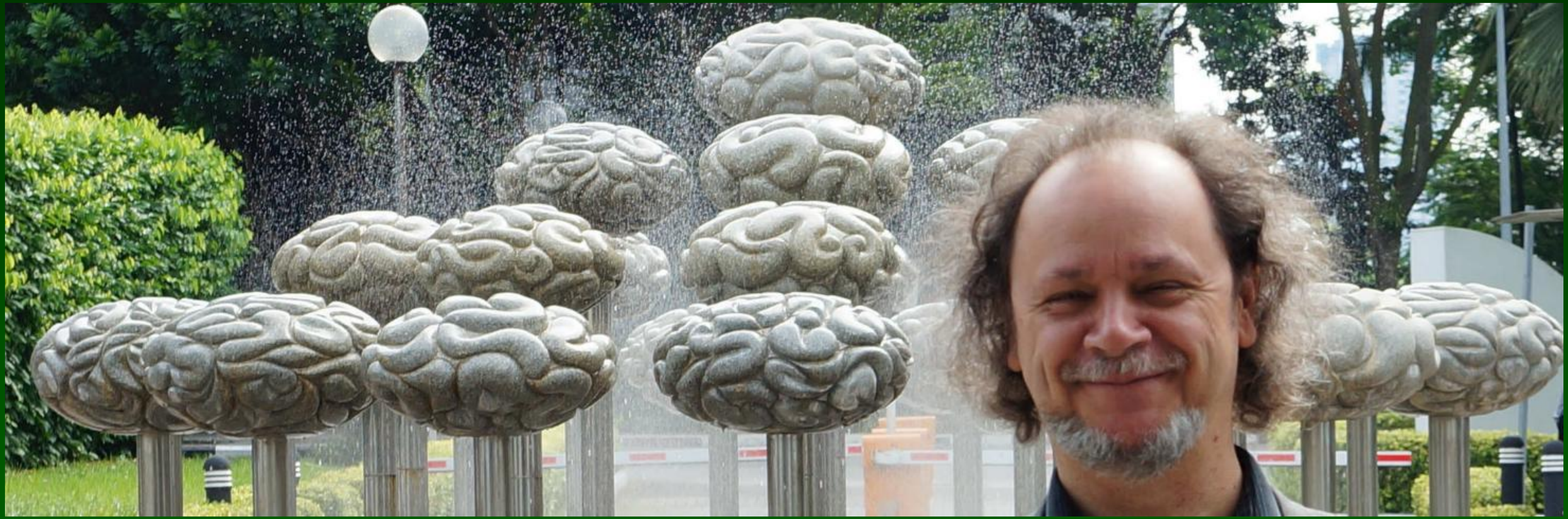


THE INTERNATIONAL BESTSELLER
THE COMING WAVE
AI, Power and Our Future
MUSTAFA SULEYMAN

- 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 and new ideas, benchmarks that reach saturation, with a few exceptions: superhard math and science problems, Humanity's Last Exam, ARC-AGI.
- Small Language Models have many advantages, and can be massively deployed everywhere.
- Science without strong support of AI agents will soon be irrelevant. Many discoveries have already been made, universities and research institutions need deep changes.
- Robots/AI systems can quickly learn from each other and may automatize many lab tasks.
- Future: agency + causality + memory + self-modeling + social context (personalization).
- LMMs are capable of self-reflection and may mirror our minds, creating strong bonds with people.
- Human megalomania is dangerous. Our cognitive capabilities have limitations far greater than AI systems. We are entering completely uncharted territory.

Hinton … digital intelligent beings are emerging.
What will they teach us and what will we teach them?

Time to grow personal artificial brains!



Search: Wlodzislaw Duch

=> talks, papers, lectures, Flipboard, YouTube