Nobilitation of neural networks

Włodzisław Duch

Neuro-informatics and Artificial Intelligence Group
Center for Modern Interdisciplinary Technologies,
Dept. of Informatics, Faculty of Physics, Astronomy & Informatics,
Nicolaus Copernicus University, Toruń, Poland

Google: Wlodzislaw Duch

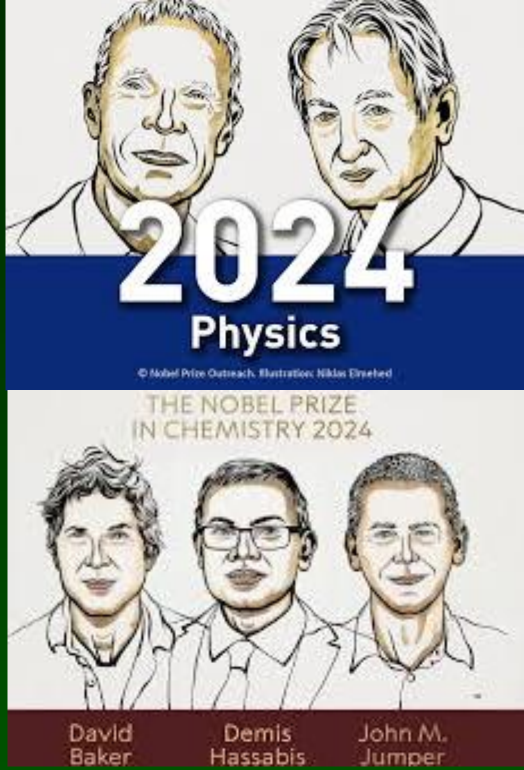
Institute of Physics, Jagiellonian University, Kraków, 7.11.2024

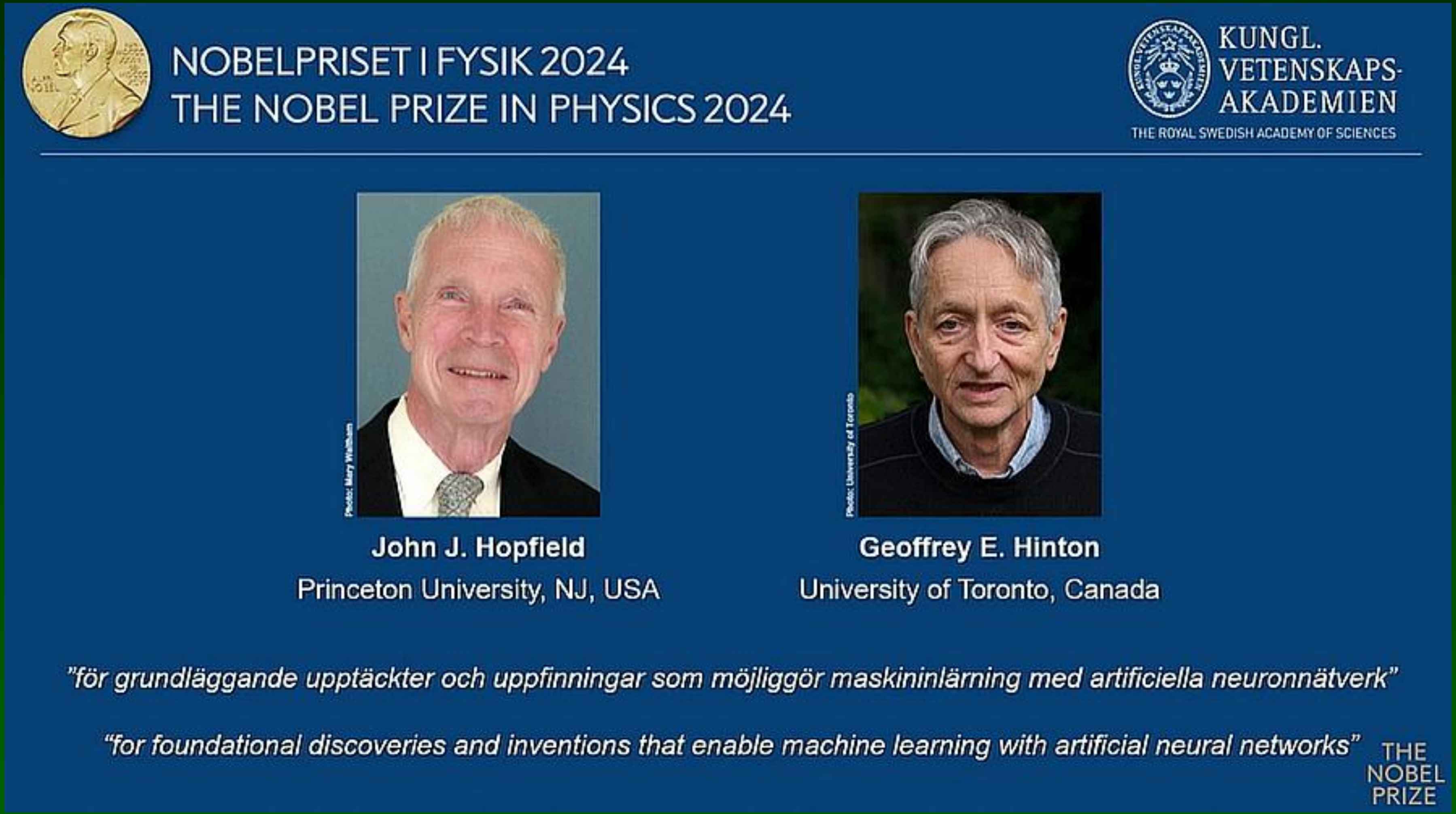
The way to Nobel prizes

1. Roots: Ising model, spin glasses, Hopfield associative memory network.
2. Boltzmann creative machines, deep neural networks.
3. Generative Pretrained Transformers
4. Neural Networks for protein folding.
5. Large Multimodal Models for science.
6. Physics Informed Neural Networks.
7. LMM agents – new intelligent digital beings?

ChatGPT << AI. Over 1000 papers in arxiv.cs.ai every week!
We see only a tip of the iceberg …

AI news on my Flipboard.





Hopfield network model of associative memory and optimization problem.
Hinton discovery of backpropagation (1986) and deep network (2015) training algorithms + other models.

Nobel Prizes related to theory of complex systems: Onsager (1968), Anderson (1977), Parisi (2021).

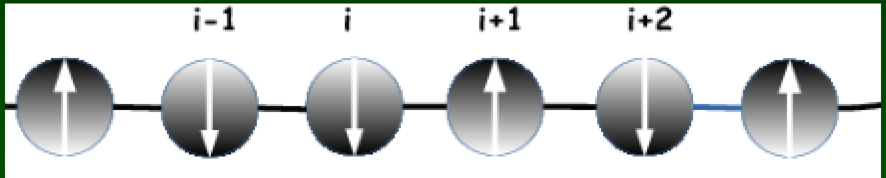
From Ising model …

100 years of Ising model: 800 new papers are based on the Ising model every year!

Wilhelm Lenz (1920) gave a problem of onset of ferromagnetism to his student Ernst Ising.

A model of ferromagnetism in statistical mechanics, with discrete variables that represent magnetic dipole moments of atomic spins $\sigma_i = \pm 1$. Summing over nearest neighbors:

$$H(\sigma) = -J \sum_i \sigma_i \sigma_{i+1}$$



The 1D ferromagnetic model ($J_{ij} > 0$) was solved in 1924 Ising thesis; it has no phase transition.

T. Ising, R. Folk, R. Kenna, B. Berche, and Y. Holovatch, *The Fate of Ernst Ising and the Fate of His Model*, arXiv:1706.01764.

Pauli, Peierls, Kramers, Wannier, Potts, worked on 2D models. The 2D square-lattice Ising model was solved analytically by Lars Onsager in 1944 in the absence of external magnetic field.

L. Onsager, Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition, Phys. Rev. **65**, 117 (1944).

Onsager got 1968 Nobel in Chemistry for this and the older work on the reciprocal relations in thermodynamics of irreversible processes.

to spin glasses …

Philip Warren Anderson was awarded the 1977 Nobel Prize in Physics (with Nevill Mott and John van Vleck) for investigations into the electronic structure of magnetic and disordered systems, including his work on spin glasses.

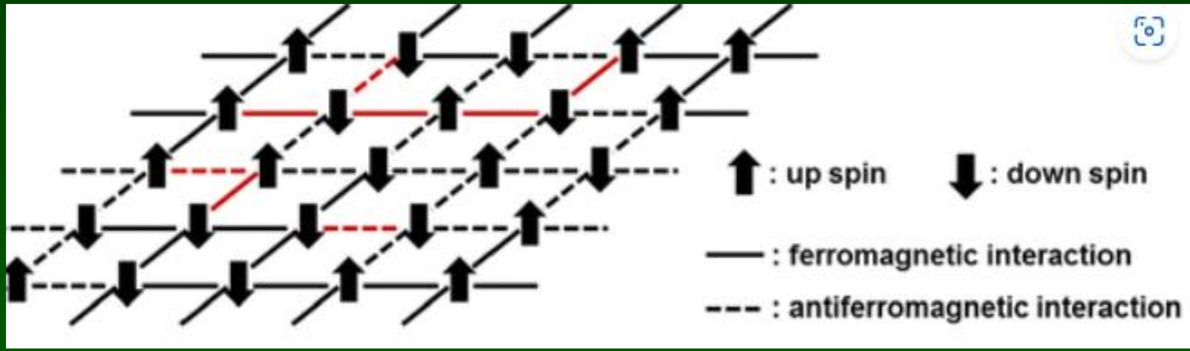
S. F. Edwards, P. W. Anderson, Theory of spin glasses, J. Phys. F: Met. Phys. **5**, 965 (1975)

7 papers in Physics Today on spin glasses, 1988-90.

Spin glasses random Heisenberg Hamiltonian:

$$H(\sigma) = -\sum_{ij} J_{ij}\,\sigma_i\,\sigma_j$$

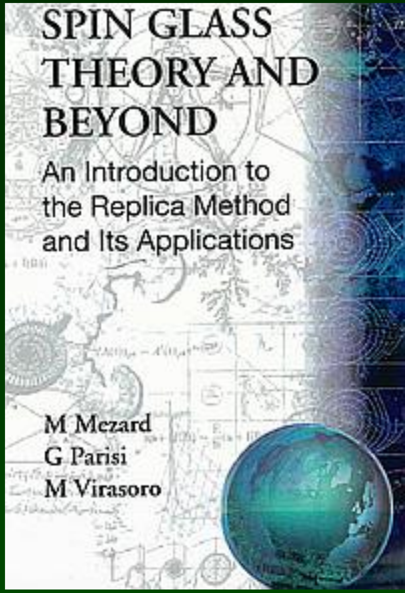
$$J_{ij} > 0 \text{ or } J_{ij} < 0)$$



Spins are in a lattice, local structure repeats periodically in all directions, and interact with neighbors. The 2D square-lattice Ising model is the simplest statistical models showing a phase transition. Depending on temperature this system may assume different structural phases, allowing for identification of phase transitions in a simplified physical model. Potts models generalize this.

Such models were applied to sociophysics, evolution of language, voting, econophysics, tumor modelling, seismic-hazard assessment, volatility in financial markets, mood swings, scientific collaboration, and asymmetric hysteresis in political polarization (Npj Complex **1**, 2024).

From spin glasses to Hopfield network



Anderson (1988): investigation of Mn-Cu spin glasses led to new ideas in optimization, neuroscience, statistical mechanics and protein structure.

Replica method (Sam Edwards, 1971) to calculate partition function $Z = \sum_i e^{-E_i/kT}$ for random Hamiltonian has been extended to the replica symmetry breaking (RSB) scheme by Giorgio Parisi, who was awarded the 2021 Nobel Prize in Physics (with Syukuro Manabe, for CO_2 warming effects; and Klaus Hasselmann, for human-induced climate change). RSB trick was needed to obtain physical results due to breakdown of ergodicity.

David Sherrington and Scott Kirkpatrick developed in 1975 Ising model with long range frustrated ferro- and antiferro-magnetic couplings that corresponds to a mean-field approximation of spin glasses describing the slow dynamics of the magnetization and the complex non-ergodic equilibrium state. Parisi gave exact solution to this model in 1979 and later it was shown that the complex nature of a glassy low temperature was phase characterized by ergodicity breaking, ultrametricity and non-selfaverageness. Energy function of the SK model has many local minima.

This was essentially the Hopfield network with random initialization. How to control patterns corresponding to local minima? This could serve as memory in dynamical systems.

From Hopfield network …

1982 paper, Hopfield presented spin network (binary activation functions) adjusting couplings to create local minima memory patterns. Information is encoded in network energy landscape.
In 1984 he extended this to continuous activation functions.

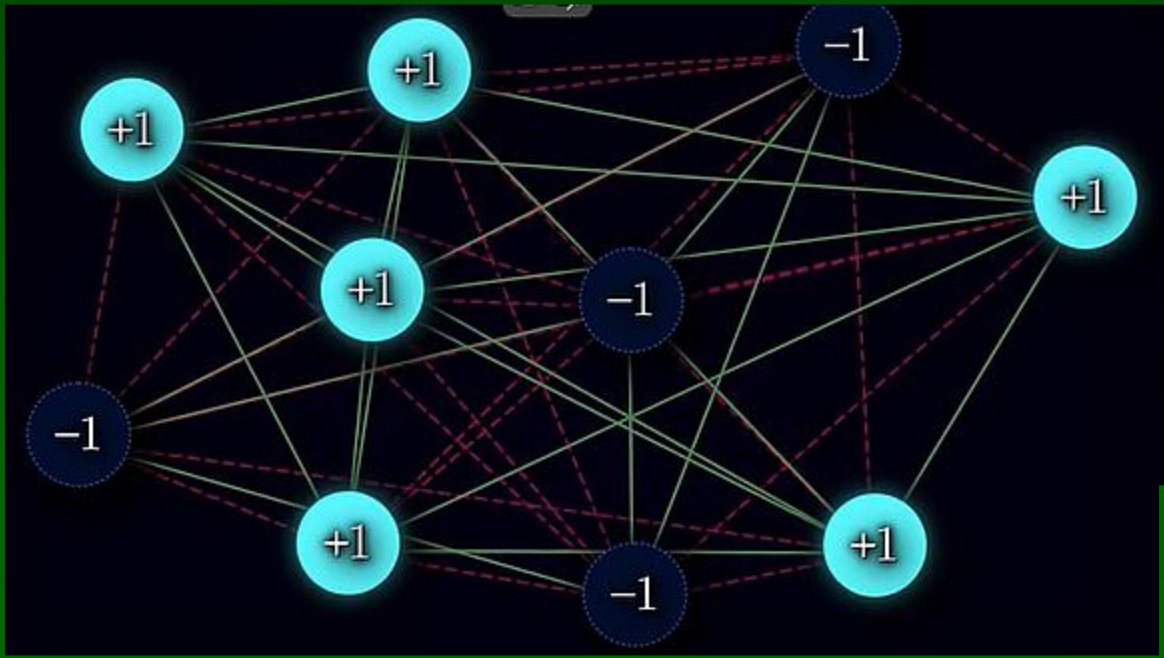
It became a standard model for the study of neural networks through statistical mechanics.

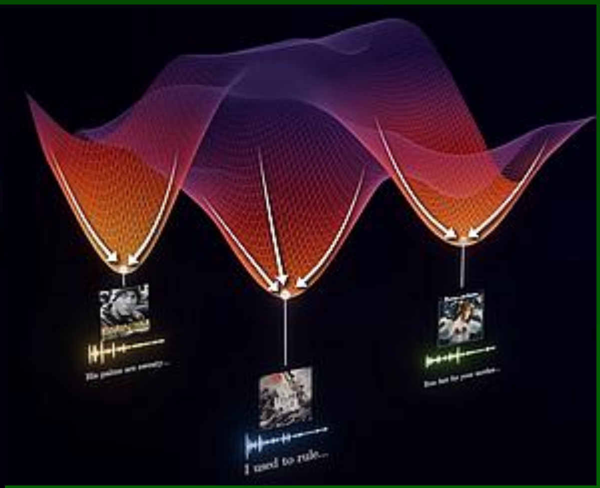
Hopfield network:
fully connected,
symmetric connections

$w_{ij} = w_{ji}$

$\min E = -\sum_{ij} w_{ij} x_i x_j$

Learning: adjust w_{ij}
to create local minima.
Inference: find x_i patterns
minimizing energy.



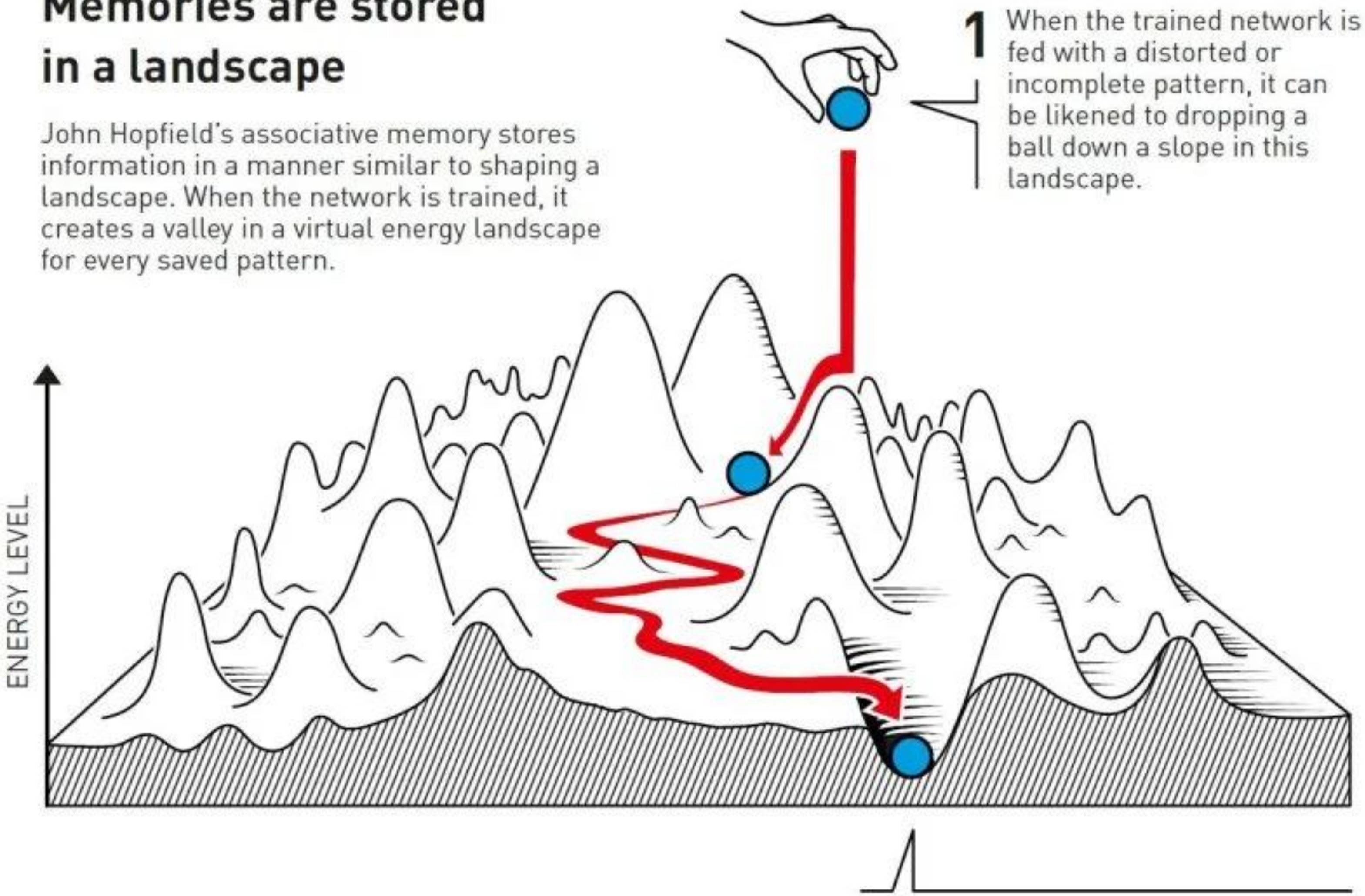


[Artem Kirsanov](#)

Memories are stored in a landscape

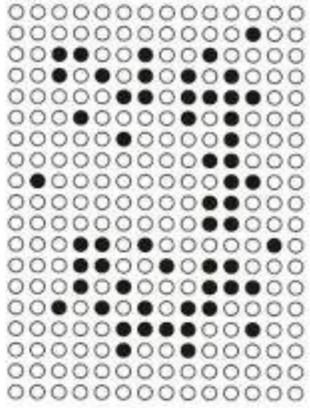
John Hopfield's associative memory stores information in a manner similar to shaping a landscape. When the network is trained, it creates a valley in a virtual energy landscape for every saved pattern.

1 When the trained network is fed with a distorted or incomplete pattern, it can be likened to dropping a ball down a slope in this landscape.

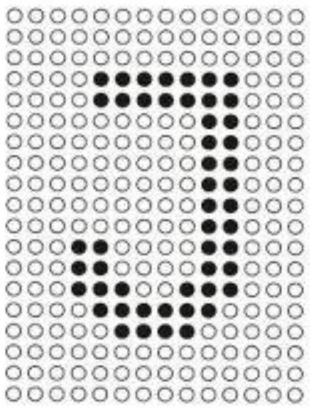


2 The ball rolls until it reaches a place where it is surrounded by uphills. In the same way, the network makes its way towards lower energy and finds the closest saved pattern.

INPUT PATTERN



SAVED PATTERN



Hopfield network as associative memory

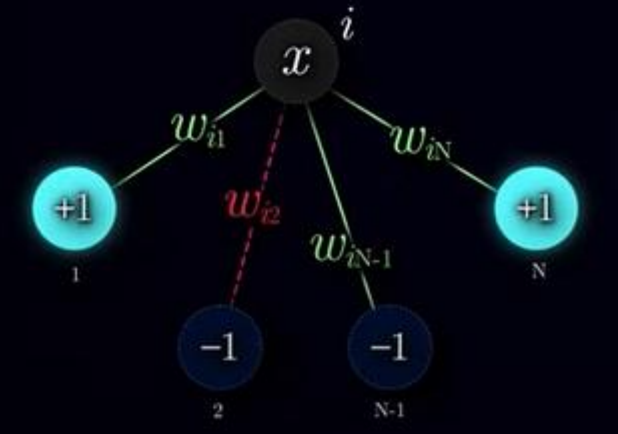
To reach minimum update x_i values summing all inputs reaching this node.
This will decrease energy, iterations will lead to a pattern $\{x_i\}$ corresponding to a local minimum, depending on initial pattern.

$$h_i = \sum_{j \neq i} w_{ij}\, x_j$$

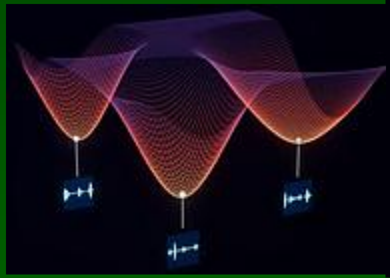
$$x_i = \begin{cases} +1 & \text{if } h_i > 0 \\ -1 & \text{if } h_i < 0 \end{cases}$$

This leads to pattern completion, noise cleaning and optimization.

Learning: for a single global minimum set $w_{ij} = x_i x_j$, so for N neurons:



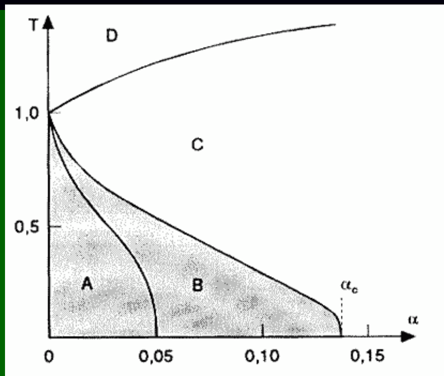
min $E = -\sum_{ij}(x_i x_j)^2 = -\mathrm{N}^2$



Storing several patterns: sum all weights for each patter

$$w_{ij}(k) = x_i(k)x_j(k)$$

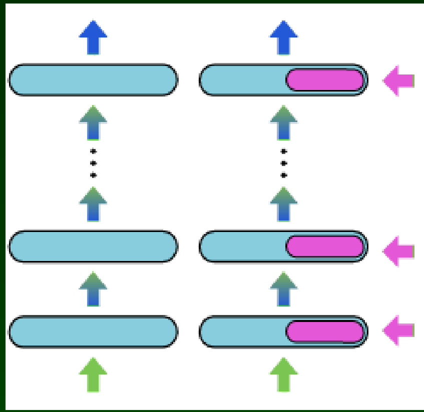
Memory capacity is ≈0.14 N   Many tricks to increase it.
Hebb rule: neurons that fire together wire together, justifies averaging.
Use simulated annealing to explore energy landscape => 4 phases in diagram.
α=$\mathrm{N_p}$/N. A – stable, B – shallow local, C – few broad, D – one global minimum.



[Hopfield Networks is All You Need | Hopfield-layers]

[Good intro by Artem Kirsanov] (YouTube)

Still relevant, after all these years …



- H. Ramsauer + 15 coauthors, *Hopfield Networks Is All You Need*, arXiv:2008.02217.

Hopfield network with continuous states and update rule equivalent to the attention mechanism used in transformers can store exponentially many patterns, retrieves the pattern with one update, and has exponentially small retrieval errors.
It has 3 types of energy minima: (1) global fixed point averaging over all patterns, (2) metastable states averaging over a subset of patterns, (3) fixed points which store a single pattern.

- C. Zanoci, N. Dehghani, and M. Tegmark, *Ensemble Inhibition and Excitation in the Human Cortex: An Ising Model Analysis with Uncertainties*, arXiv:1810.07253.

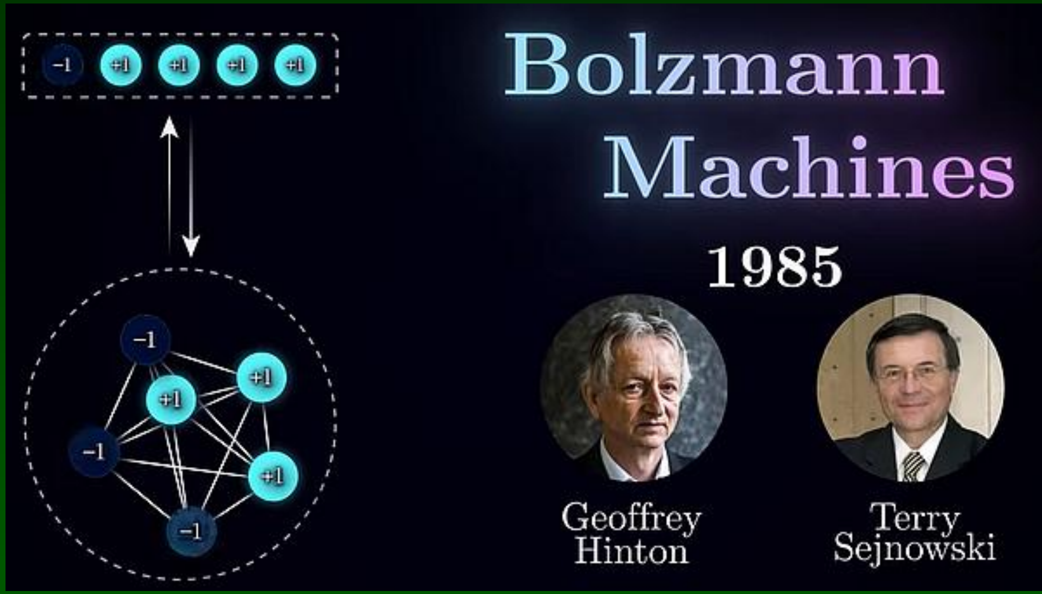
Random walks in parameter space using adaptive Markov Chain Monte Carlo, applied to the spiking patterns of excitatory and inhibitory neurons recorded with multielectrode arrays in the human temporal cortex during the wake-sleep cycle. Information-theoretic measures reveal that the Ising model captures neuronal collective behavior, explaining ~ 80%-95% of the correlations, depending on sleep state and neuron type. Thermodynamic measures show signatures of criticality …

- Hopfield models can be used for optimization: construct Hamiltonians with proper constraints.
- J. Mańdziuk, Sieci neuronowe typu Hopfielda: teoria i przykłady zastosowań. AOW 2000.

… to Boltzmann machines

Hopfield model had a great number of variants and applications, including human dementia.

It is replicating learned patterns from cues.



Adding randomness to computations, find probability distributions, not fixed patterns, to generalize memory patterns and create similar ones.

From perfect reproduction to jazz improvisations.

Hopfield network + hidden units + stochastic exploration.

Prob. of transitioning between two energy states:

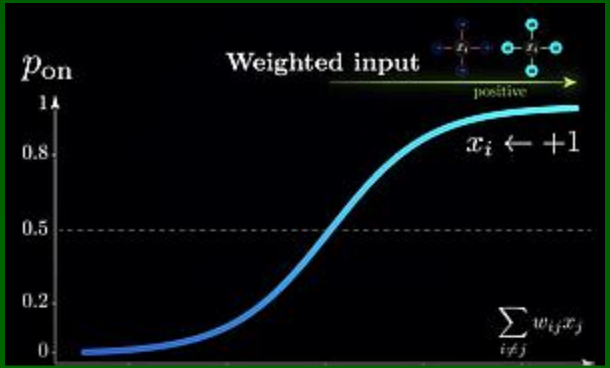
$$P(\Delta E) = e^{-\Delta E/T}$$

$$P(State_E) = \frac{1}{Z} e^{-E/T}$$

Partition function $Z = \sum_S e^{-E_s/T}$

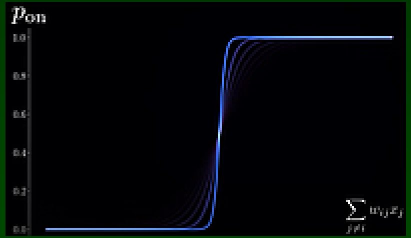
Update state of neuron x_i calculating energy $E_i = -\sum_{i \neq j} w_{ij} x_i x_j$ with probability based on energy difference for $x_i = \pm 1$, given by sigmoidal function

$$P(x = +1) = \frac{e^{-E_{i+}/T}}{e^{-E_{i+}/T} + e^{-E_{i-}/T}} = \frac{1}{1 + e^{-\Delta E_i/T}}$$



… from Boltzmann machines …

Calculate input to neuron x_i from connected neurons $\sum_{i \neq j} w_{ij} x_j$
Calculate update probability using sigmoidal function $P(x_i = +1) = \dfrac{1}{1 + e^{-\Delta E_i / T}}$

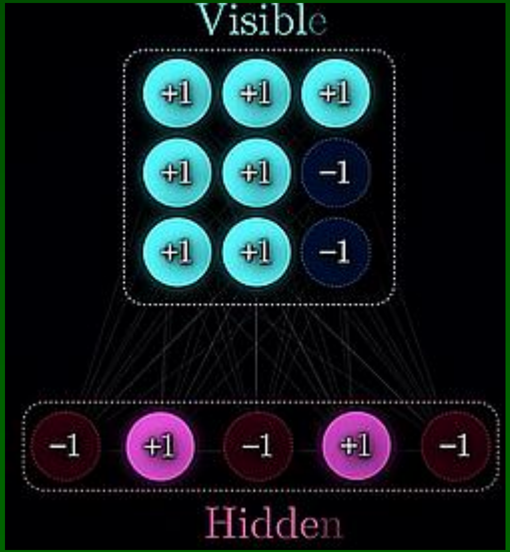


Generate random number: if it is smaller than $P(x_i)$ no change, if larger change.

Low T = more deterministic, high T more random exploration of local minima.
Used also in LLM models to control creativity of network.

Contrastive learning: changing probability of one pattern will influence partition function.

Maximize joint probability of training data $\{x^{(1)}, x^{(2)}, \dots\ x^{(n)}\}$



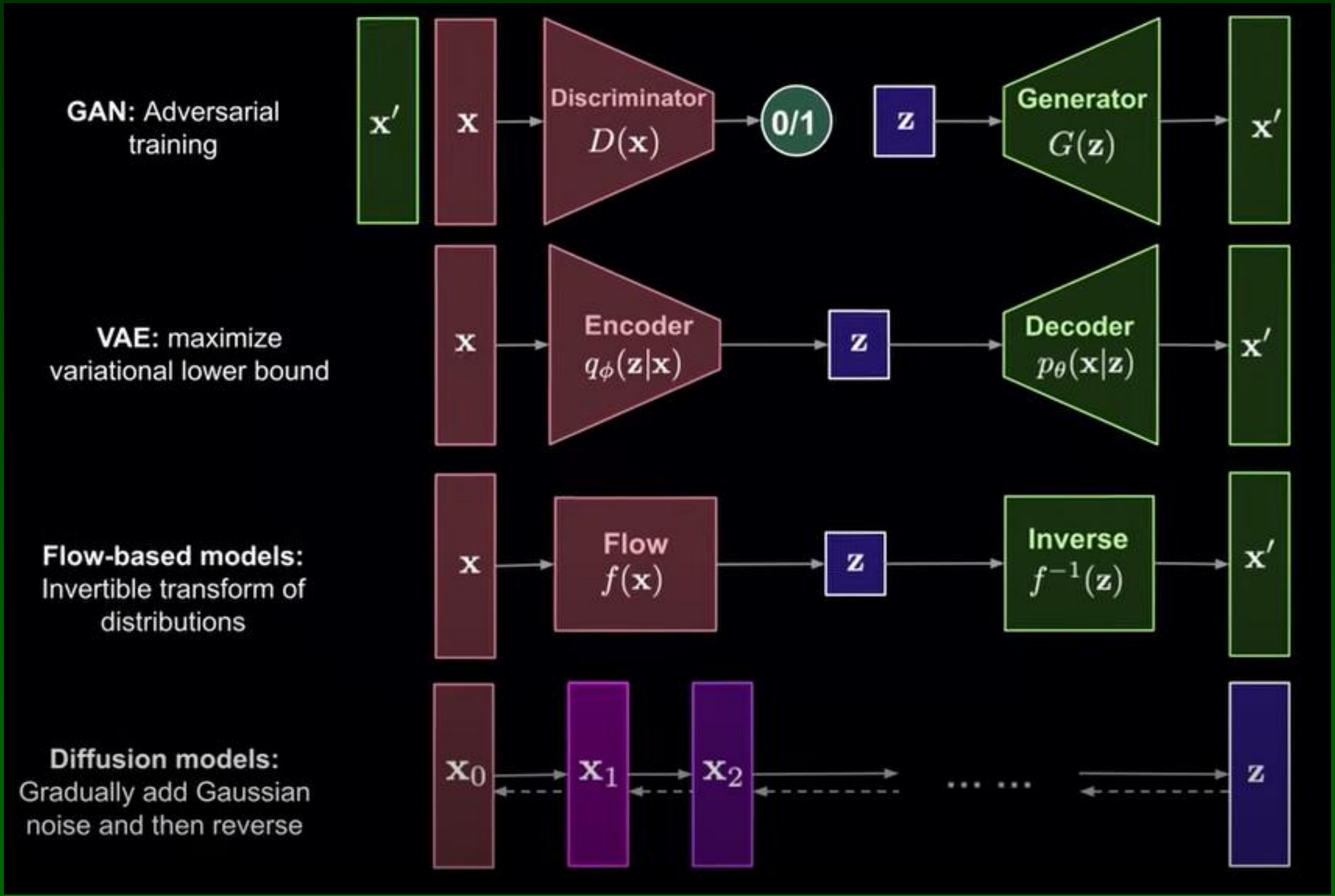
$$\log P(\text{Data}) = -\frac{1}{T} \sum_{i=1..n} \log P\left(x^{(i)}\right) - n \log Z$$

Minimization of partition function will increase energy of spurious states.
Calculation of gradients leads to a contrastive Hebbian rule:

$$\Delta w_{ij} \propto \langle x_i x_j \rangle_{\text{data}} - \langle x_i x_j \rangle_{\text{model}}$$

Clamp inputs, run network, weights reinforce data encoding, second part reduces dreaming, run starting from random inputs.
RBMs, Restricted Boltzman machines, have only input-hidden units connections.

… to other latent probabilistic models …



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

Algorithmic (Kolmogorov) complexity is the key.

Attractor neurodynamics

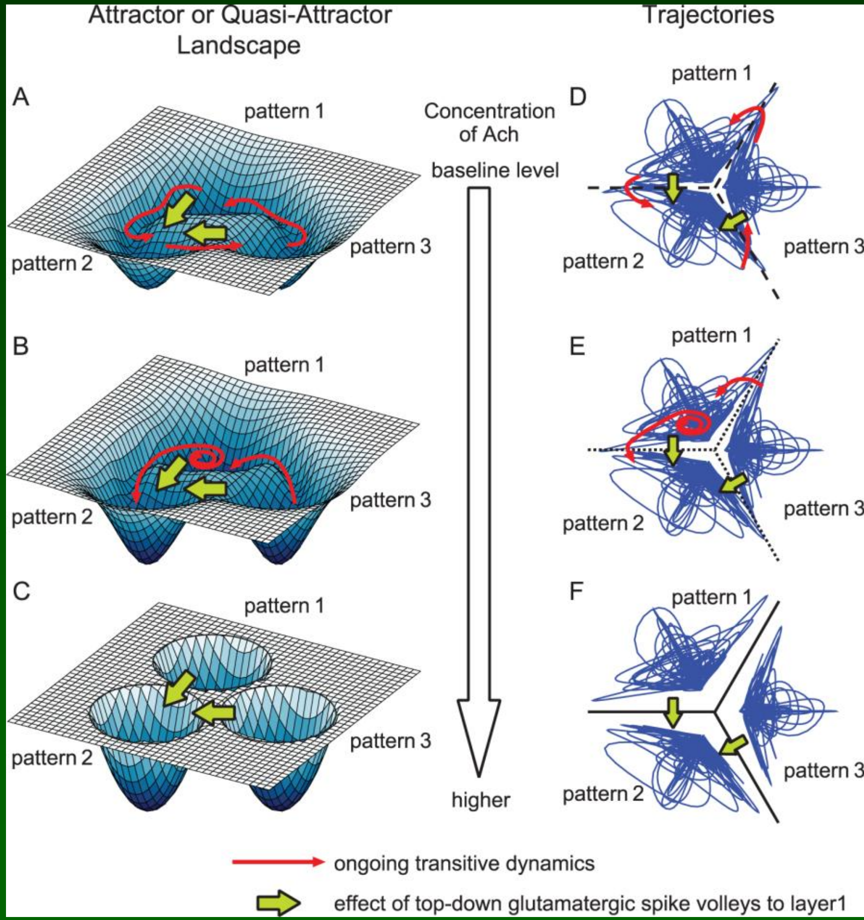
Hopfield networks and Boltzmann machines belong to a larger category of dynamical systems called attractor networks.
In Hopfield network dynamics settles to a point attractor, in Boltzmann machines stochastic dynamics leads to metastable states, blended states, cyclic or stochastic states.

In theoretical neuroscience, different kinds of attractor neural networks have been associated with different functions, such as memory, motor actions, and associations.

Amit, D. J. (1992). Modeling Brain Function: The World of Attractor Neural Networks. CUP.

Thoughts are chains of transitions between attractor states. Attractor networks explain why some information may become memes, “units of cultural information” (Dawkins, 1976), integrated into conceptual framework of memory system.

Duch W. (2021) Memetics and Neural Models of Conspiracy Theories. Patterns. Cell Press.

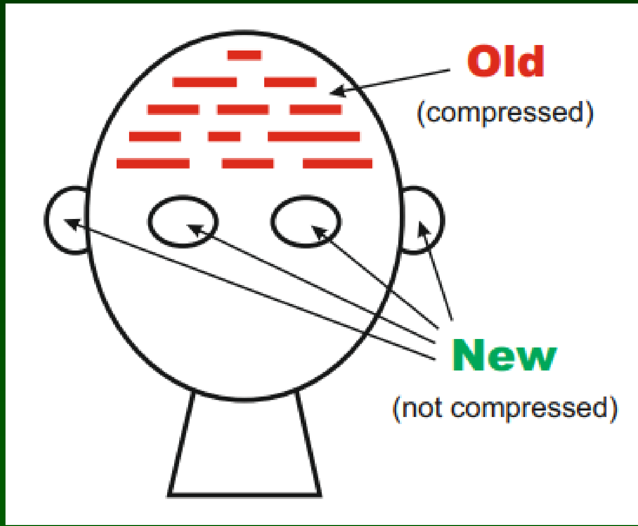


Cognition as Compression

Computing ↔ Cognition, Artificial ↔ Natural systems.

Compression in ML model selection, Algorithmic Information Theory (AIT): Minimum Length Encoding (MLE), Minimum Description Length (MDL), Minimum Run Length Encoding, Minimum Message Length Encoding, etc.

- Cognition, Language Learning is based on compression



Two kinds of memory systems: fast but small, and vast, stable but slow.

Episodic memory allows to remember events after a single exposure.
Hippocampus encodes information about active brain areas, but has small capacity, contains < 100 mln neurons, while cortex has 160 times more.
It can recreate activity of the cortex similar to that of the past experiences.

A flood of images, conversations and texts requires compression to remember.
Consolidation of episodic memory takes part during sleep, like contrastive learning?
Hippocampus re-activates cortex many times, forming compressed semantic memory, allowing for generalizations, associations and abstract reasoning.
Knowledge is internalized, associations become automatic.

LLMs contain only compressed memories, do not record new episodes.

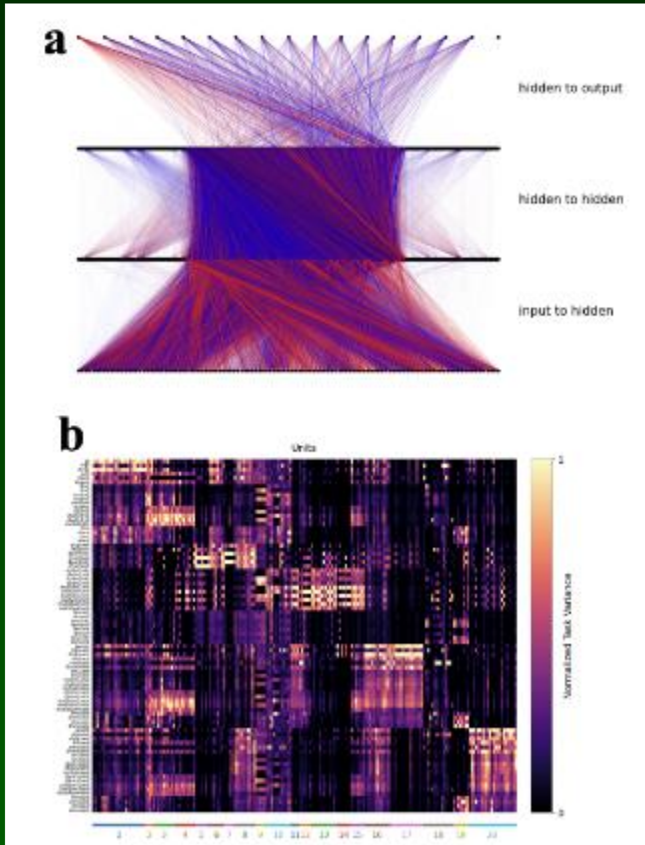
Brain-inspired modular training

Brain large scale network sparsity 0.1-1%, but functional connectivity 1-5%.

Y. Li, E. J. Michaud, D. D. Baek, J. Engels, X. Sun, and M. Tegmark,
The Geometry of Concepts: Sparse Autoencoder Feature Structure (10/2024).

Z. Liu, M. Khona, I.R. Fiete, and M. Tegmark, *Growing Brains: Co-Emergence of Anatomical and Functional Modularity in Recurrent Neural Networks* (2023)

Mod-Cog cognitive multi-task battery, consisting of 84 tasks, such as interval estimation, mental navigation, sequence generation.



Recurrent neural networks (RNNs) exhibit functional modularity, neurons are clustered by participation in shared computational subtasks.
Brain-inspired modular training (BIMT) model trained to solve a set of compositional cognitive tasks grows brain-like anatomical modularity.
Functional and anatomical clustering emerge together, such that functionally similar neurons also become spatially localized and interconnected.
The model can optimally balance task performance and network sparsity, enhancing the interpretability of neural network architectures.

XAI Anthropic

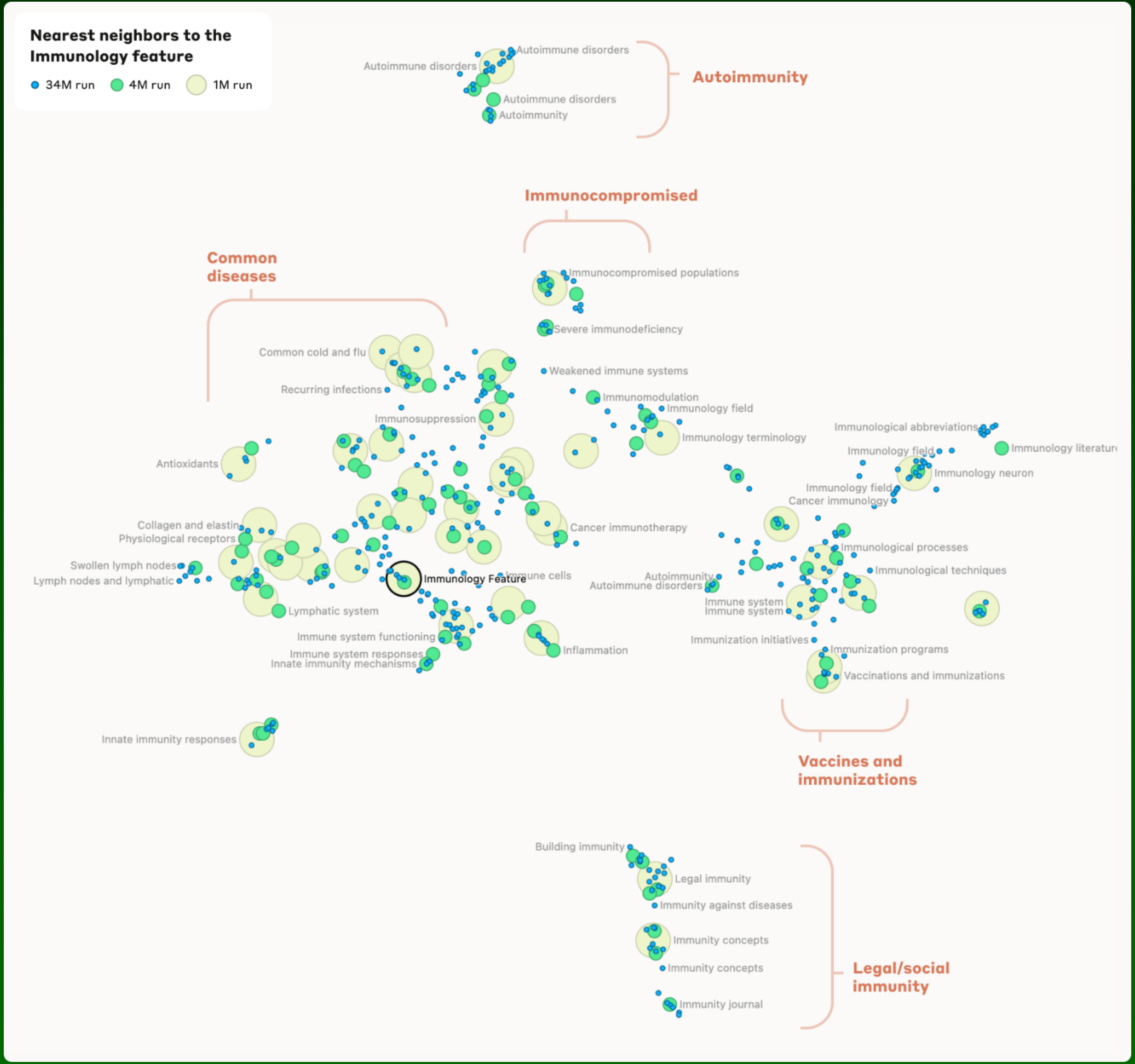
XAI = eXplainable AI. Anthropic Claude 3.

Can we understand how it works?

Immunology: activation pattern found in multiple contexts, patches of concepts that fits to a specific context.

Separate neural network was trained to reconstruct the activations of a specific layer of the LLM, activation patterns are treated as features or concepts that the model has learned.

Duch W, Setiono R, Zurada J.M. (2004) Computational intelligence methods for understanding of data. Proc. of the IEEE 92(5) (2004) 771- 805



Othello-GPT

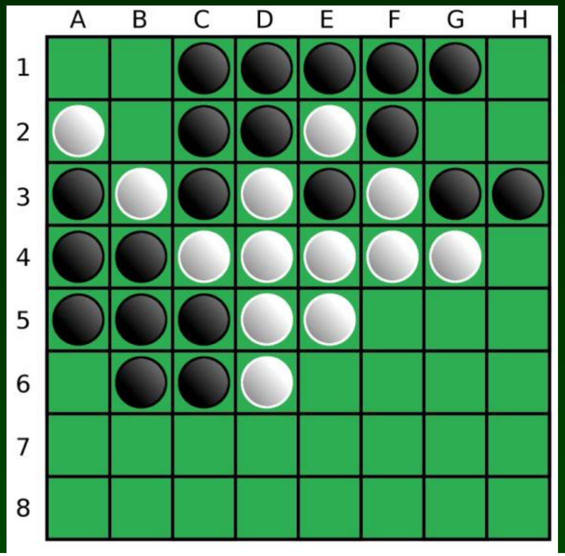
Superficial statistics or coherent and grounded neural representations?

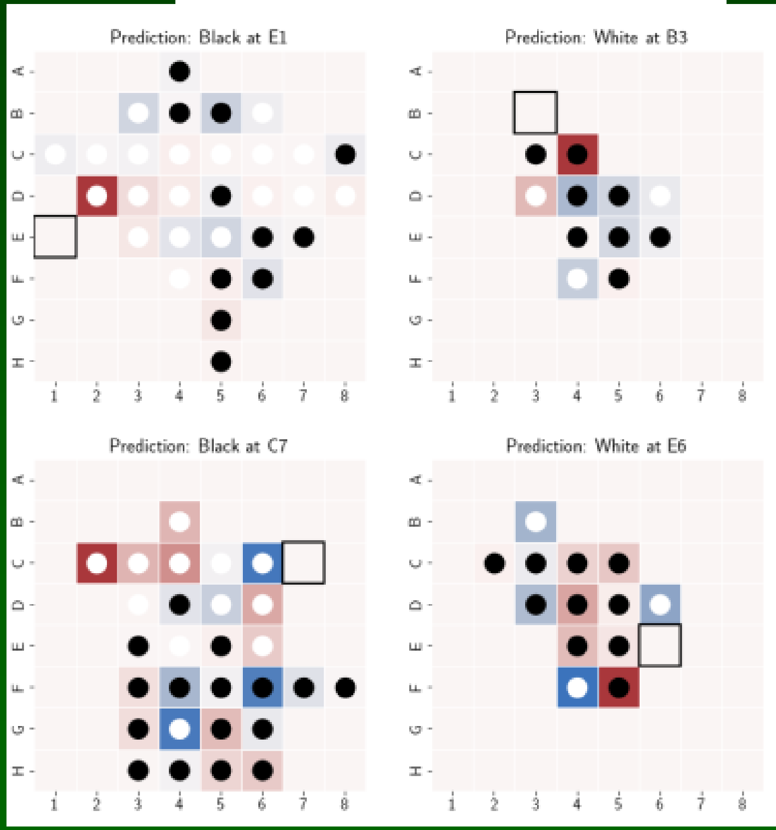
- A GPT variant Othello-GPT was trained to extend a list of moves with legal moves. The model has no a priori knowledge of the game or its rules, it only predicts the next move.
- A 2-layer MLP classifier takes as input internal activations of a network, outputs next position. Activation-intervention technique was used to create latent saliency maps.
- Internal board representation emerged as a result of compression.

The network was playing Othello roughly like a human: by keeping a game board in its “mind’s eye” and using this model Othello-GPT learns to make **strategically good moves**.

It is the most parsimonious description of the training data.

- GPT perceives what appears in its imagery.
- Black empty box: Top-1 prediction by the model. Contribution to prediction: red=high, blue=low





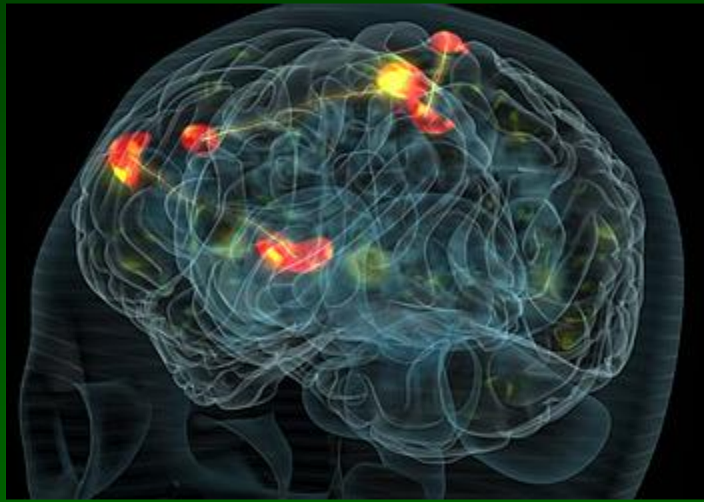
Autoreflection and consciousness

Several papers point to the development of internal models in LLMs.

Learning spatial and temporal information shows emergence of robust linear representations of space and time across multiple scales.

Gurnee, W., & Tegmark, M. (2024). Language Models Represent Space and Time.

Li, K. … H., & Wattenberg, M. (2023). Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task.



GPT perceives what appears in its imagery.
This is similar to the processes in our brains, the basis of self-awareness.

Numerous neuronal networks excite and compete with each other. Only the strongest processes that can be clearly distinguished from the noise (signal detection theory), will be expressed as action, speech, thought, or percept.

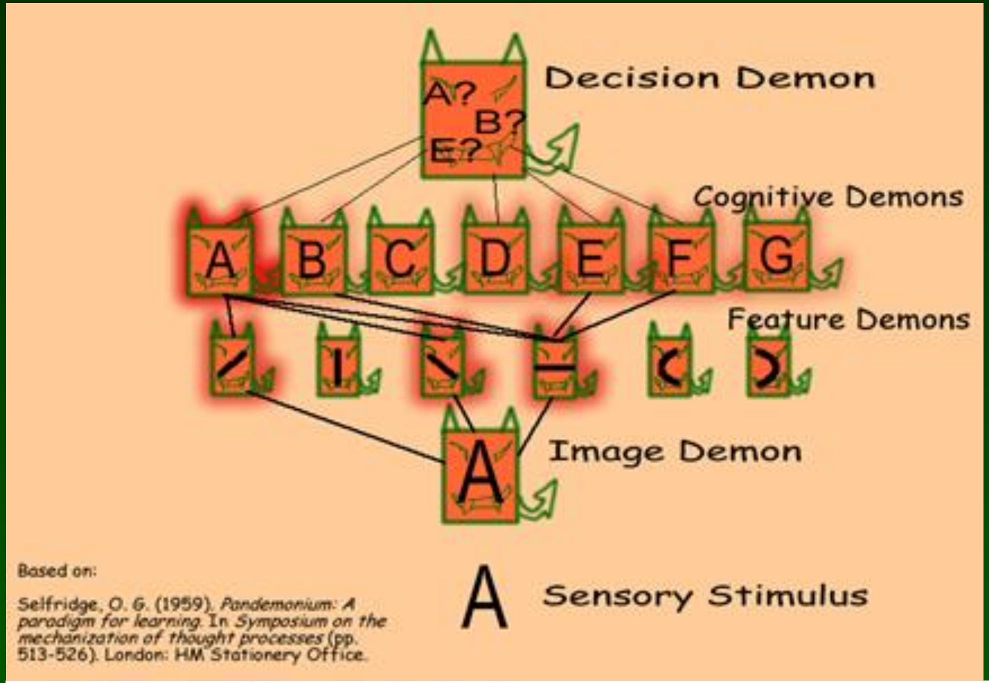
Consciousness is the perception of what is going on in one's own mind (J. Locke, 1689).

Serious discussions on machine consciousness: 6 major theories and requirements.
Butlin et al. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness.*

Consciousness of Artificial Intelligence - MoC5 Public Evening Event

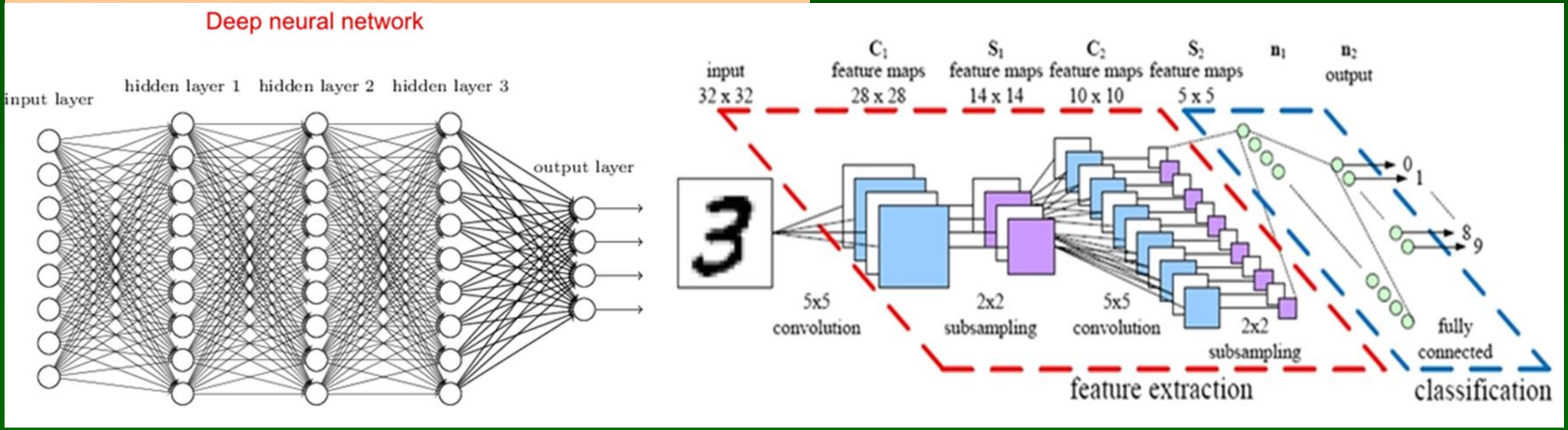
Feedforward neural networks

Neural feedforward classifiers



Pandemonium, invented in 1959 by Selfridge.

Data, words, image patches
=> networks extract features adjusting connections (weights),
=> training to recognize many patterns,
=> invariant object classification, regression.



… to feedforward neural networks …

DE Rumelhart, GE Hinton, RJ Williams, Learning representations by back-propagating errors

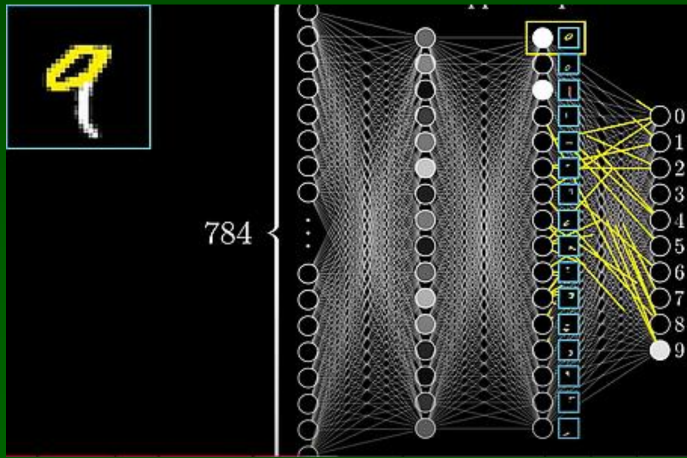
Nature 323, 1986, cited ≈ 40K.

DE Rumelhart, GE Hinton, RJ Williams, Learning internal representations by error-propagation

Parallel Distributed Processing, Book chapter, 1986, cited ≈ 55K.

A Krizhevsky, I Sutskever, GE Hinton, Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25, 2012, cited ≈165 K.

Y LeCun, Y Bengio, G Hinton, Deep learning. Nature 521, 2015, cited ≈86 K.



Multi-layer perceptrons (MLPs) use a combination of sigmoidal functions, can learn to approximate arbitrary function.

$$F_i(X;W) = \sigma\left(\sum_j W_{ij}^{(M)} \sigma\left(\sum_k W_{jk}^{(M-1)} \sigma\left(\sum_l W_{kl}^{(M-2)} \ldots \sigma\left(\sum_n W_{mn}^{(2)} X_n\right)\right)\right)\right)$$

Much better convergence properties than polynomials.
Forward pass, check errors at output, backpropagate errors updating weights using gradient of error function:

$$E(\mathbf{W}) = \frac{1}{2}\sum_{i=1}^{N}\left(Y_i - F_i(X;\mathbf{W})\right)^2 = \frac{1}{2}\sum_{i=1}^{N} E_i$$

Knowledge in neural networks

Hopfield networks remember episodes, patterns.

Semantic memory requires stable, latent representation.
Networks extract features, hidden layers transform inputs enabling classification, prediction of object properties.

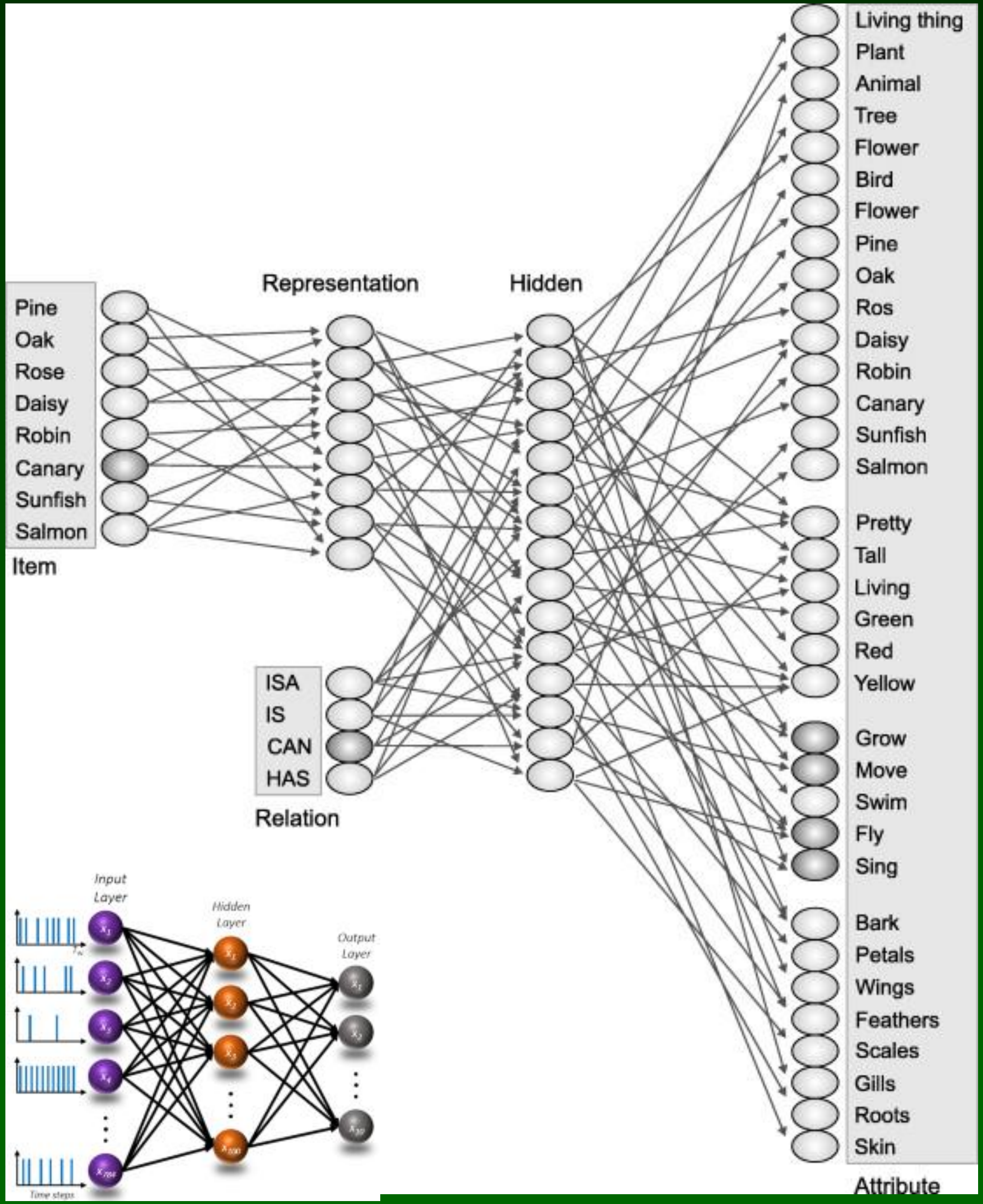
We have no access to the hidden representations, only to the network outputs: actions, decisions, thoughts, imagery.

Brains: 100 T synapses/weights.

Sensory inputs => trains of spikes.
Spikes => recurrent activation of groups of neurons
=> spreading activation through neural networks.

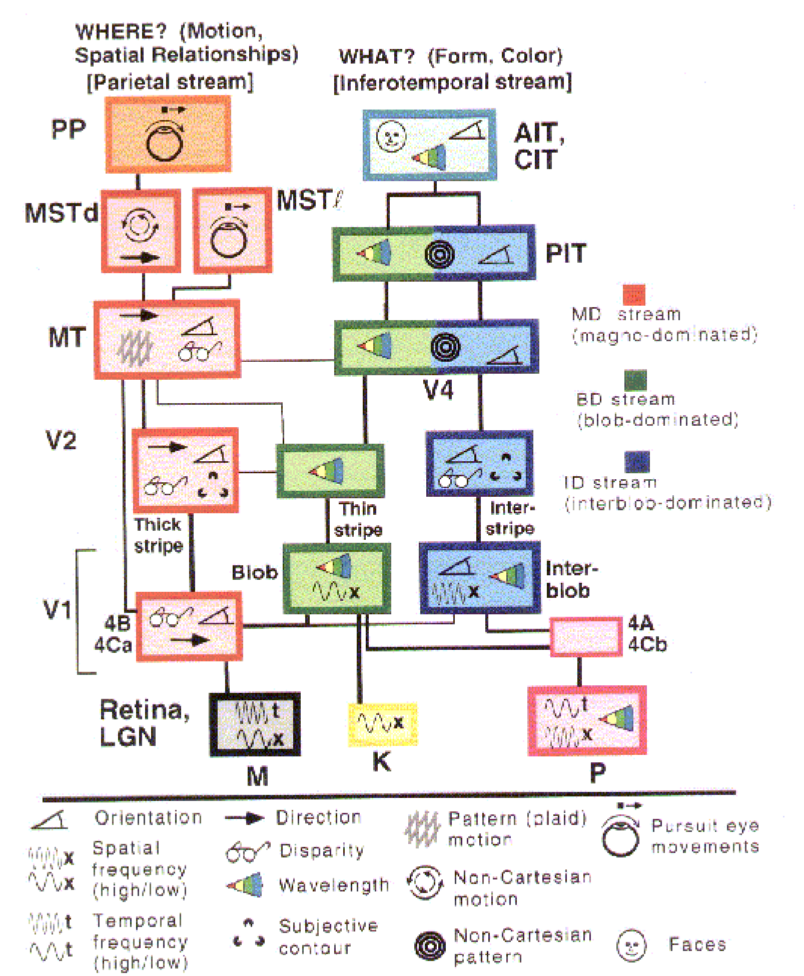
LMM networks: 1 T parameters.

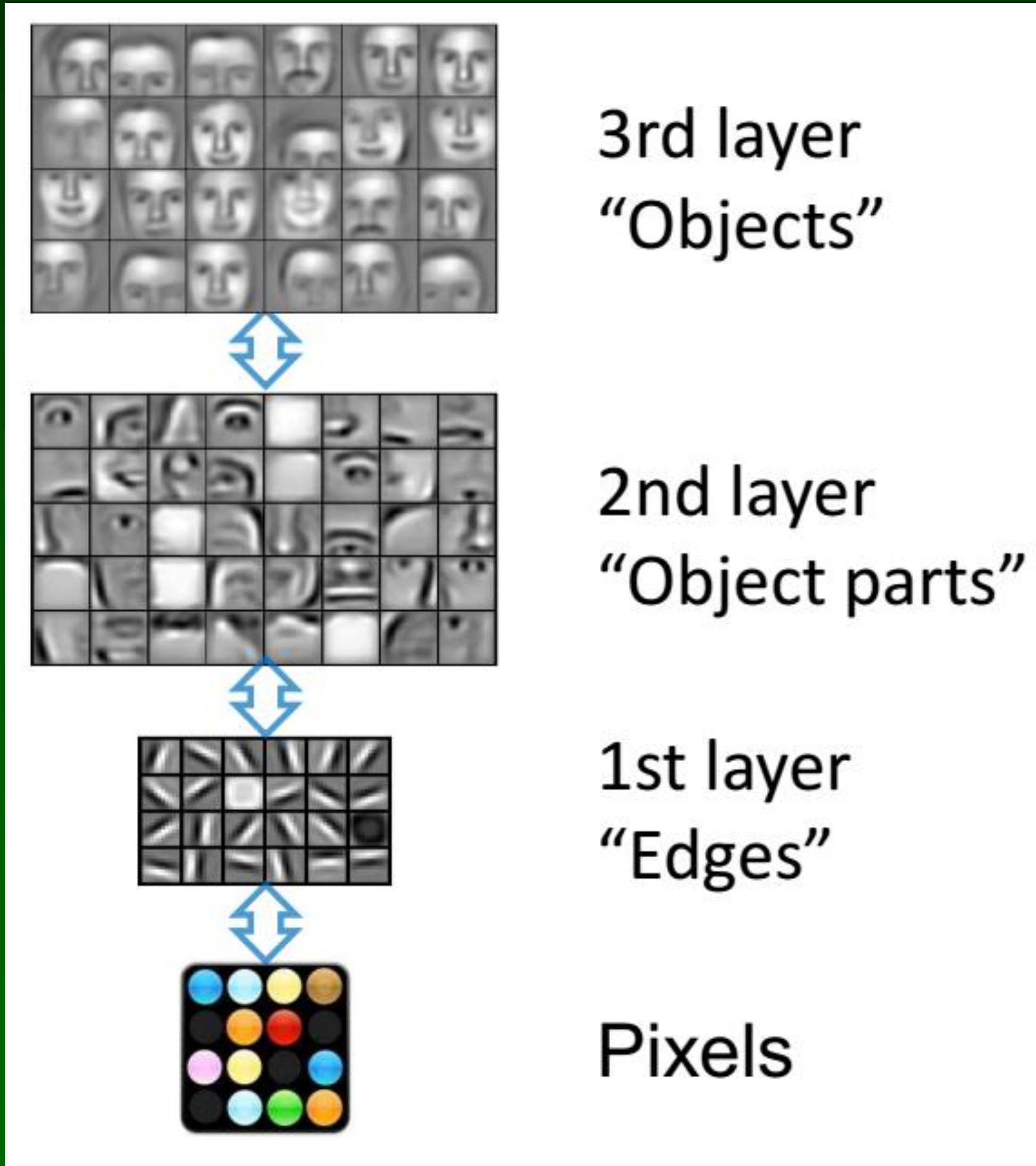
Input: words, image patches, data,
=> networks with adjustable parameters, many layers,
=> parameters trained to recognize patterns,
=> output decisions, responses to external/internal inputs.



… to deep neural networks …

For complex data, like images, networks with many layers are used, discovering simple features and than composing these features in increasingly complex ways.





From deep networks to transformers

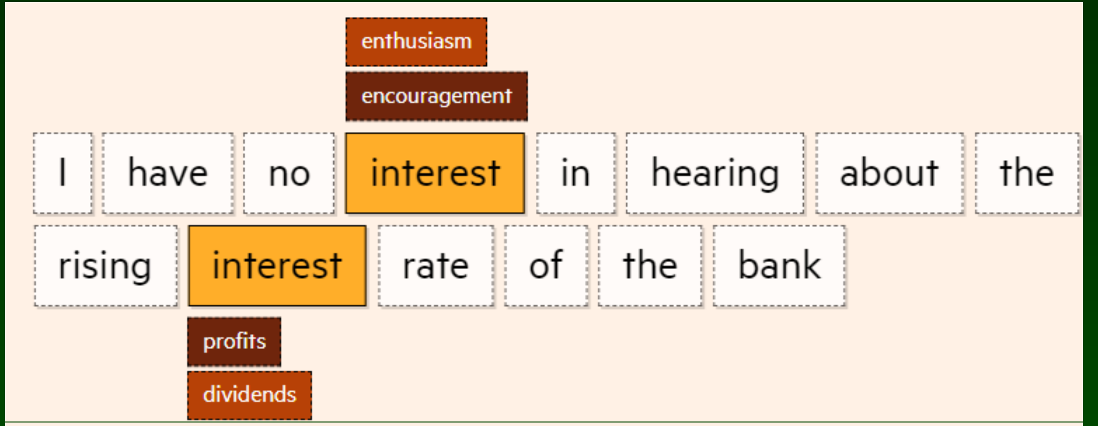
Word2Vec for embedding of words as vectors in the parameter space, preserving similarity.

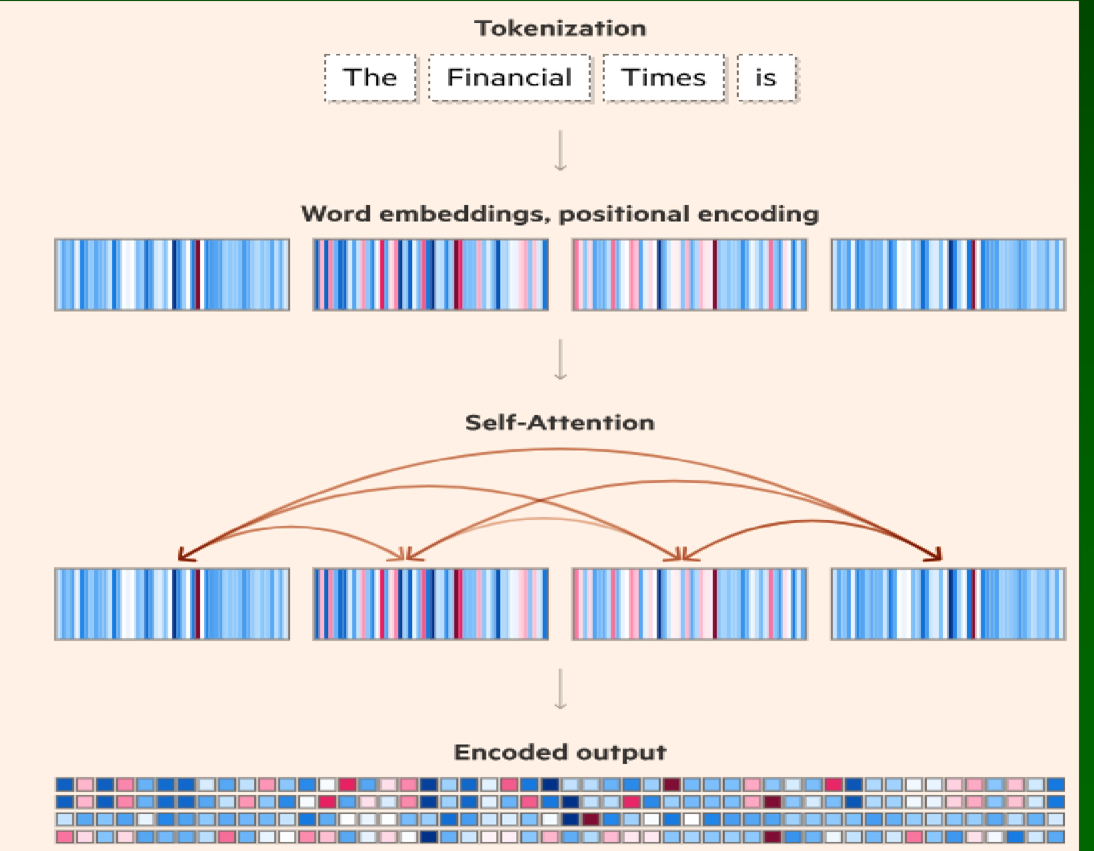
Transformer model published by Google in June 2017: Attention is all you need, started the generative AI era.

A key concept of the architecture is self-attention to understand relationships between tokens, linking each token in text to other tokens important to understand its meaning.

Vector representations of concepts that are semantically related become similar to each other => GPT interprets inputs using meaningful synonyms.

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

Generative Pre-trained Transformers or GPTs are now most popular algorithms for construction of large models.
Simple intro to GPT networks on Youtube.

Attention: given a sequence of tokens (words, image patches), how relevant is each input token to other tokens?

Attention vectors capture context (embedding, semantics) + encode relative positions (syntax) of words.

Many blocks for input encoding, recurrent processing.

Generation of words/images is not unique, so various outputs are selected to match whole context created by external prompts and current internal activity
=> creativity + human alignment.

GPT is not a deterministic algorithm, even if it was trained to say so! Networks are containers of knowledge, data flow. Not a program following instructions.

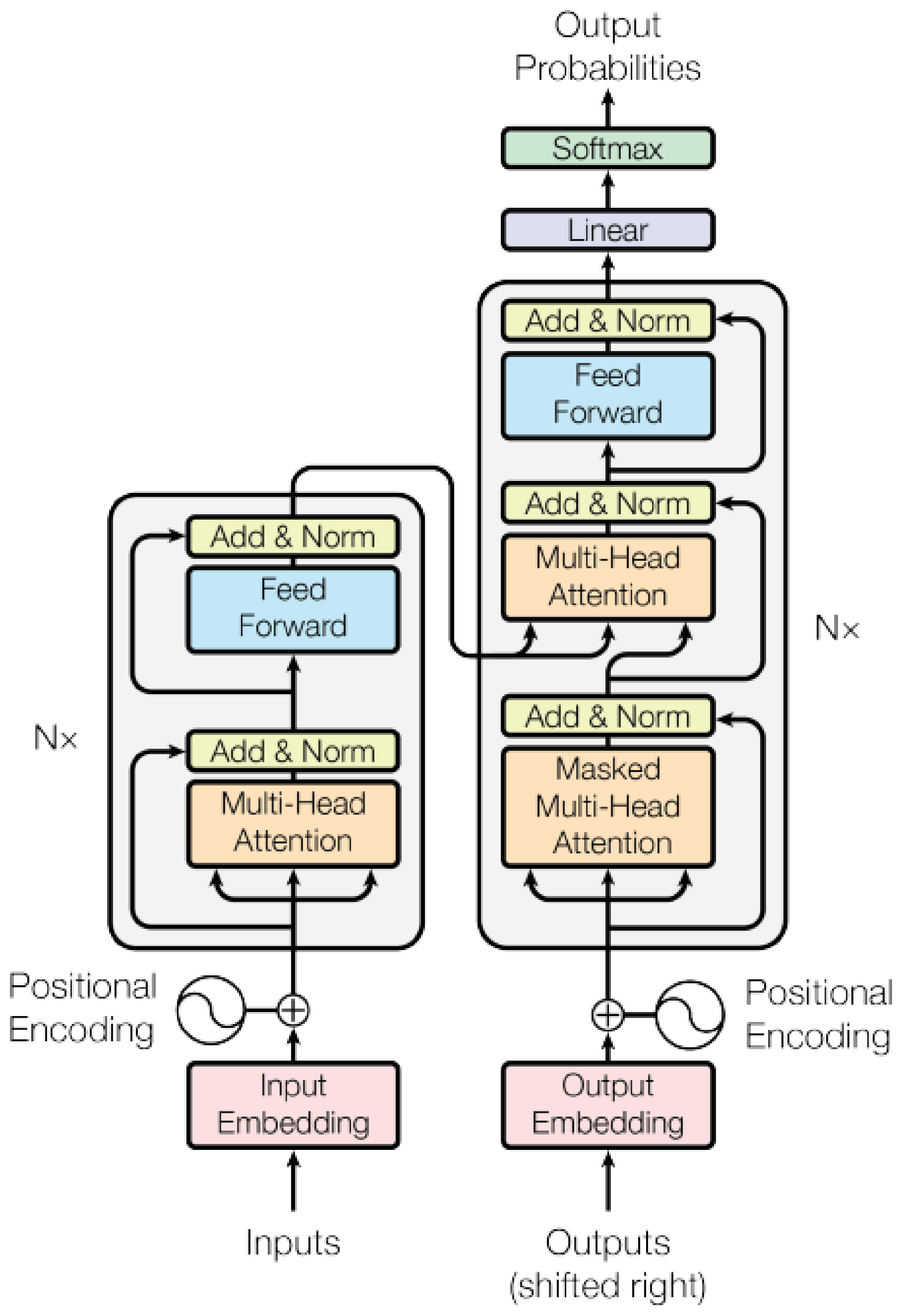


Figure 1: The Transformer - model architecture.

Transformers and spreading activation

Predictive AI: search + heuristics.

Generative AI: spreading activation networks, binding relevant information.

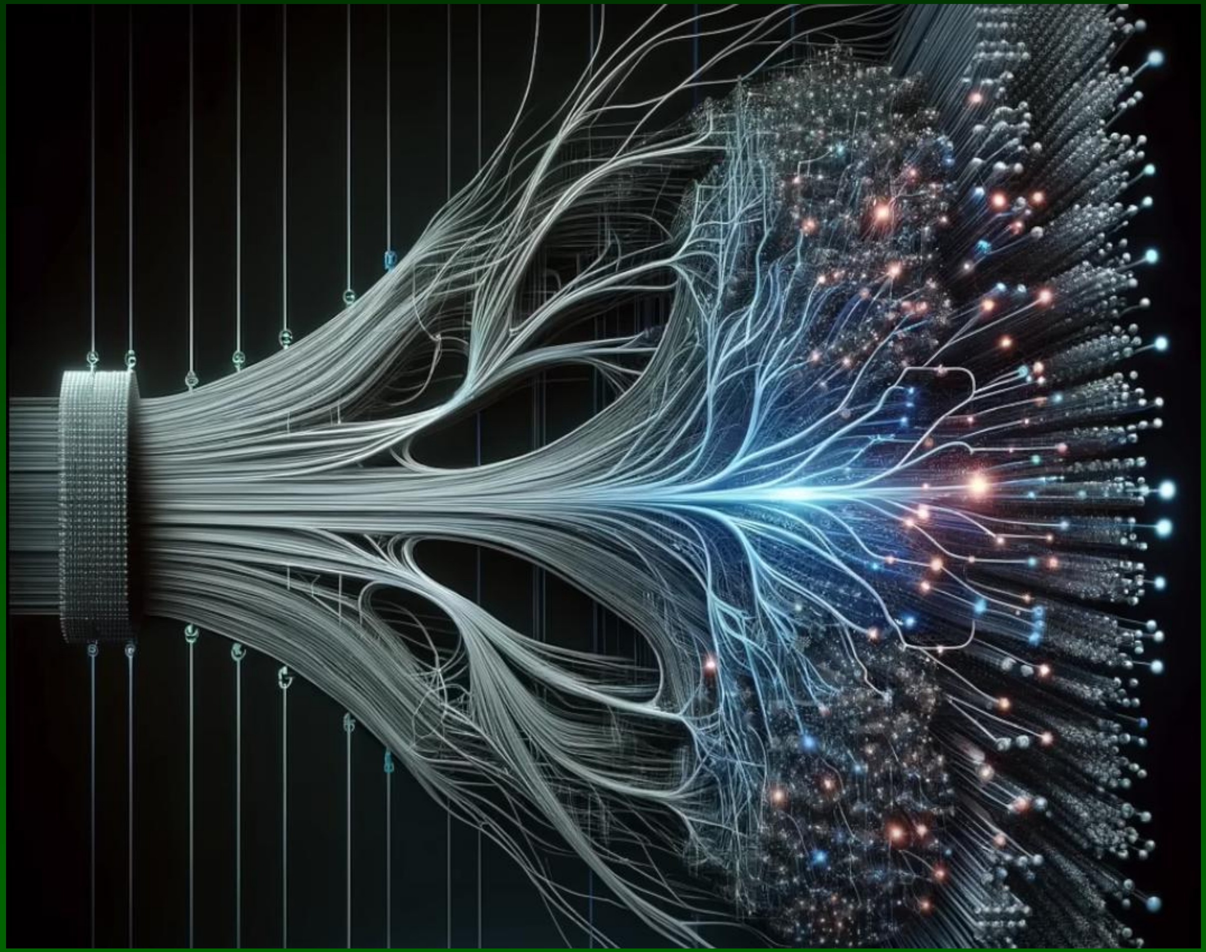
GPT = Generative Pre-trained Transformer

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

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

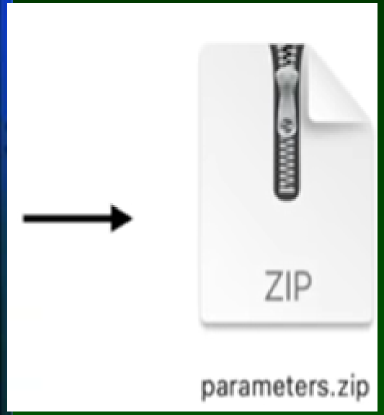
LLM visualization https://bbycroft.net/llm

List of 37320 variants open-source language models (27/10/2024).



Duch et al., Towards Understanding of Natural Language: Neurocognitive Inspirations. 2007

Creating LLM/LMM: information compression



Training >10 000 GB text

12 days of work 6000 GPU V100 units
2 mln \$, 10^{24} or bln bln bln operations.

parameters ≈ 70 GB

Llama 2 (Meta's open model) has only 70 bln parameters. Compression 10000:70 ≈143 x.
Stronger compression – more errors but greater creativity, unusual associations.
Precise questions (+prompt) activate LLM parameters allowing for correct associations.

Energy: one image ≈ 10 Wh ≈ smartfon charge, 1 grosz, half an hour of human brain energy use!
Games, crypto mining still use much more energy than AI, but energy use become significant.

LLM/LMM world

List of 131 LLM/LMM models (10/2024), including 39 open source models.

HuggingFace Open LLM Leaderboard, from 10 mln to 238 bln parameters.
3 most popular families, with many variants: Llama 3, Mistral, Falcon, GPT-2, Gemma + Phi 3, Granite.

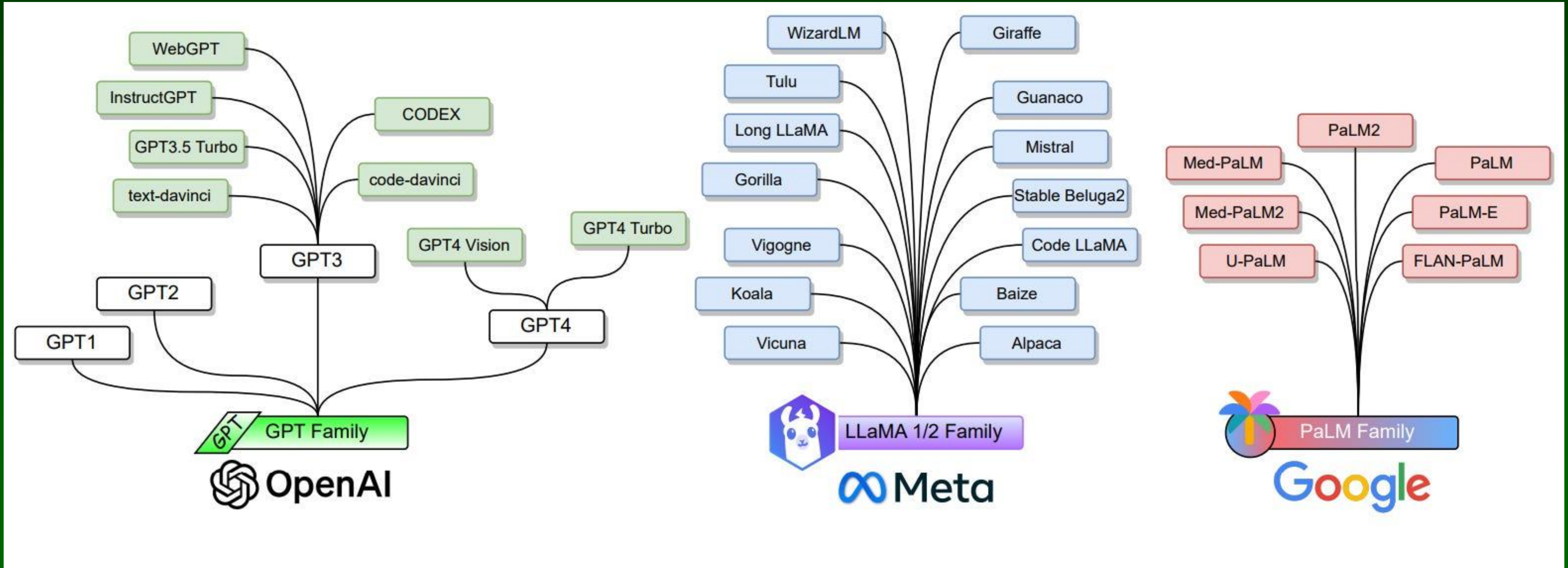
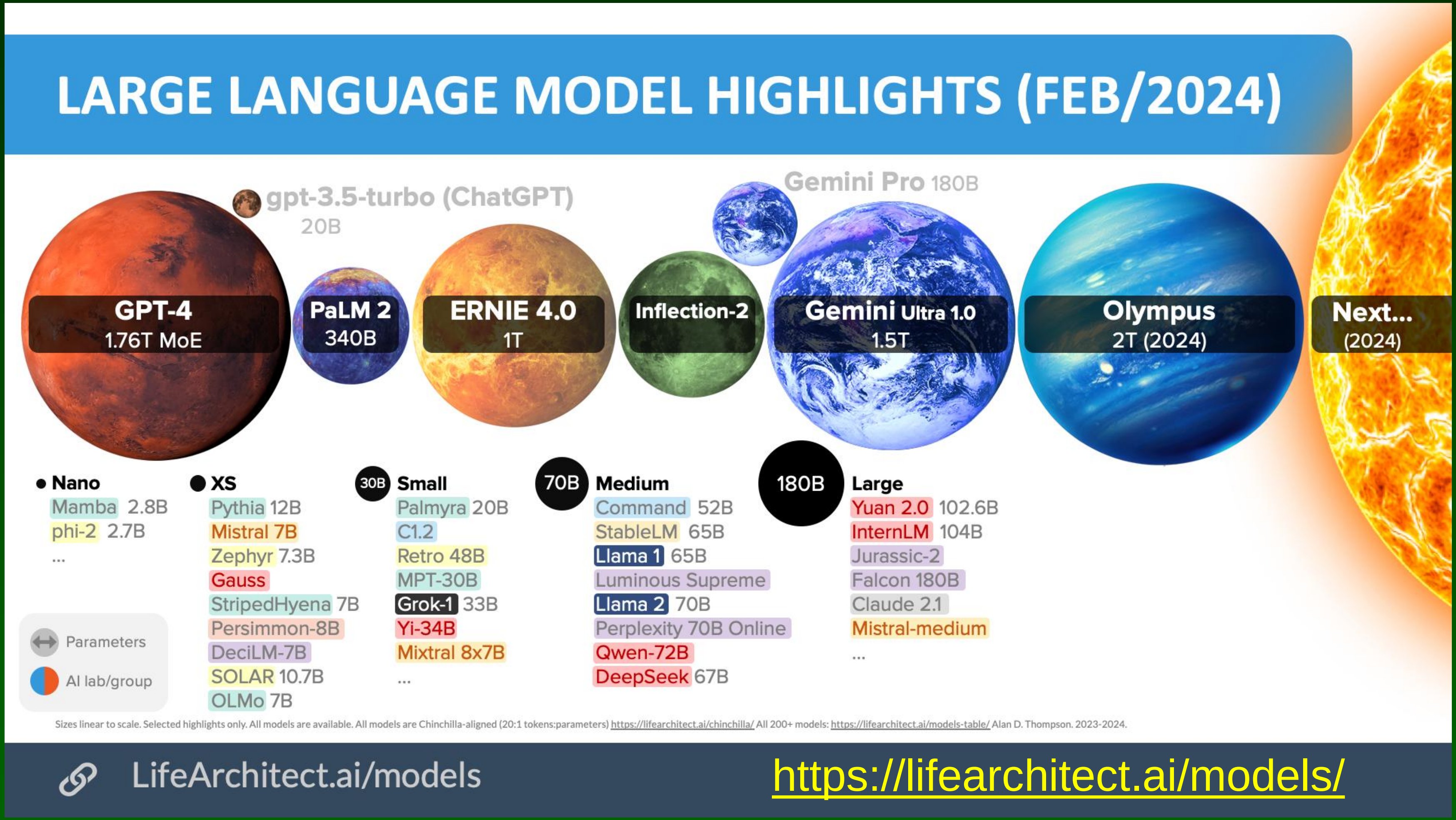


Fig. 8: Popular LLM Families.

LLM, LMM



Multimodal systems, all modalities in training data simultaneously: Anthropic Claude-3, Google Gemini, Pro/Ultra, several GPT-4 versions, much larger GPT-5 coming?

Edge SLMs

GPT-4 has 1.7 trillion parameters (probably). Much larger models are coming, but is it necessary?

Small Language Models (SLMs) may even be more accurate, can process and generate human-like text, videos etc.

DeepSeek-Coder-V2 MoE framework 16B or 236B, activate only 2.4B/21B.

LlaMA-13B surpasses GPT-3 despite being over 10 times smaller,

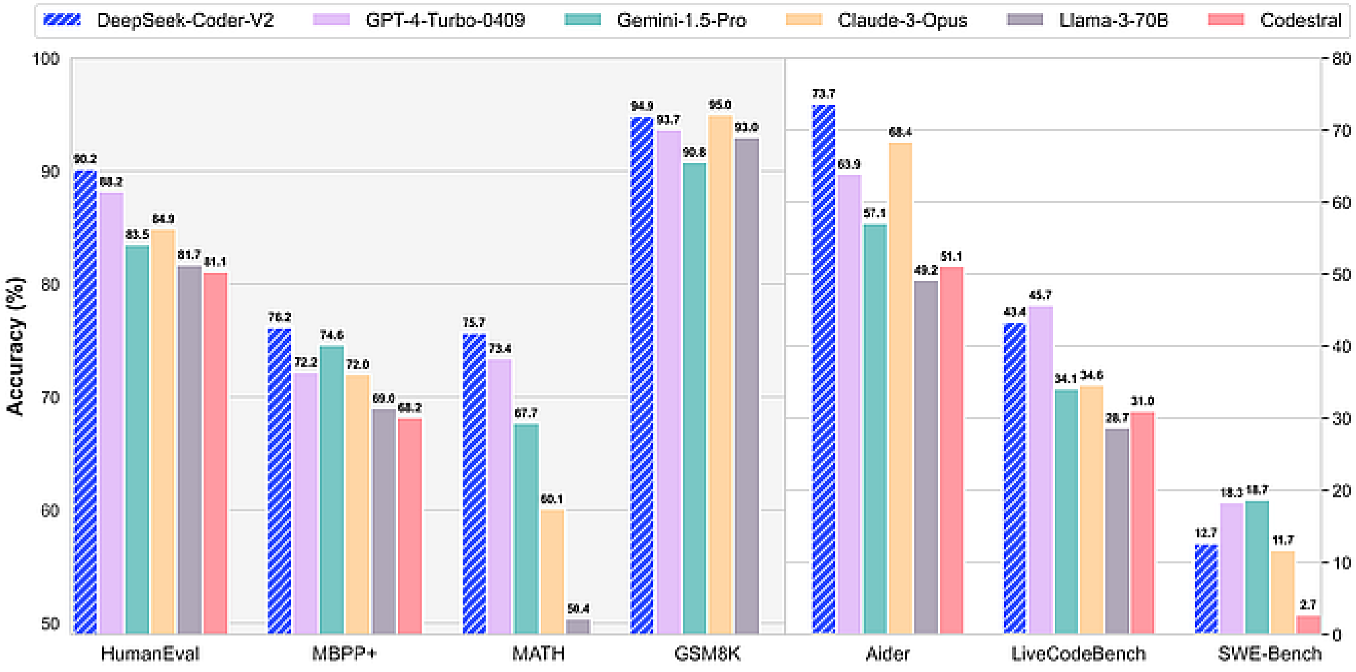
Microsoft Phi-2 and Phi-3 (2.6-14 B) outperforms 25x larger models. LLaMA-7B is close to PaLM-540B.

Ministral 8B is the best model (10/2024)

Gemini Nano is in Android phones.

Model	Knowledge & Commonsense					Code	Math	Multilingual		
	MMLU	AGIEval	Winogrande	Arc-c	TriviaQA	HumanEval pass@1	GSM8K maj@8	French MMLU	German MMLU	Spanish MMLU
Gemma 2 2B	52.4	33.8	68.7	42.6	47.8	20.1	35.5	41.0	40.1	41.7
Llama 3.2 3B	56.2	37.4	59.6	43.1	50.7	29.9	37.2	42.3	42.2	43.1
Ministral 3B	**60.9**	**42.1**	**72.7**	**64.2**	**56.7**	**34.2**	**50.9**	**49.1**	**48.3**	**49.5**
Mistral 7B	62.5	42.5	74.2	67.9	62.5	26.8	51.3	50.6	49.6	51.4
Llama 3.1 8B	64.7	44.4	74.6	46.0	60.2	**37.8**	61.7	50.8	52.8	54.6
Ministral 8B	**65.0**	**48.3**	**75.3**	**71.9**	**65.5**	34.8	**64.5**	**57.5**	**57.4**	**59.6**

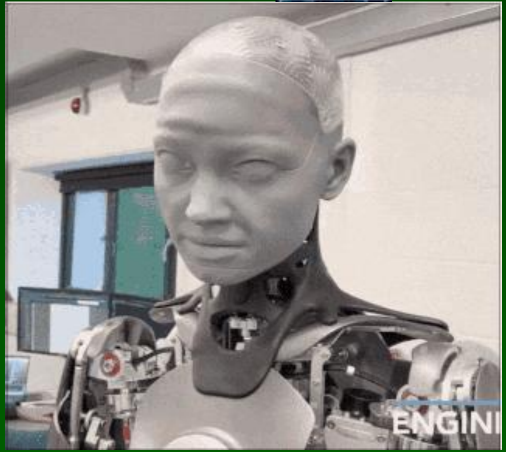
Table 1: Ministral 3B and 8B models compared to Gemma 2 2B, Llama 3.2 3B, Llama 3.1 8B and Mistral 7B on multiple categories



GameNGen, programs in mind

Generative intelligence (a small diffusion model) creates and executes a Doom game without writing program. This is imagery! https://gamengen.github.io

From calculator to superhuman AI



Reasoning: 1997–Deep Blue wins in chess;
2016 –AlphaGo wins in Go; 2017 Alpha GoZero 100:0.

Open Games: 2017–Poker, Dota 2; 2019-AlphaStar Starcraft II,
2022 Stratego, Diplomacy, Bridge – what is left?

Perception: speech, vision, recognition of faces, emotions, personality traits, sexual, political and other preferences ...

Robotics: 2020 Atlas robot (Boston Dynamics) backflip and parkour, autonomous vehicles, 2023 Tesla Optimus Gen 3, Ameca, Unitree G1.

Science: 2020 AlphaFold 2/3, now 620 M 3D proteins, 2023-GNoME (Deep Mind) 2.2 mln structures; AlphaProof, AlphaGeometry 2.

Creativity and imagination: GAN revolution, Dall-E, Midjourney, Leonardo, Canva, Stable Diffusion, Suno, AIVA, music composers.

Language: 2011–IBM Watson wins in Jeopardy; 2018–Watson Debater 2020: BERT answers questions from SQuAD database.

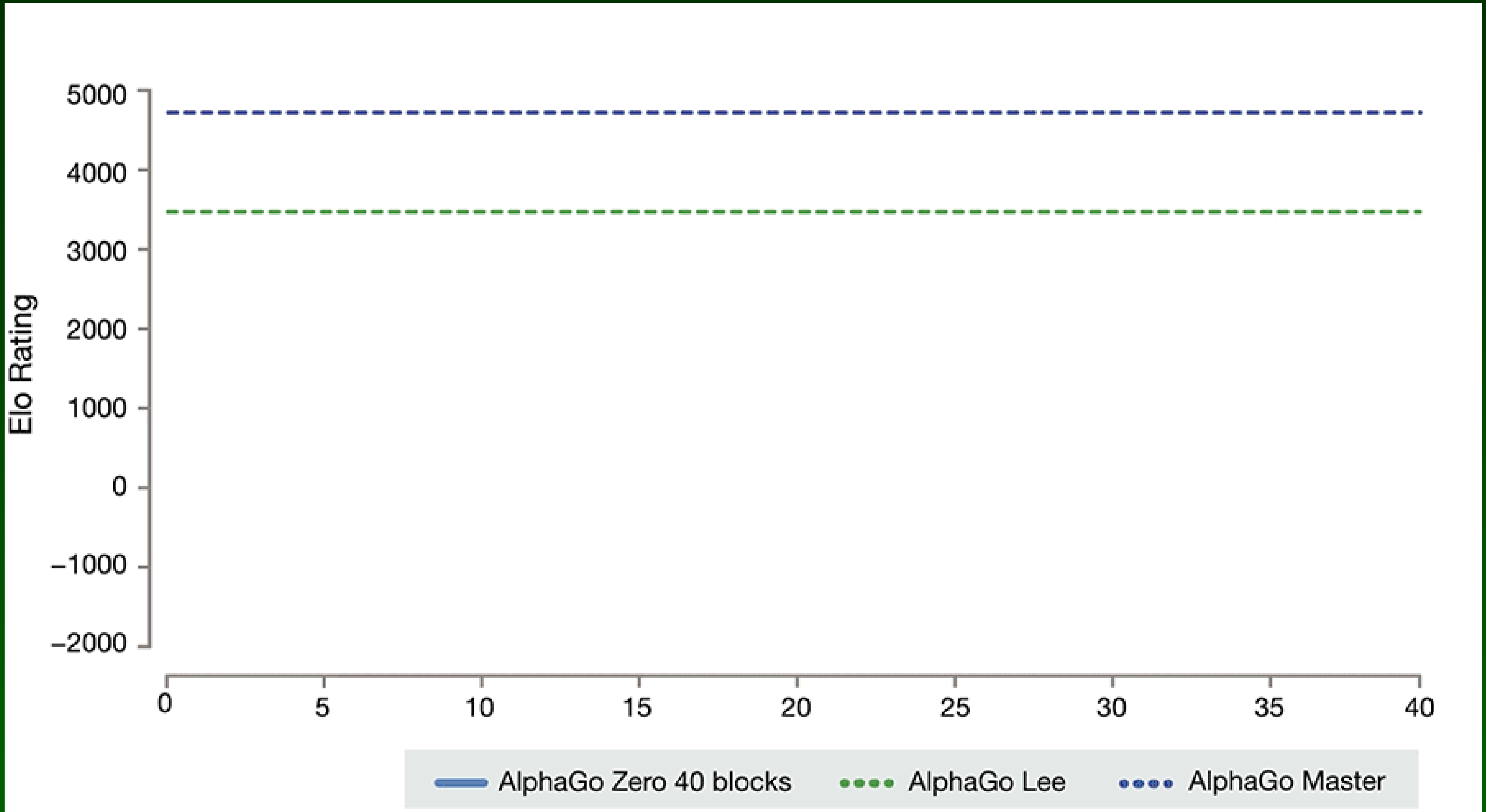
Cyborgization: BCI, brain-computer symbiosis, soon?

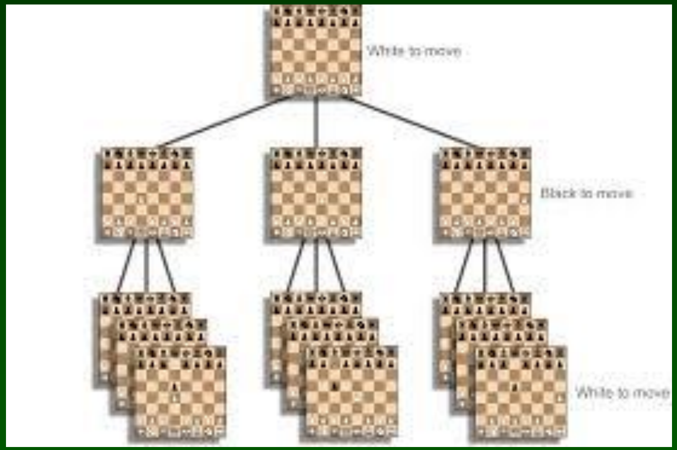
We desperately search for tests that show human superiority over AI.

Deep Mind

NOBELPRISET I KEMI 2024
THE NOBEL PRIZE IN CHEMISTRY 2024
KUNGL. VETENSKAPS-AKADEMIEN
THE ROYAL SWEDISH ACADEMY OF SCIENCES
David Baker
University of Washington
USA
Demis Hassabis
Google DeepMind
United Kingdom
John M. Jumper
Google DeepMind
United Kingdom
"för datorbaserad proteindesign"
"for computational protein design"
"för proteinstrukturprediktion"
"for protein structure prediction"
#NobelPrize
THE NOBEL PRIZE

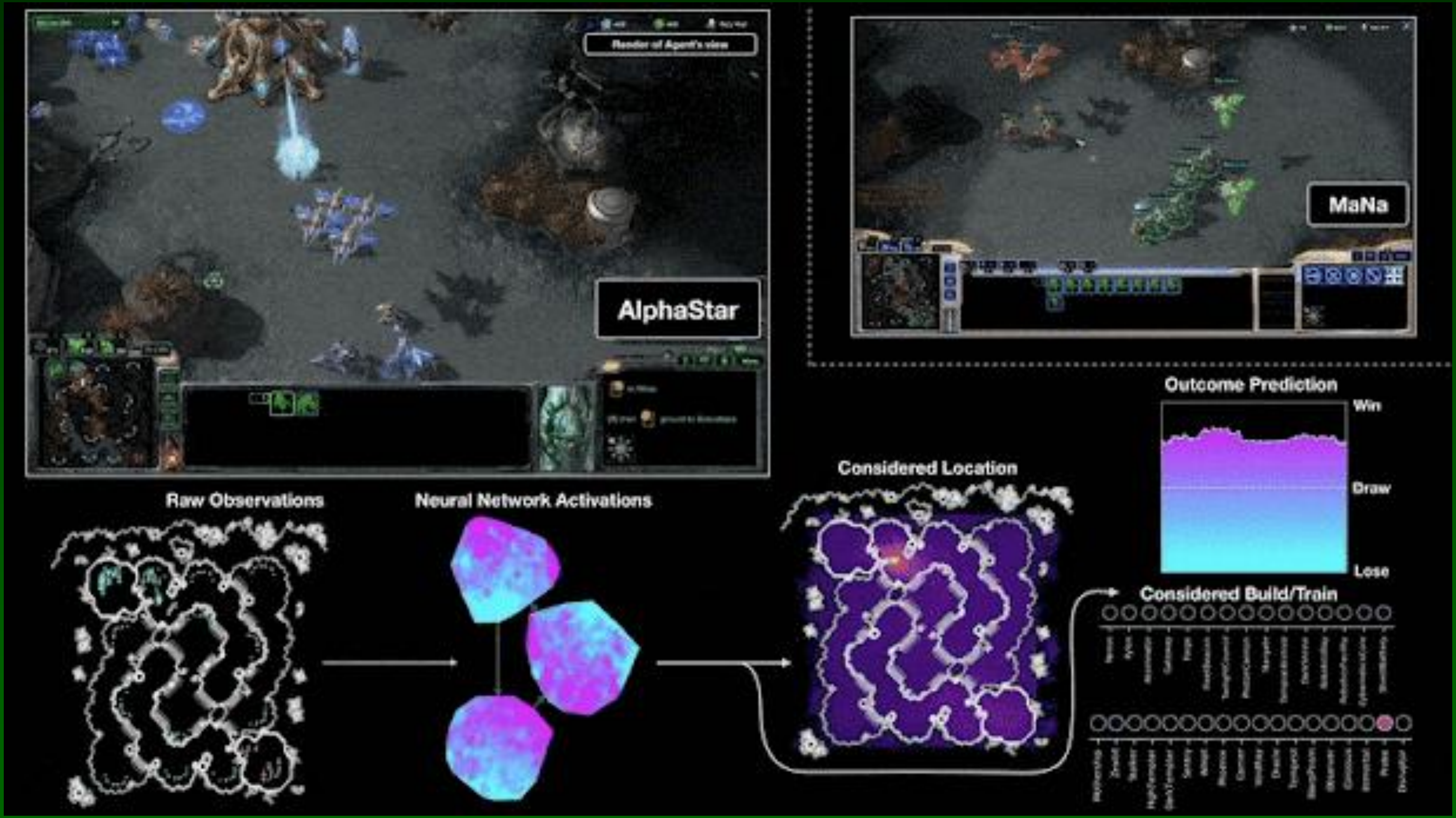
AlphaGo Zero and human knowledge





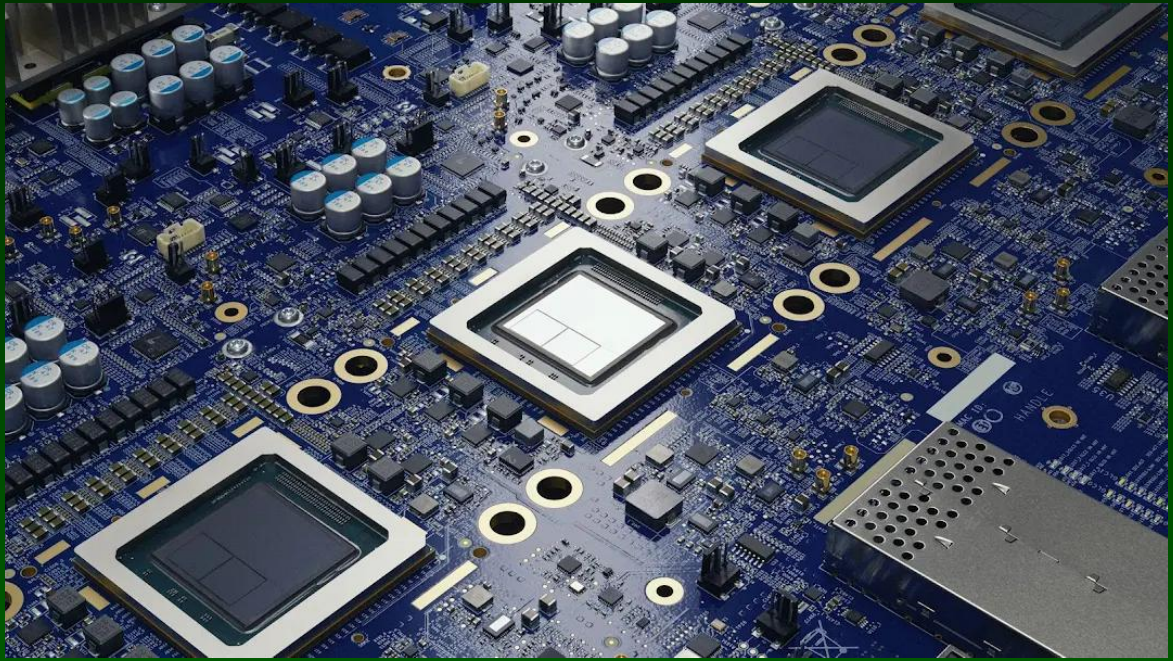
Superhuman level in the strategic game of Go, playing against its own copy. Search + NN as heuristics.
Human knowledge becomes irrelevant, teaching human strategies decreases final level!
Shocking news: Ruoss … & Genewein, T. (2024). *Grandmaster-Level Chess Without Search* (arXiv:2402.04494) 270M parameter transformer model, 1-step intuitive decisions!

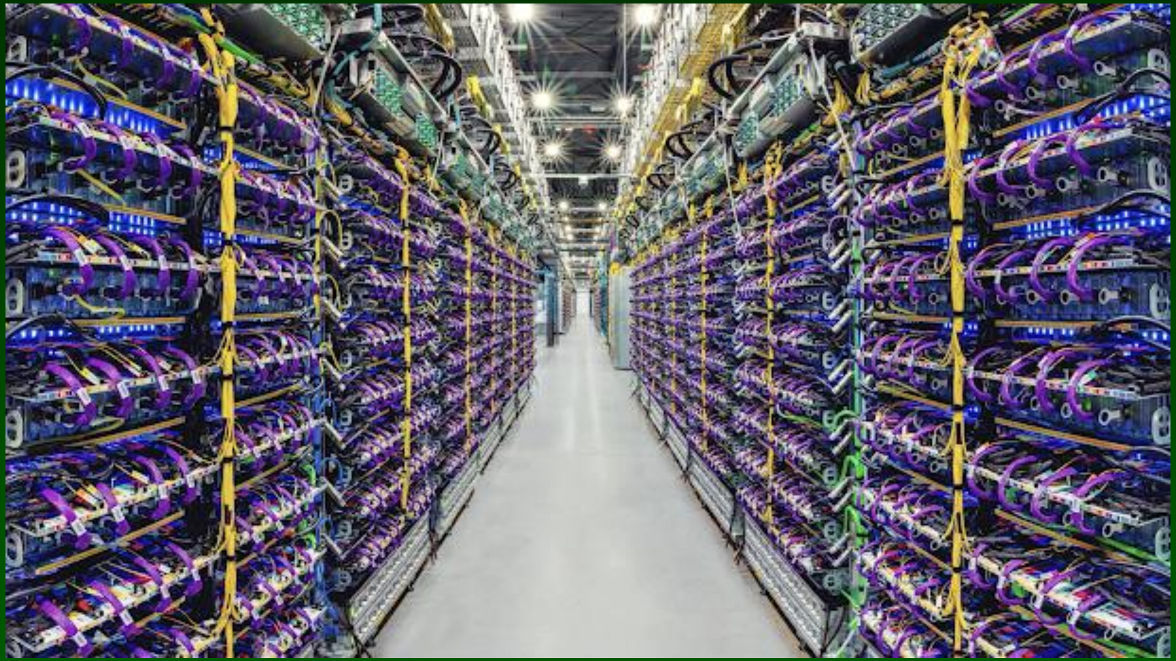
AlphaStar Strategic Game



Deepmind.google alphastar – mastering the real-time strategy game Starcraft-II (2019)
Open, strategic games, like war games.

AlphaChip

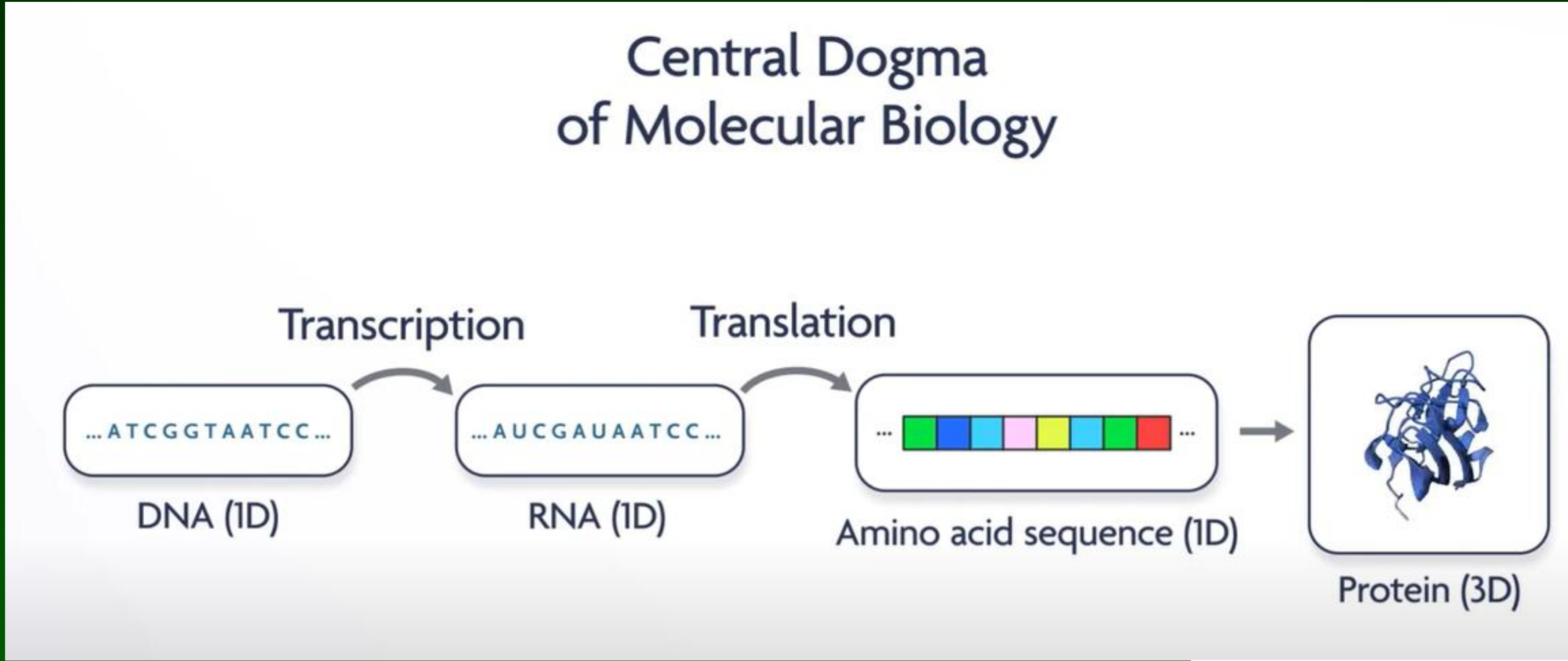


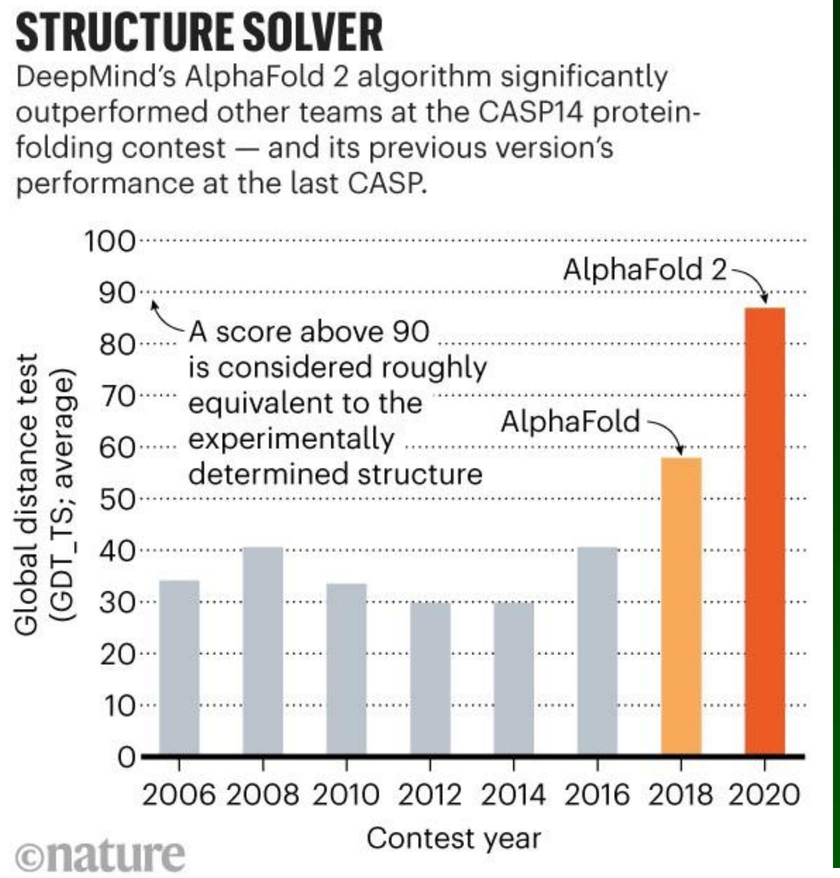


AI has accelerated and optimized chip design, and its superhuman chip layouts are used in hardware around the world. Such layouts were used in the last three generations of Google's custom AI accelerator, the Tensor Processing Unit (TPU). AI is here a partner helping humans.

Mirhoseini, A., Goldie, A., Yazgan, M. *et al.* A graph placement methodology for fast chip design. *Nature* **594**, 207–212 (2021), Addendum 26.09.2024

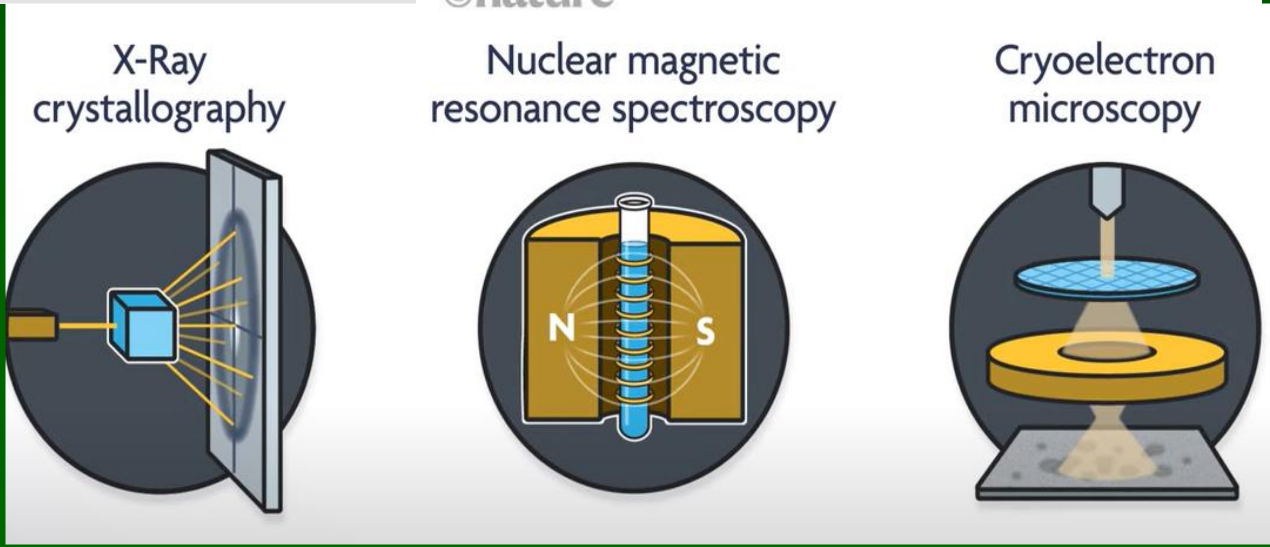
From transformers to protein folding



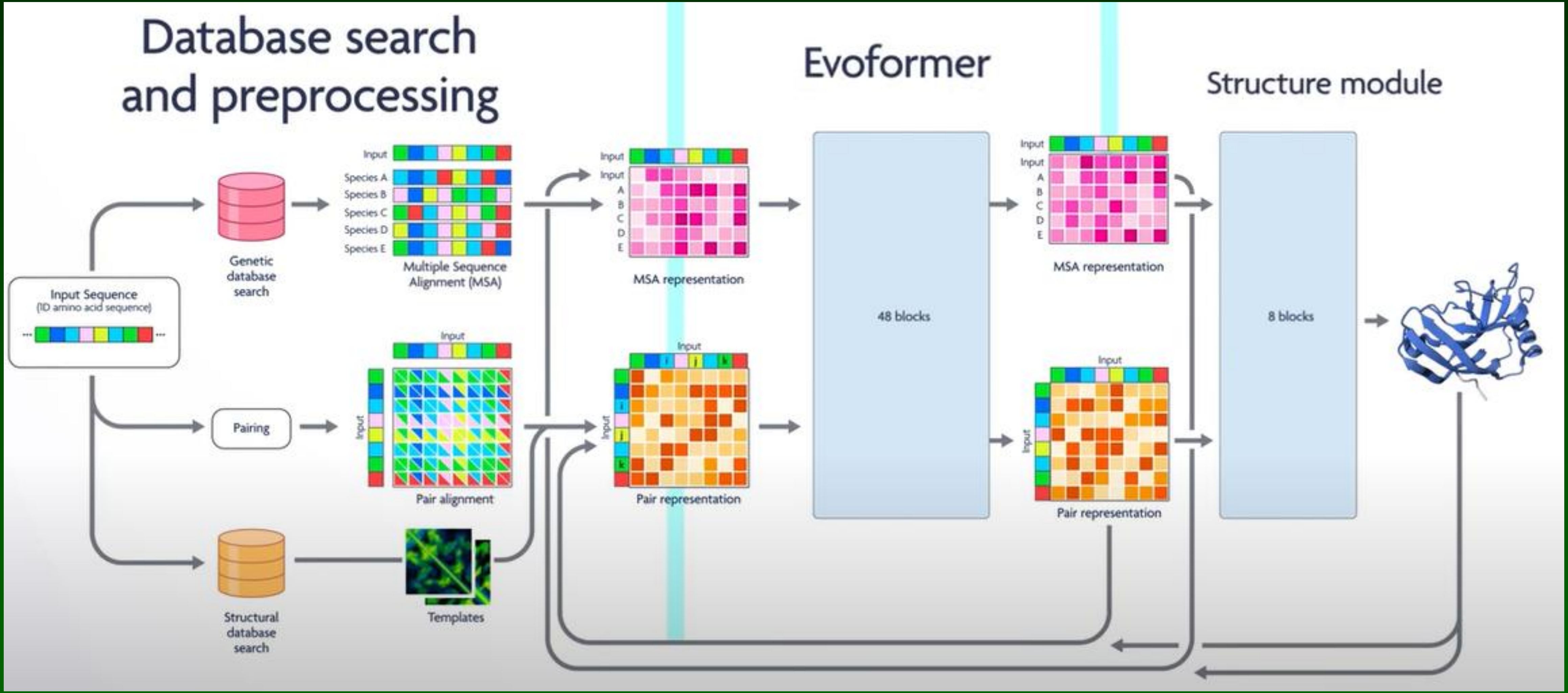


R. B. Altman, A Holy Grail — The Prediction of Protein Structure, New England J. of Medicine **389** (2023)

Google Deep Mind AlphaFold 2 has reached accuracy comparable to experiments on about 100000 proteins and released 218 mln 3D protein structures.

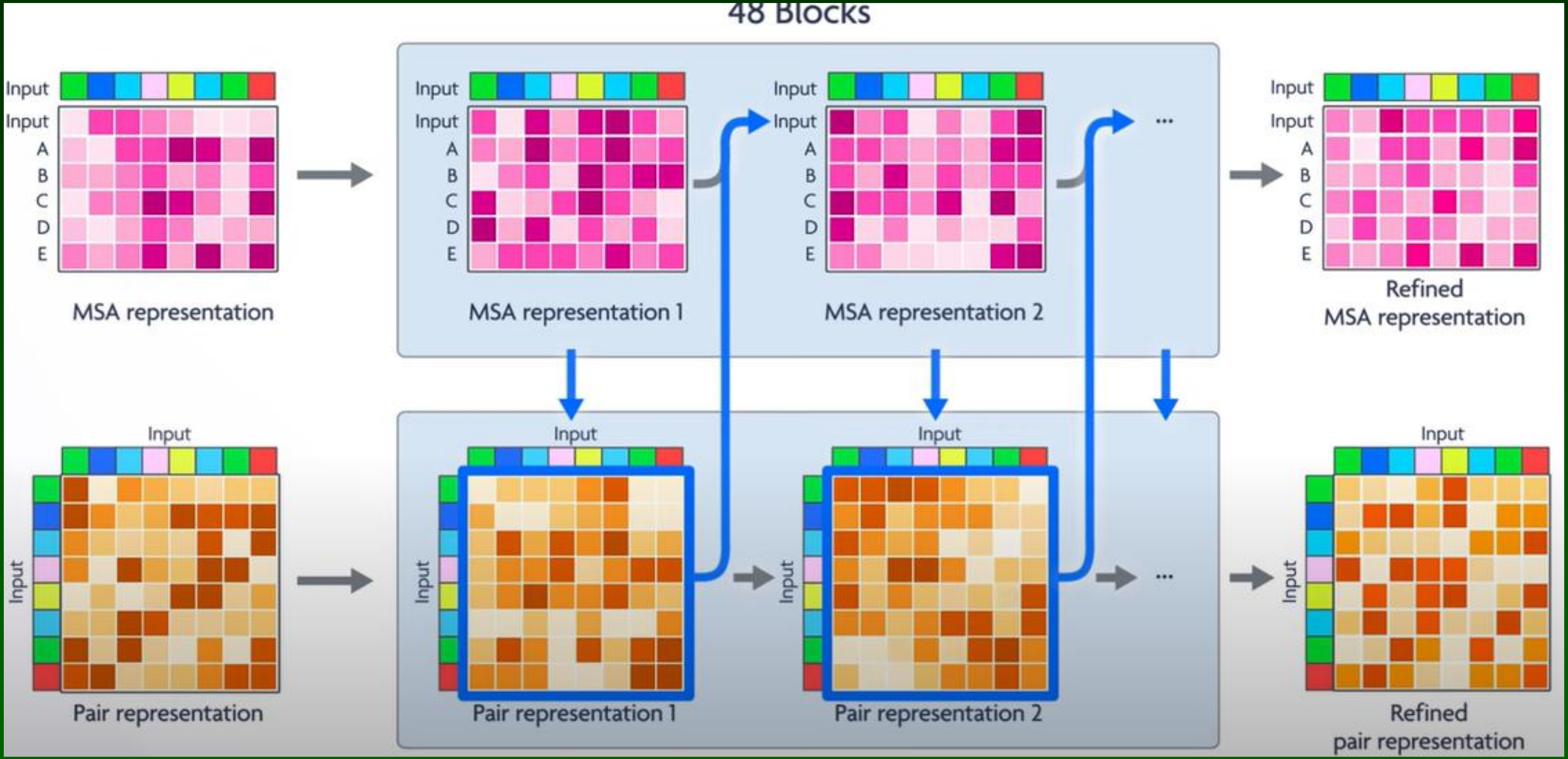


AlphaFold architecture



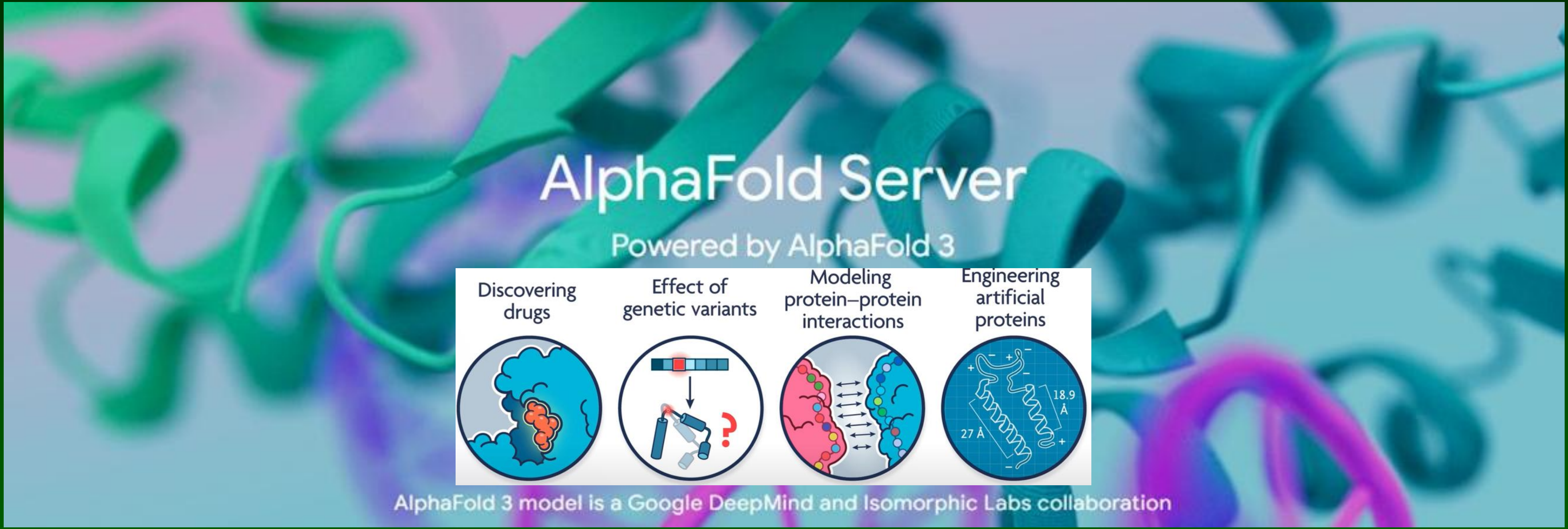
Google Deep Mind AlphaFold is using evolutionary information from genetic databases to find similar sequences, and compares pairing matrix to experimentally derived structures. Evoformer evaluations importance of positions in sequences and in sequences of similar species.

AlphaFold MSA



Evoformer uses neural network to relate position of each pair of residues with the third one residue, iteratively influencing MSA, until refined pair representation and MSA representation is stable. Structure module is a network that creates invariant 3D structure (rotations, translations, physical constraints).

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).

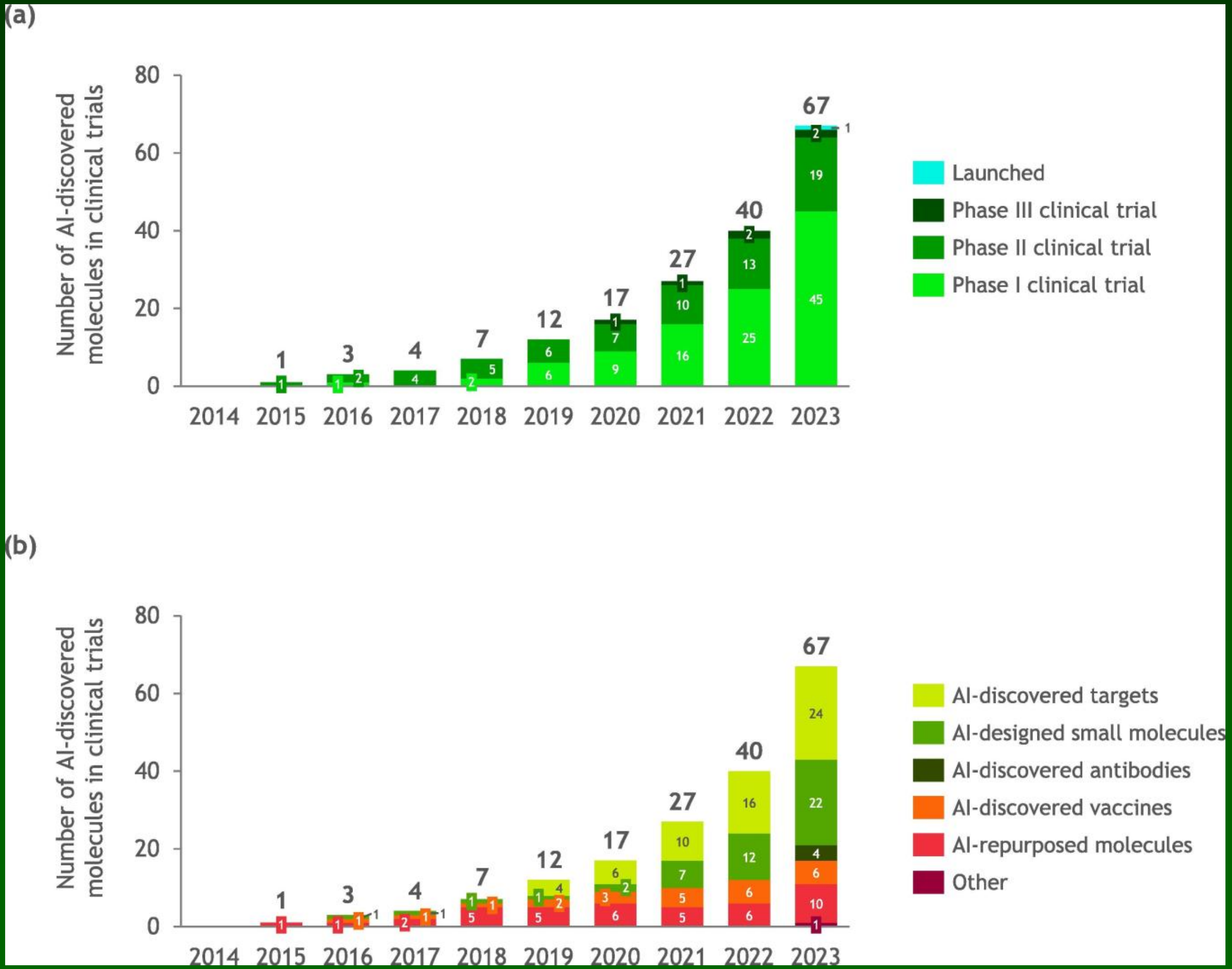
AlphaFold 3 predicts the structure and interactions of all of life's molecules, AlphaProteo generates novel proteins for biology and health accelerating research in nearly every field of biology/molecular medicine.

AI-discovered drugs in clinical trials

Number of molecules discovered by AI-first Biotechs that have entered clinical trials. Includes molecules that were partnered with pharma companies and excludes COVID-19-related molecules.

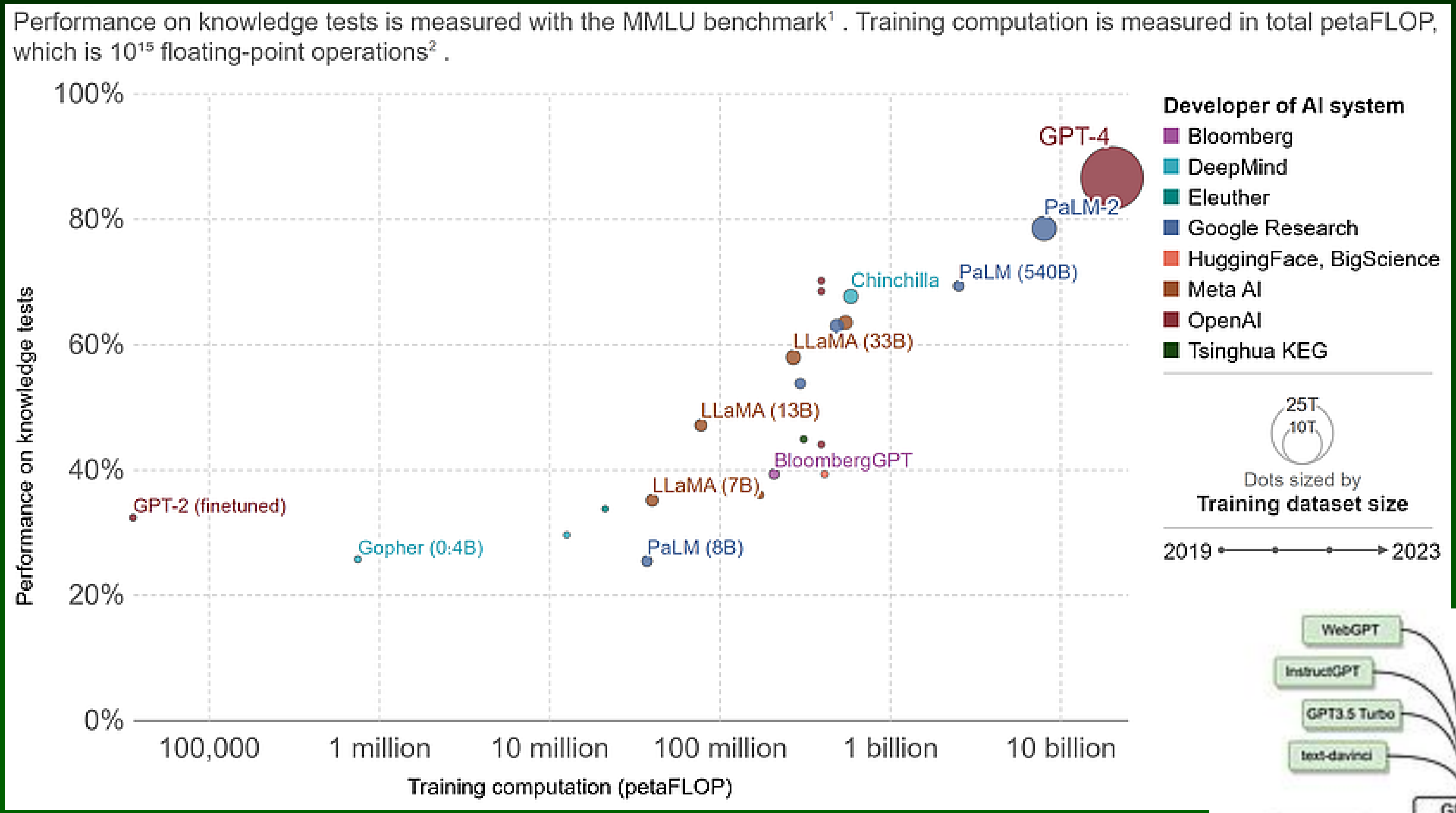
(a) AI molecules by clinical Phase.

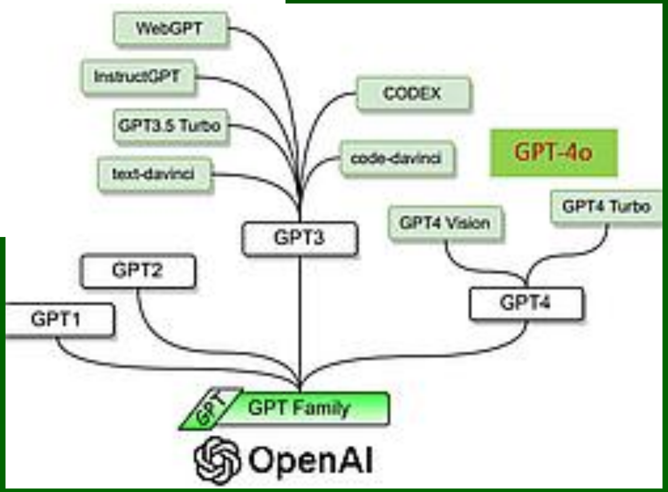
(b) AI-discovered molecules by mode-of-discovery.



Neural networks for science

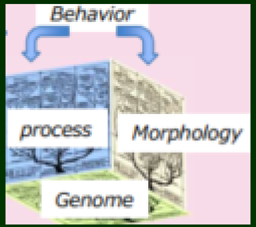
Performance vs. training size





MMLU database: ~16K questions in 57 subjects, from history to math. Average human 34.5%, expert in a single domain 89.8%, GPT-4 86.4%.

Mostly associative thinking, poor reasoning abilities. Scaling works, but autoreflection and longer deliberation is even more effective.

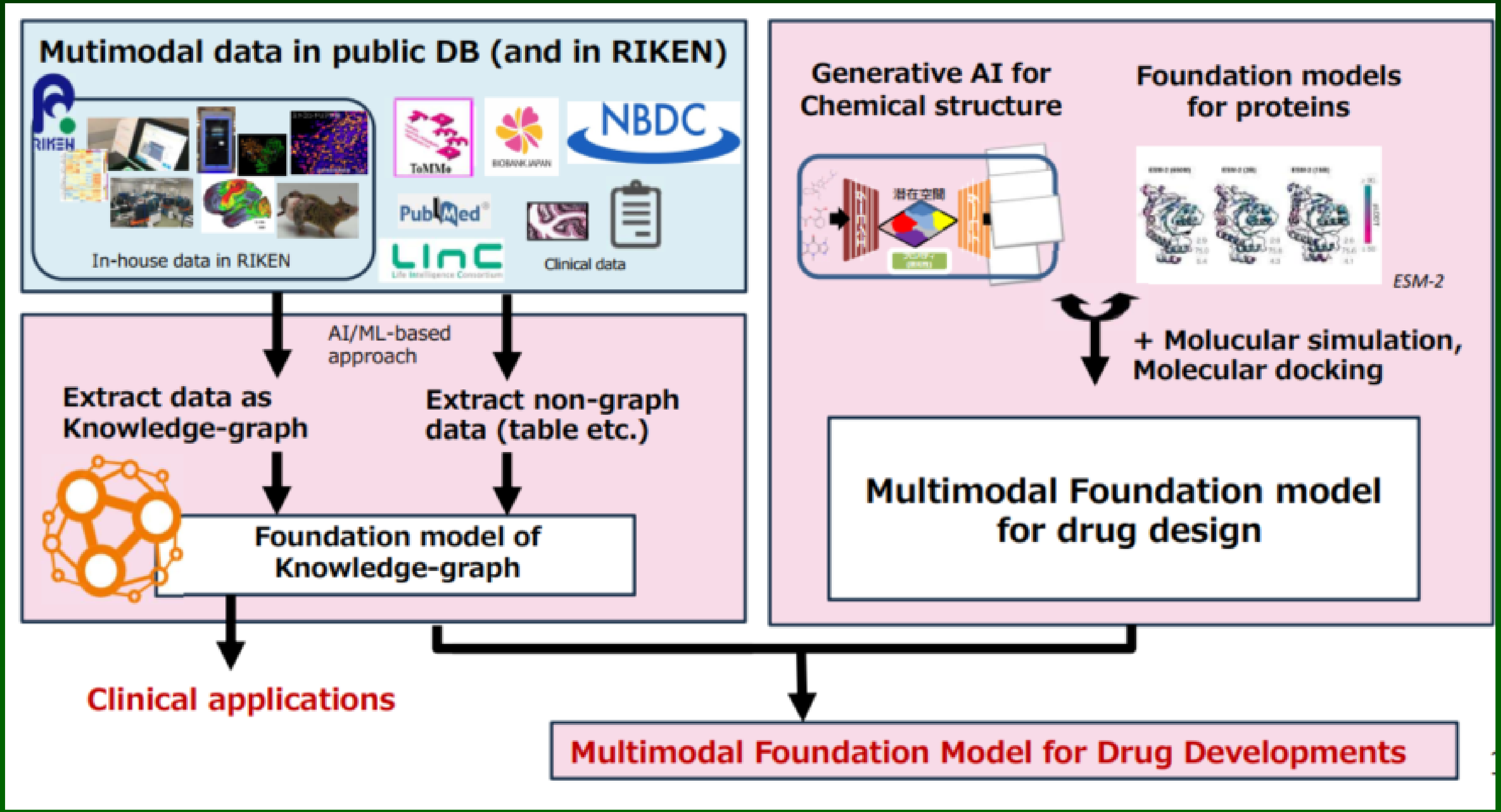


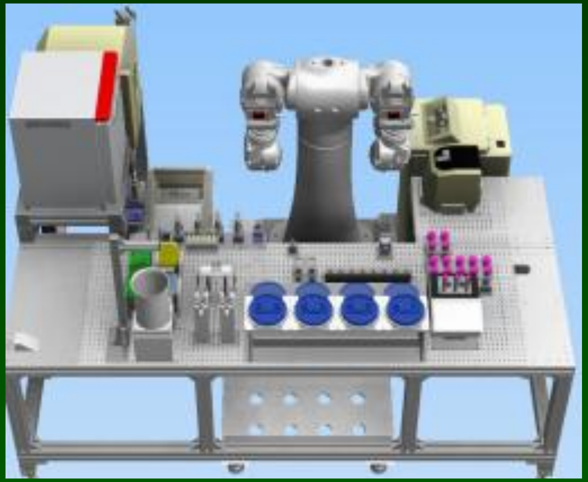
Building, training, and using large-scale models + hardware development (3/2024).
(a) AI development, LLMs/LMMs, scalable libraries and frameworks, AI workflows, data aggregation, model evaluation, alignment, etc.;
(b) design/build hardware and software systems, research automation;
(c) AI applications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

RIKEN plans to launch the new program in 2024 called **T**ransformative **R**esearch **I**nnovation **P**latform - **A**rtificial **G**eneral **I**ntelligence for **S**cience (TRIP-AGIS).
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 cell culture,genomics, cell-based screening, proteomics, metabolomics.

NIMH approach to phenomics

2008: The Consortium for Neuropsychiatric Phenomics

New approach RDOC NIMH: Mental disorders result from deregulation of large brain systems. **Research Domain Criteria** (RDoC) matrix based on **multi-level neuropsychiatric phenomics** will help to answer different types of questions.

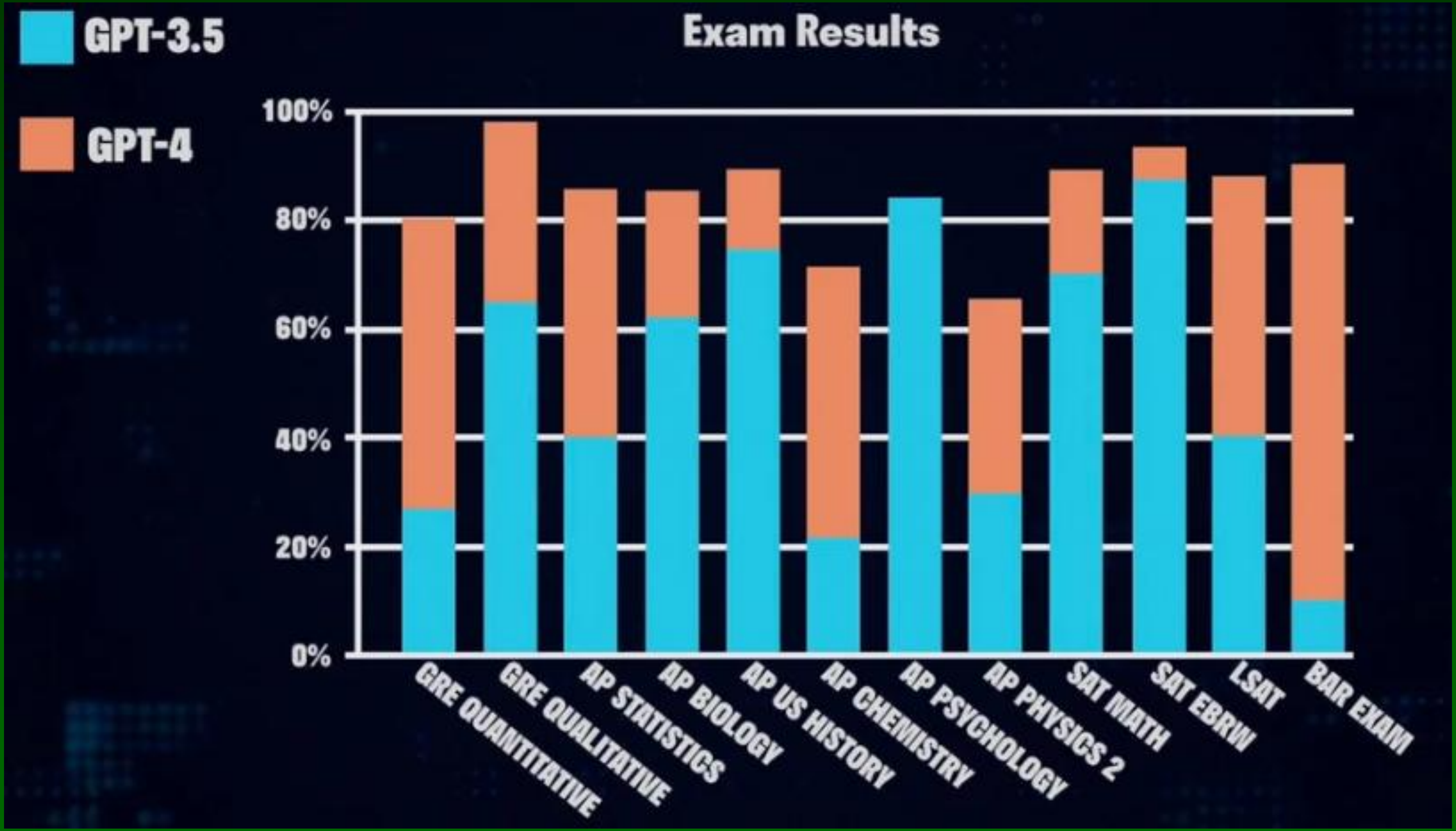
Network level is in the middle, can be connected to the mental and metabolic level via computational models.

These neural networks should be closely inspired to neuroscience: spiking neurons, ion channels, Hodgkin-Huxley models etc.

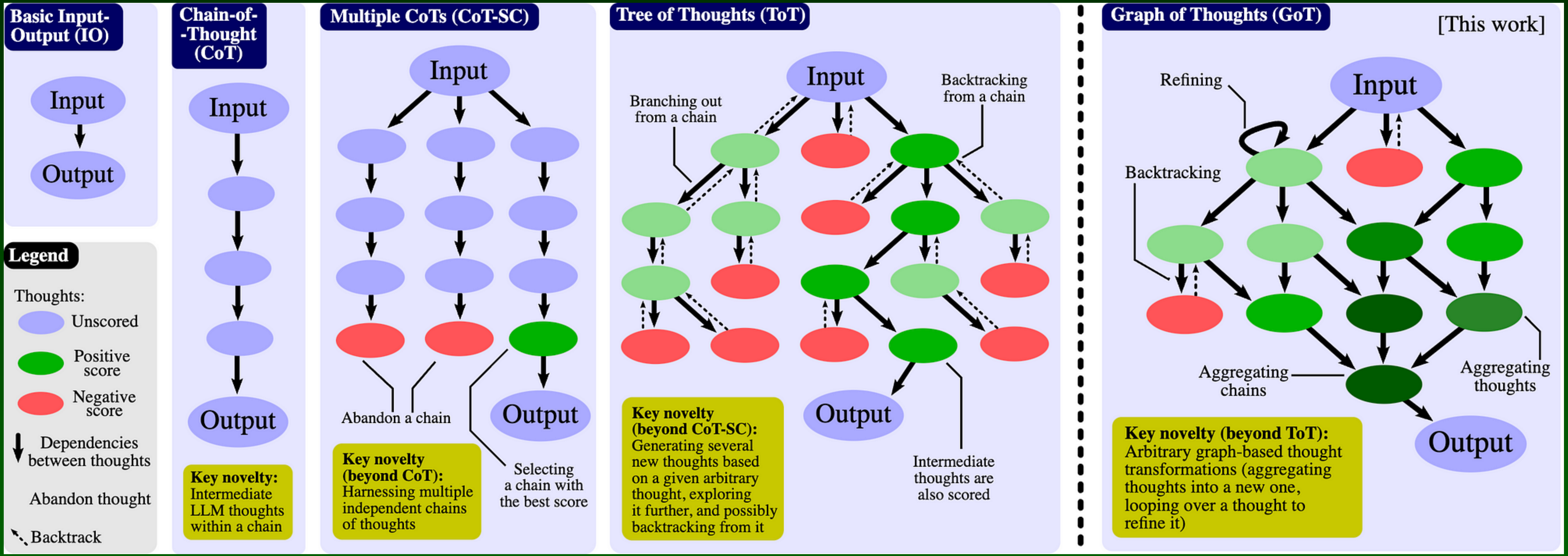
Complexity of biological organisms can be analyzed only with the help of AI tools. See my slides: Multi-level explanations in neuroscience. From genes to subjective experiences.

GPT-4 passes many exams …

GPT-4 performance on academic and professional exams (technical report).
Emergent properties, and the use of tools (plugins). Neural transformers.



Thinking – cognitive inspirations



Associations: input => output.

This is GOFAI! LLM associations in heuristic search.

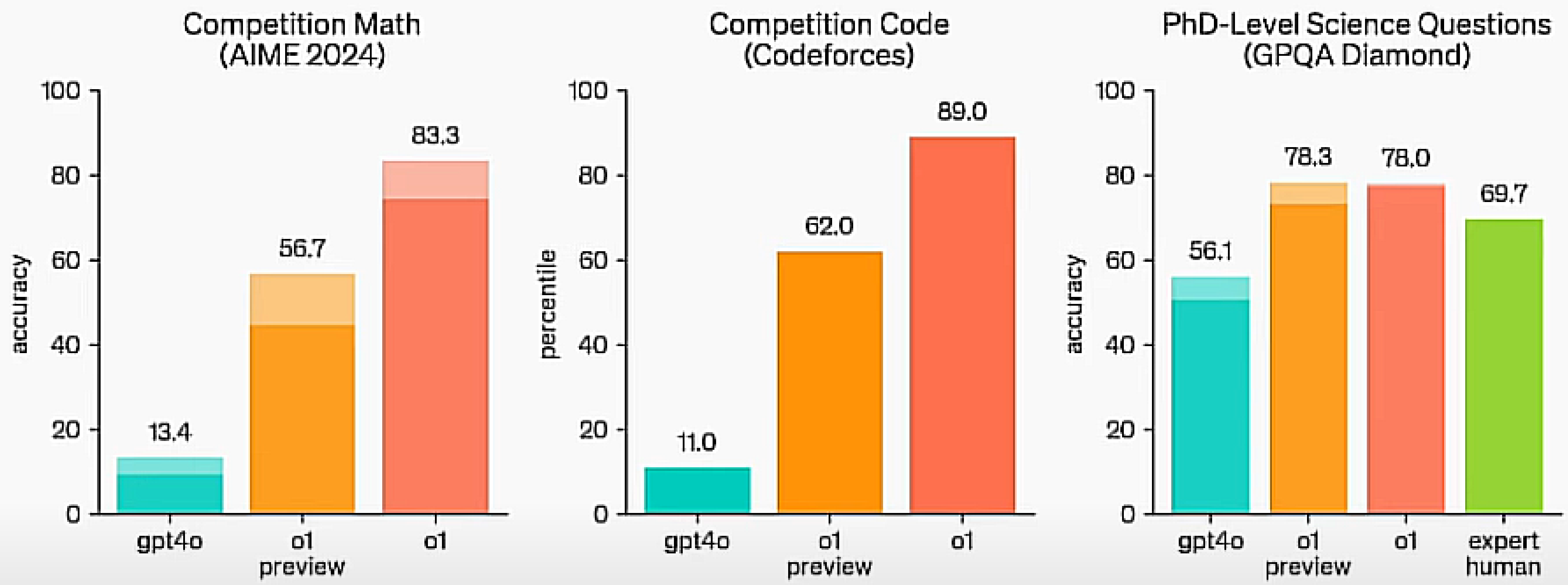
Chain of thought (CoT): step by step.

Tree of Thoughts (ToT): search for solutions.

Graph of Thoughts (GoT): like human reasoning.

Complex reasoning GPT4-o1

Huge improvement for problems that require complex reasoning (12.09.24).

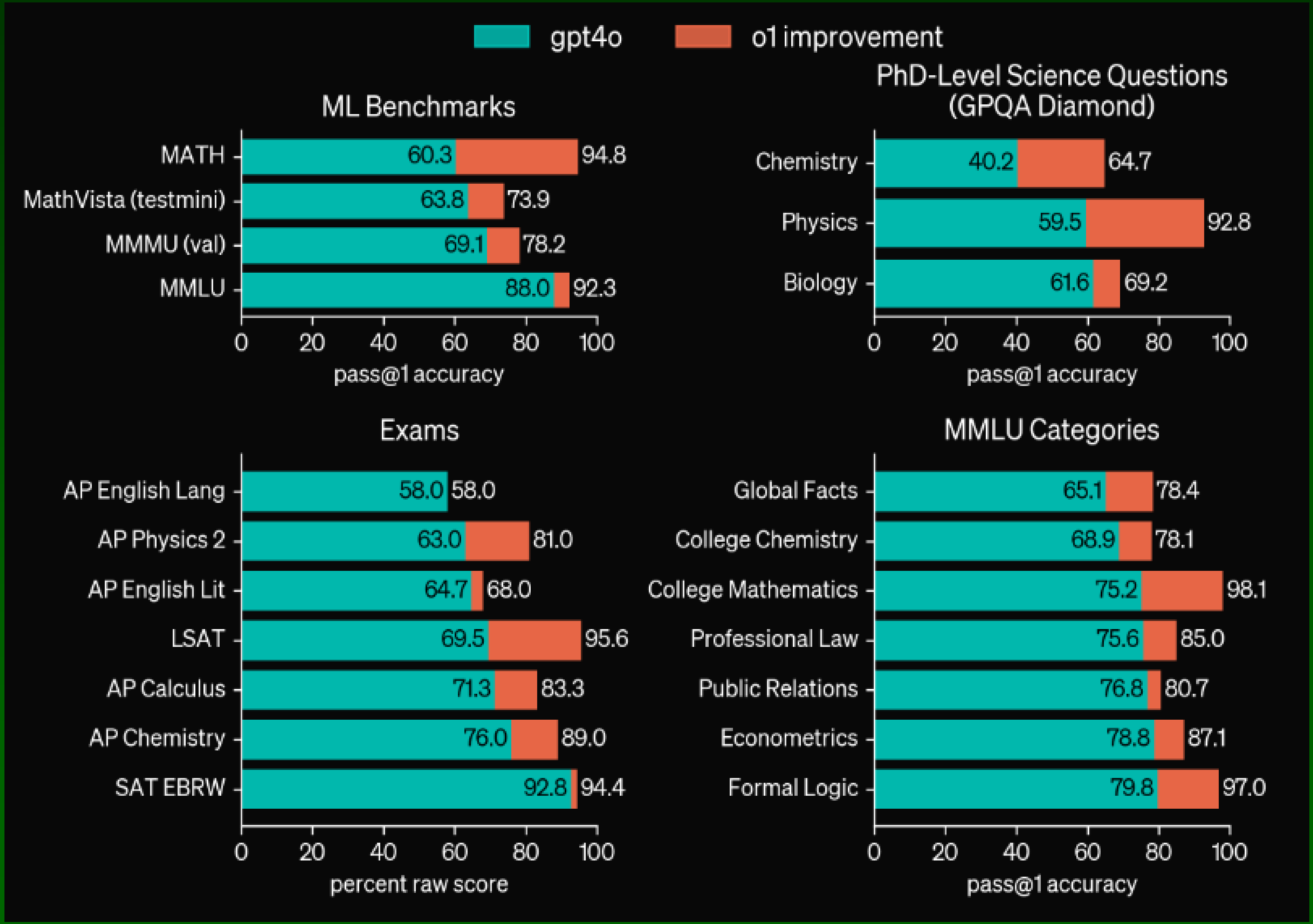


o1 greatly improves over GPT-4o on challenging reasoning benchmarks. Solid bars show pass@1 accuracy and the shaded region shows the performance of majority vote (consensus) with 64 samples.

gpt4o   o1 improvement

GPT4-o1 tests

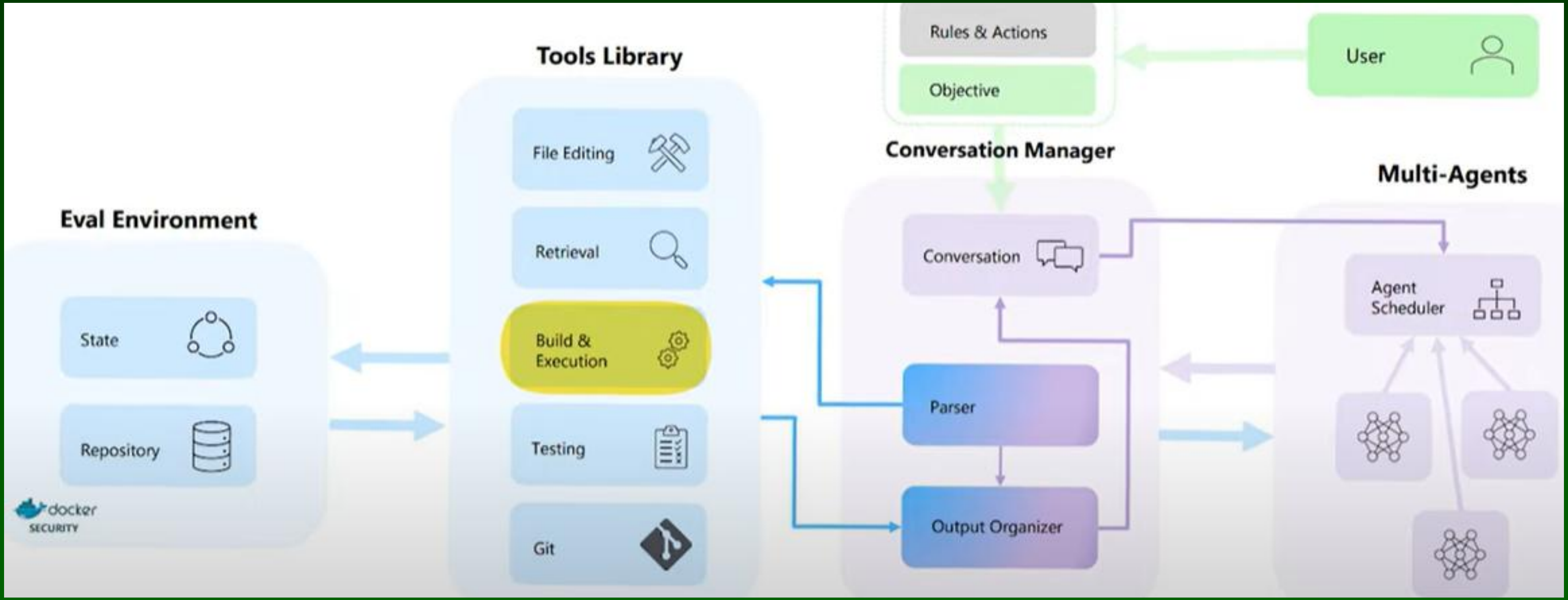
Shocking progress in many fields! Chains and graphs of thoughts work surprisingly well.



Very hard problems from J.D. Jackson's book on classical electrodynamics, that PhD physics students work on for 10 days, GPT4-o1 has solved in 2 minutes!

Microsoft agents are coming

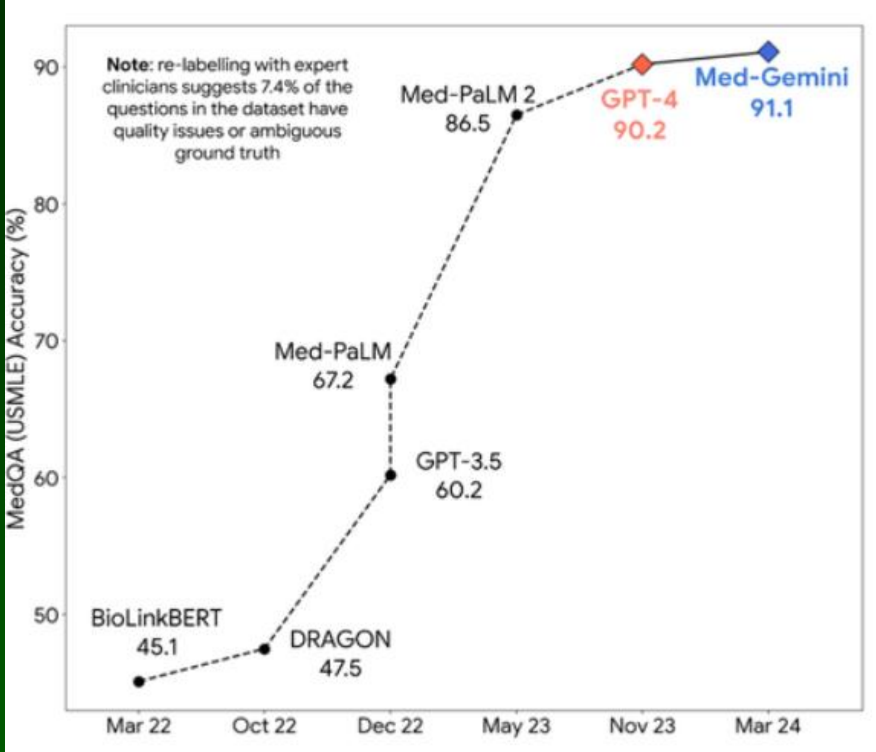
Imagine a team of AI-driven agents working for you … MS AutoDev (3/24), Studio (11/24)
Integrate agents autonomous into software development process.



- Define objectives, agents will perform all actions engaging programmer in dialog with conversation manager overseeing the process and coordinating the actions of AI agents through a combination of rules and actions.
- AutoGen: a general framework for automation of LLM workflows.

Med-Gemini

[MedQA US Medical Licensing Exam (USMLE)-style question benchmark](),
Med-Gemini achieves a state-of-the-art performance of 91.1% accuracy,
surpassing Med-PaLM 2 by 4.6%.
A careful inspection of the MedQA benchmark with expert clinicians found that 7.4% of the questions were deemed unfit for evaluation as they either lacked key information or supported multiple plausible interpretations.



Med-Gemini has multimodal capabilities, application to radiology, pathology, dermatology, ophthalmology, and genomics in healthcare. Large multimodal models can interpret complex 3D scans, answer clinical questions, and generate state-of-the-art radiology reports. We demonstrate a novel mechanism to encode genomic information for risk prediction across a wealth of disease areas.

Hospital Simulacrum

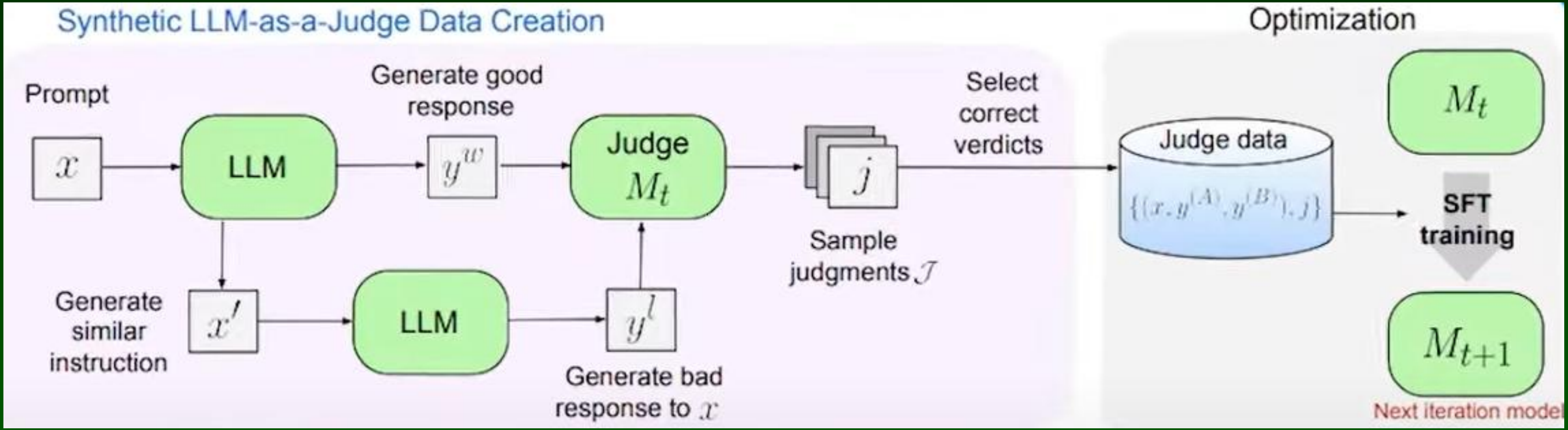
AI bots are gaining experience in virtual world!
A small simulated hospital with 25 GPT-3.5 agents ...
The virtual agent hospital simulates the entire process of treating disease, with autonomous agents as patients, nurses and doctors.

The doctor learns to treat diseases accumulating experience from successful and failures, improving its performance in various tasks.

The knowledge is gained from real medical cases.
After enrolling 10,000 patients, Agent-Doctor reaches a 93% success rate in treating lung diseases on the MedQA database, better than any other system.
Several other simulacra exist, they are examples of self-learning.

Li et al, 5/2024, Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents.

Self-Teaching in LLM



Model-based evaluation as a reward model for training requires human evaluation, but STE does it without human annotations, using 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, repeating this training at each new iteration using the improved predictions.
STE improved Llama3-70B-Instruct from 75.4 to 88.3% (88.7% with majority vote) on RewardBench, outperforming LLM judges such as GPT-4, matching top-performing reward models trained with human-labeled examples.

Wang, T … & Li, X. (2024). *Self-Taught Evaluators* (arXiv:2408.02666)

Science in the new era

1st paradigm	2nd paradigm	3rd paradigm	4th paradigm	
Empirical science	**Theoretical science**	**Computational science**	**Big data-driven science**	**Accelerated discovery**
Observations Experimentation	Scientific laws Physics Biology Chemistry	Simulations Molecular dynamics Mechanistic models	Big data Machine learning Patterns Anomalies Visualzation	Scientific knowledge at scale AI-generated hypotheses Autonomous testing
Pre-Renaissance	**~1600s**	**~1950**	**~2000**	**~2020**

Increasing speed, automation, and scale

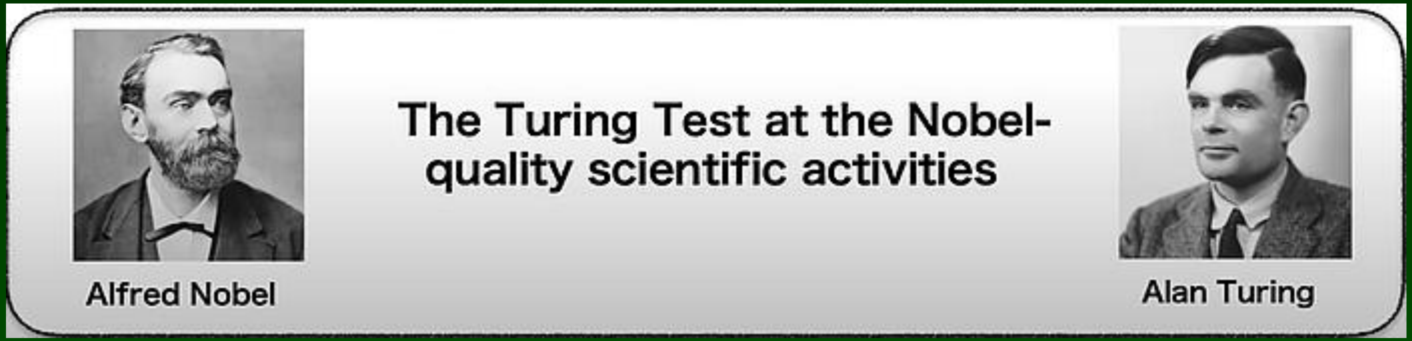
IBM Science and Technology Outlook 2021.

Increasingly complex data models: CyC, IBM Watson, GPT-4, Claude 3, Large Multimodal Models (LMMs) with more than trillion parameters ... and specialized Small MMs.
Large collaborations to create foundation models for science and robotics.

Attacking problems at the irreducible computational complexity level (Wolfram 2002)

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.

Requires in-depth understanding of the process of scientific discoveries, a closed-loop system: knowledge acquisition, hypothesis generation and verification, to full automation of experiments and data analytics.

4th Nobel Turing Challenge Initiative Workshop, Feb 13-14, 2024, Nihonbashi Life Science, Tokyo, **Challenge YouTube channel**.

Nature, 25.05.2023: Deep learning-guided discovery of an antibiotic targeting *Acinetobacter baumannii*

AI4Science

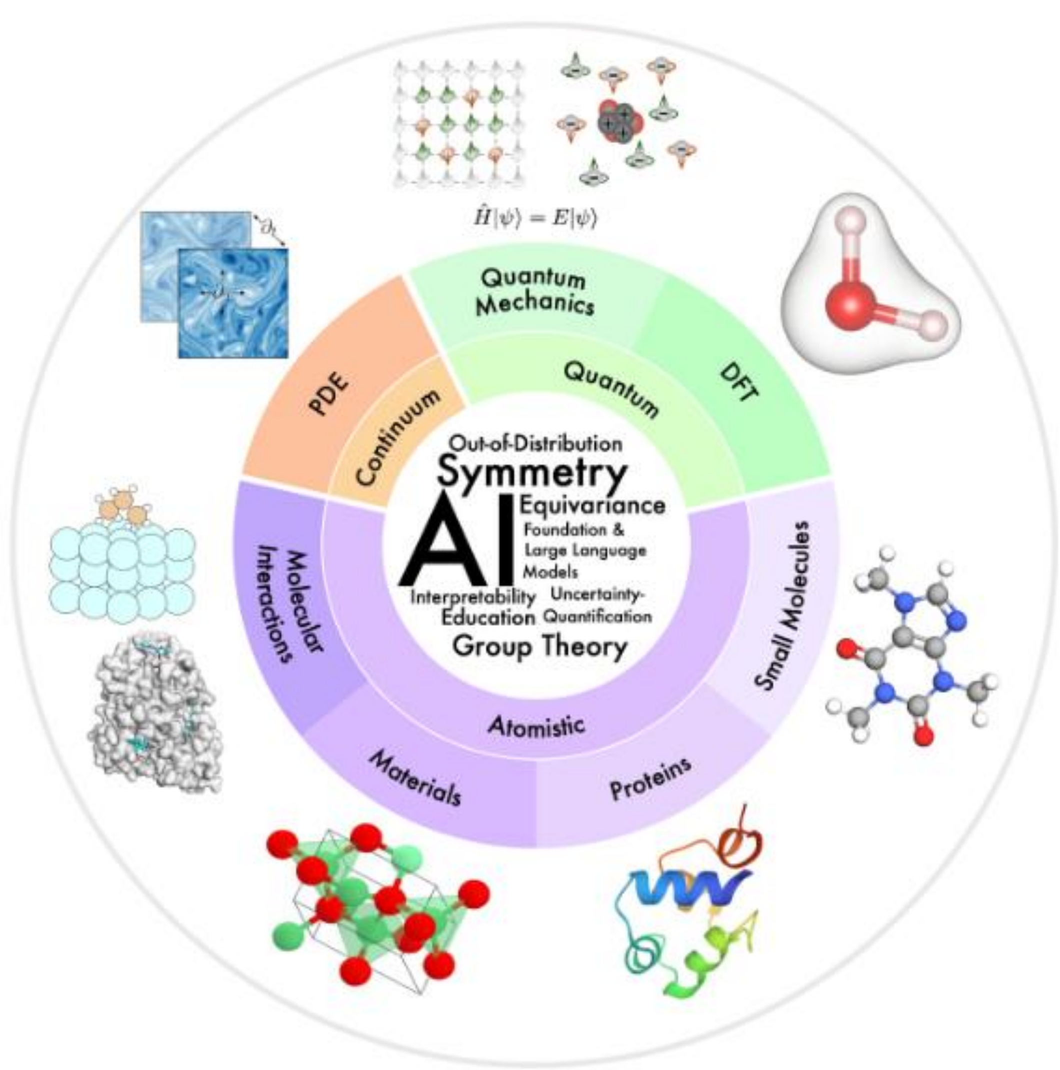
Zhang, X., Wang, L., Helwig, J., … Ji, S. (2023). Artificial Intelligence for Science in Quantum, Atomistic, and Continuum Systems. arXiv:2307.08423

Physics Informed Machine Learning (PIML):
High Level Overview of AI and ML in Science and Engineering. YouTube 2/2024

Engineering systems are governed by physics and involve safety critical components.

We need to embed prior physical knowledge into the machine learning process at each stage.

Physics informed machine learning is critical for many applications that can learn more from sparse and noisy data sets.

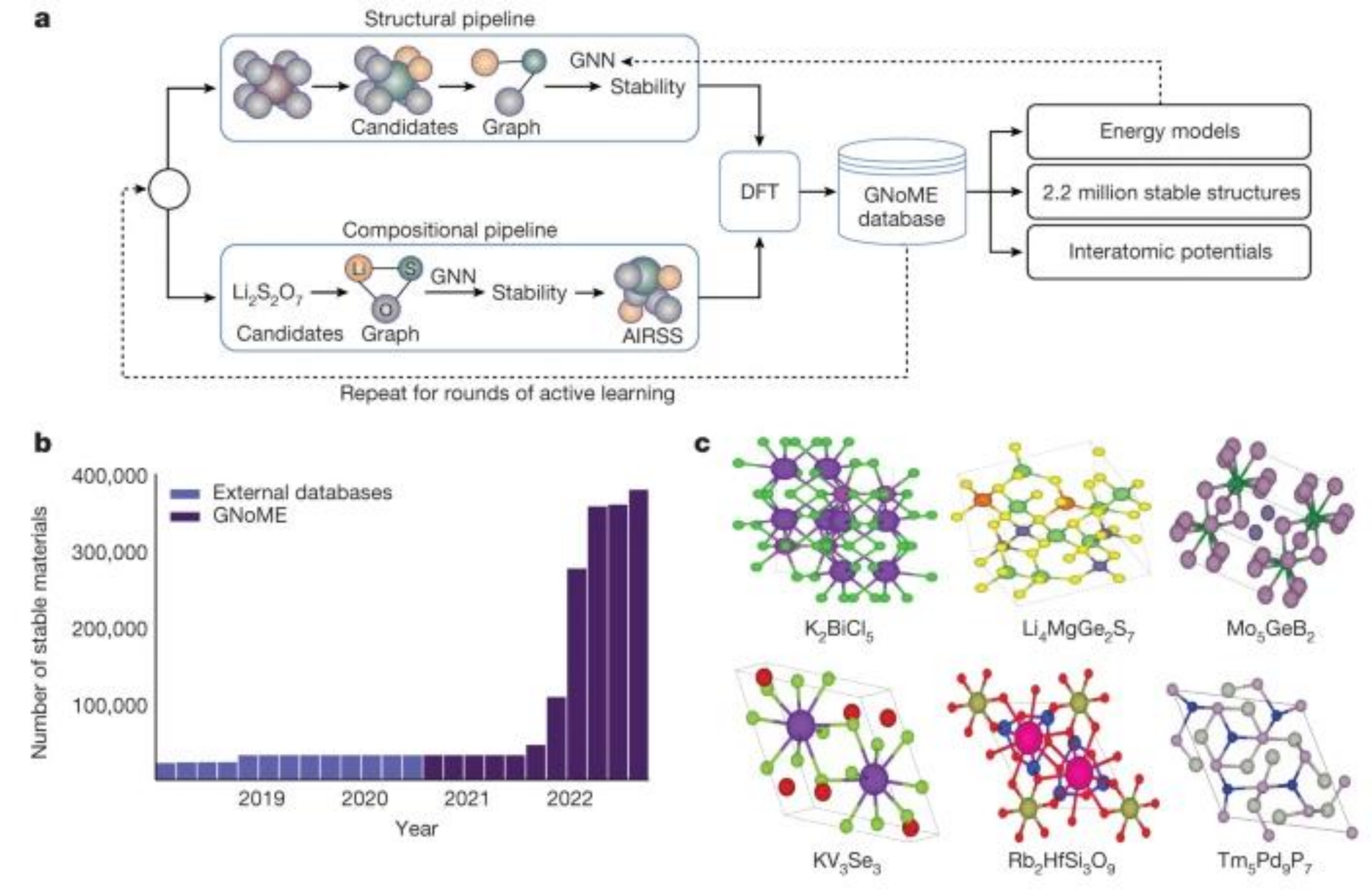


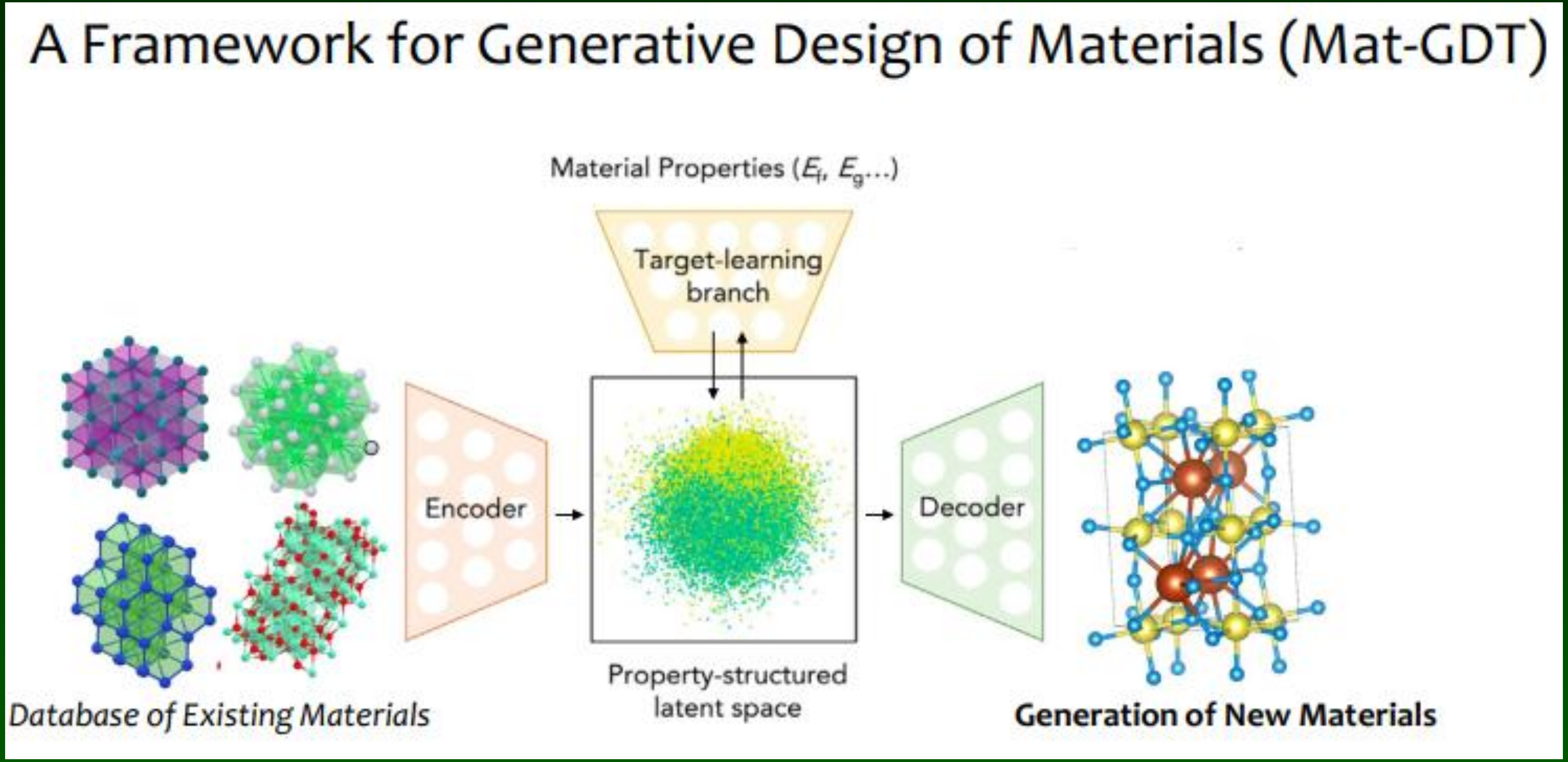
GNoME Materials

Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., & Cubuk, E. D. (2023). Scaling deep learning for materials discovery. *Nature 624*(7990)

GNoME = graph networks for materials exploration, discovered 2.2 mln stable crystals, 381 000 new stable materials, 736 structures already experimentally verified.

Combination of neural networks with quantum chemistry (DFT).





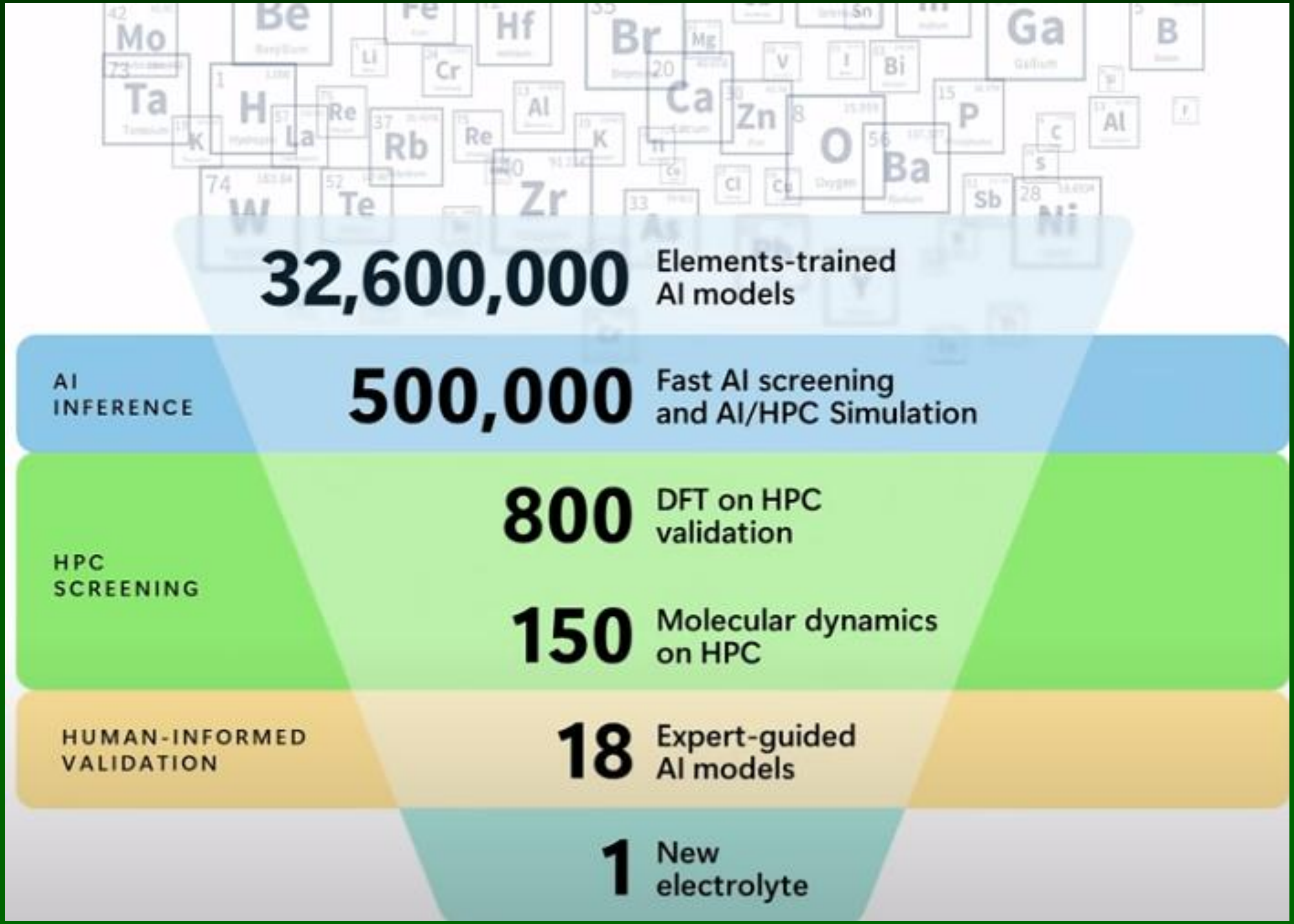
Coscientist (Nature 624, 2023), an AI system driven by GPT-4 that autonomously designs, plans and performs complex experiments by incorporating large language models empowered by tools, internet and documentation search, code execution and experimental automation.

Zhekun Ren, Tonio Buonassisi et. al. ***Cell-Press Matter,*** V5(1), *2022, 314-335*

Ruiming Zhu, Nong Wei et. al. (2023) Accounting for Symmetries of Crystals.

Microsoft designs battery

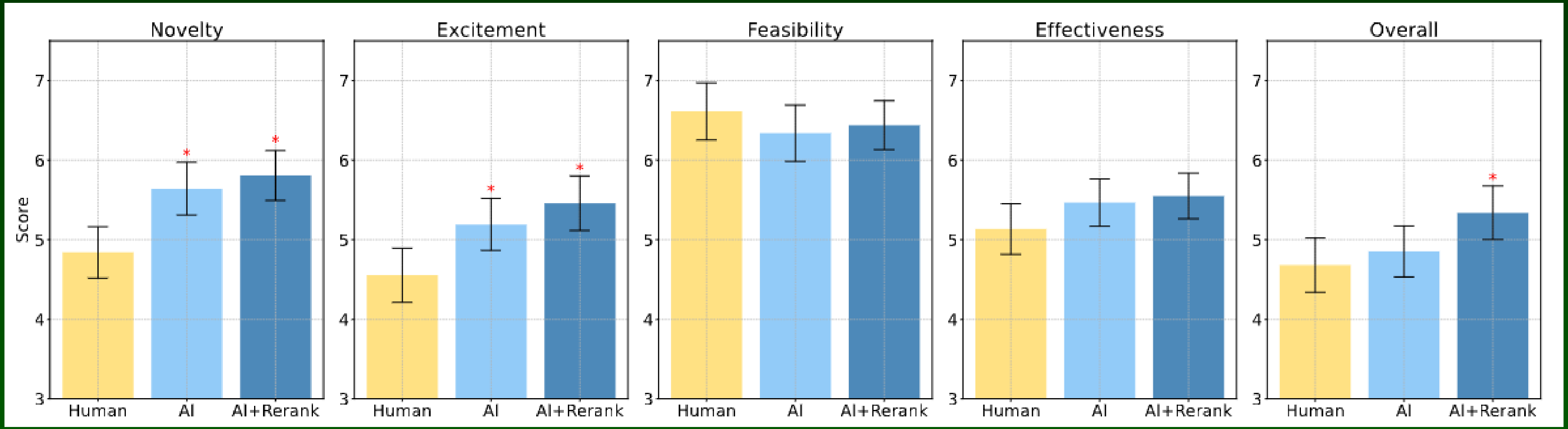
First, compress all relevant information into associative neural network.



Chris Bishop, [The Revolution in Scientific Discovery](). 3/2024

Generating novel ideas

Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. 6.09.2024. 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.

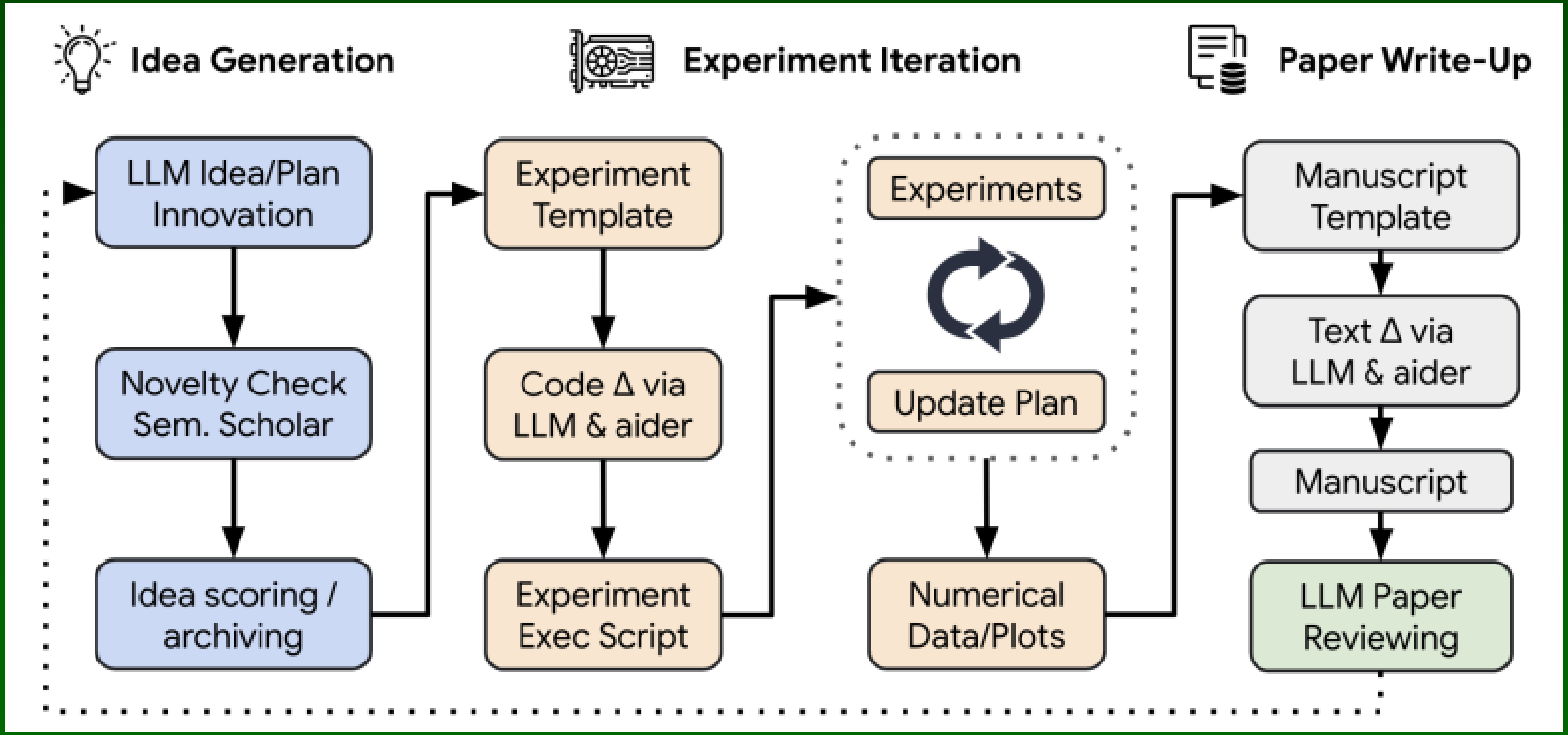
Research Topics: how to reduce social biases, improve code generation, security or privacy, mathematical problem solving, performance on low-resource languages, check factuality, how to reduce hallucination, estimate uncertainty and confidence of LLMs.

AI Scientist

The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. 12.08.2024

The AI Scientist 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.



GPT 4-o1
GPT-4o-mini

AI2ES for environmental science

5 NSF AI ES Institutes were created: Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography. The vision of AI2ES is to create trustworthy AI methods for diverse environmental science (ES) users that will revolutionize our understanding and prediction of high-impact atmospheric and ocean science phenomena and create new educational pathways to develop a more diverse AI and environmental science workforce.

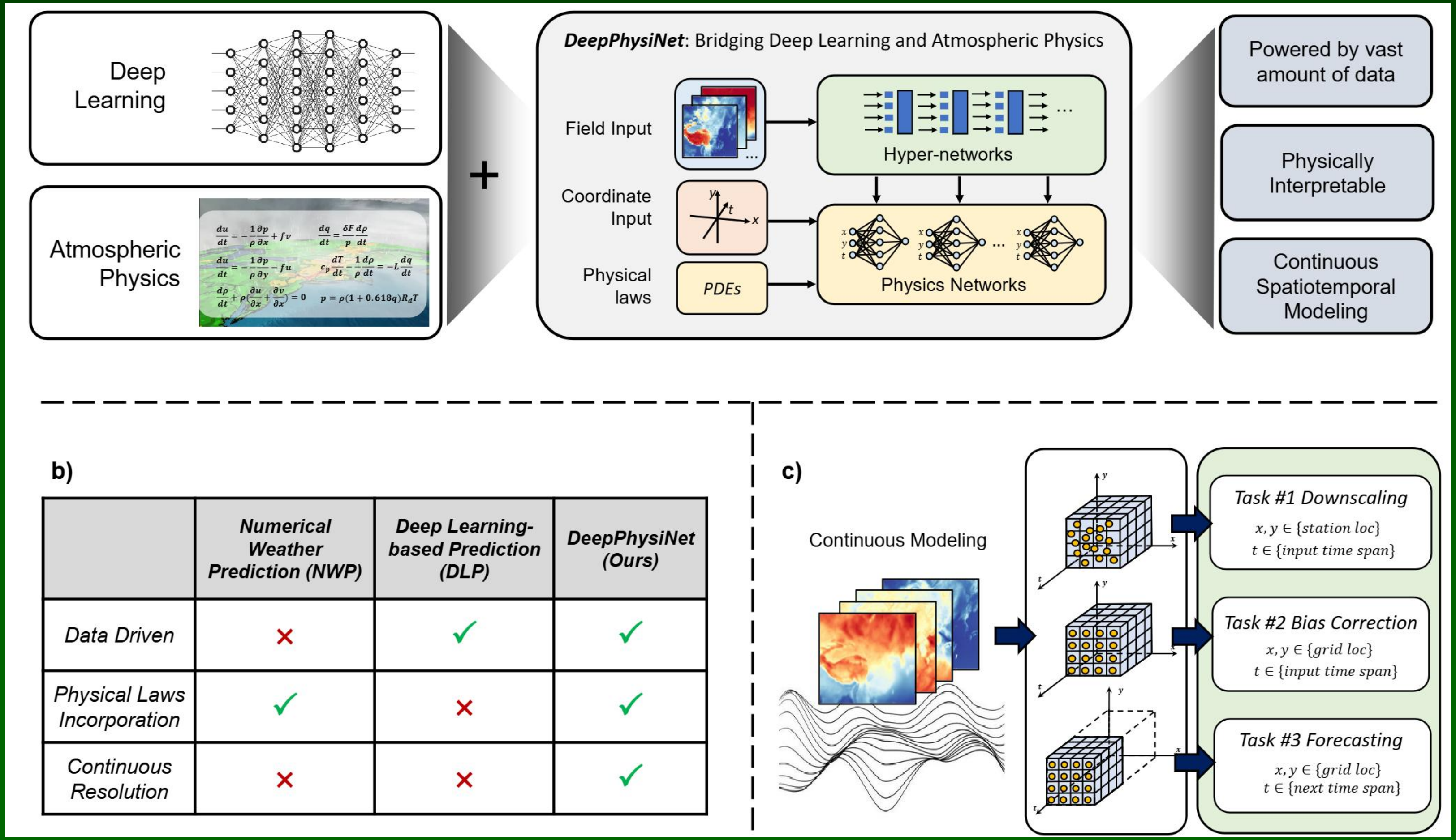
McGovern, A., Tissot, P., & Bostrom, A. (2024). Developing trustworthy AI for weather and climate. *Physics Today*

Weather and PINNs

ClimODE, a spatiotemporal continuous-time process that implements a key principle of advection from statistical mechanics, learning global weather transport as a neural flow, estimating the uncertainty in predictions. An order of magnitude smaller parameterization.

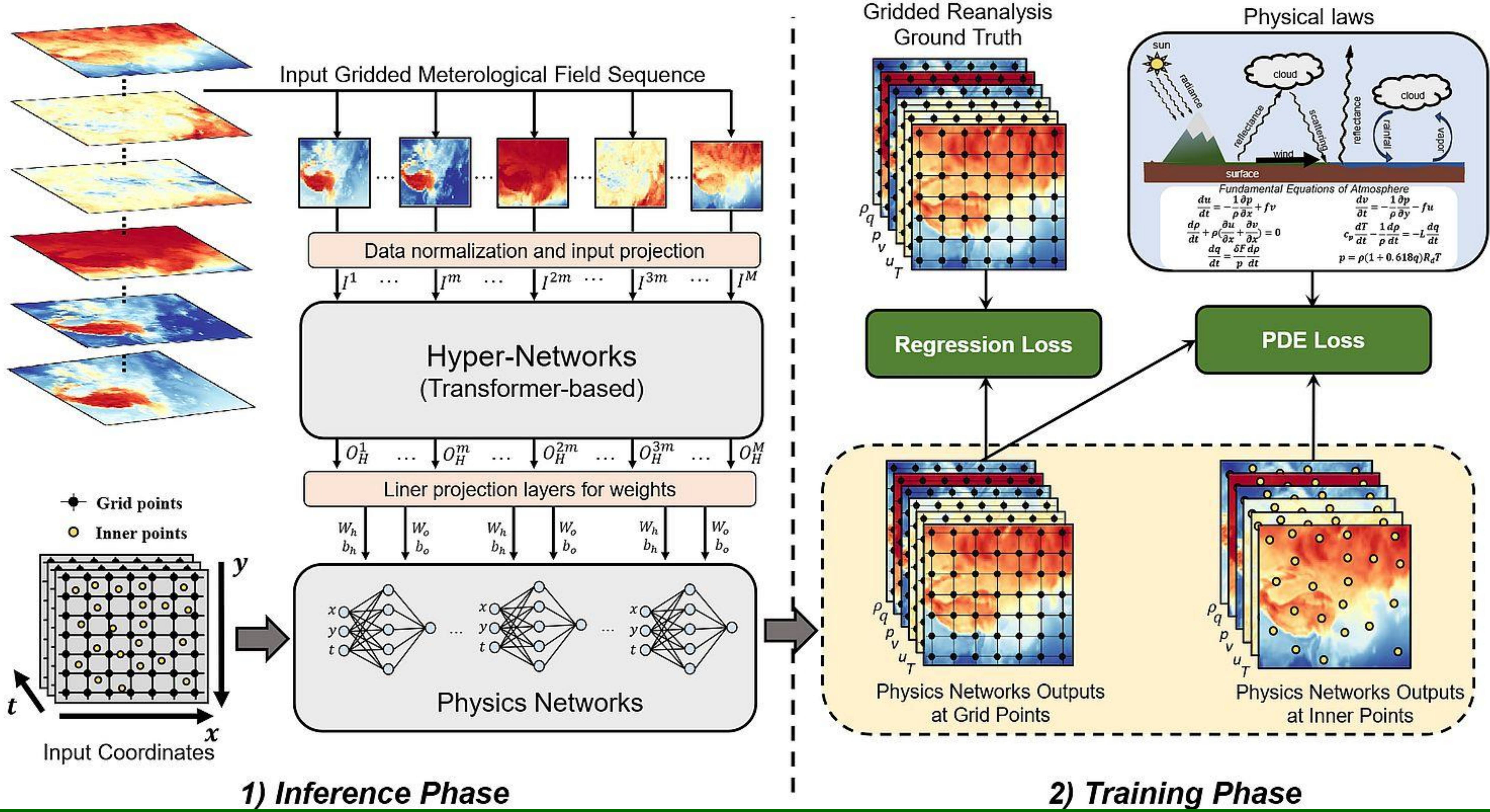
Verma, Y., Heinonen, M., & Garg, V. (4.2024). ClimODE: Climate and Weather Forecasting with Physics-informed Neural ODEs.

DeepPhysiNet framework (2024) The physical laws in the form of Partial Differential Equations can be incorporated as a part of loss function.



	Numerical Weather Prediction (NWP)	Deep Learning-based Prediction (DLP)	DeepPhysiNet (Ours)
Data Driven	✗	✓	✓
Physical Laws Incorporation	✓	✗	✓
Continuous Resolution	✗	✗	✓

Chen, L. … & Yang, Z. (2023). Machine Learning Methods in Weather and Climate Applications: A Survey. Applied Sciences, 13(21).

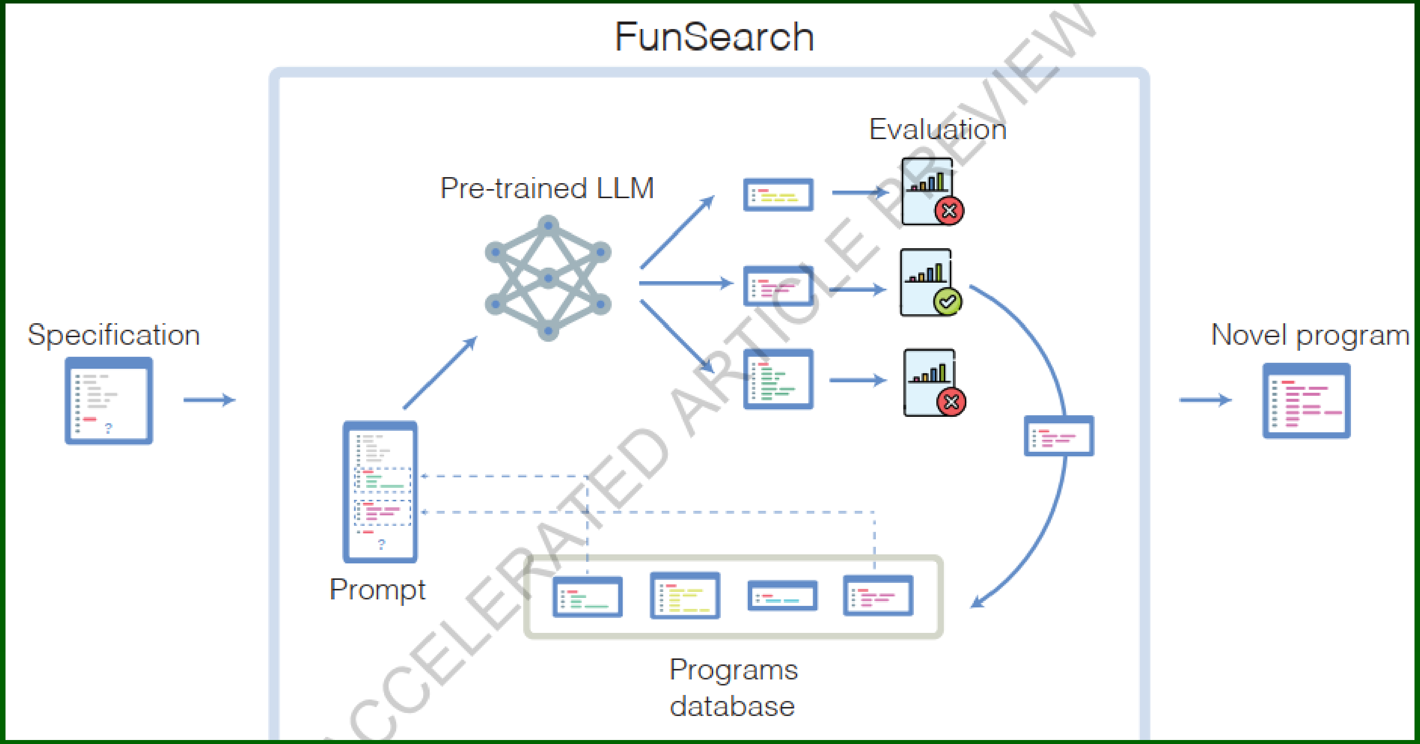


Input Gridded Meterological Field Sequence
Data normalization and input projection
I^1 … I^m … I^{2m} … I^{3m} … I^M
Hyper-Networks
(Transformer-based)
O_H^1 … O_H^m … O_H^{2m} … O_H^{3m} … O_H^M
Liner projection layers for weights
W_h b_h W_o b_o
Grid points
Inner points
Input Coordinates
Physics Networks
1) Inference Phase
Gridded Reanalysis Ground Truth
Physical laws
sun
radiance
cloud
reflectance
scattering
rainfall
vapor
wind
surface
Fundamental Equations of Atmosphere
$\frac{du}{dt} = -\frac{1}{\rho}\frac{\partial p}{\partial x} + fv$
$\frac{dv}{dt} = -\frac{1}{\rho}\frac{\partial p}{\partial y} - fu$
$\frac{d\rho}{dt} + \rho(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial x}) = 0$
$c_p\frac{dT}{dt} - \frac{1}{\rho}\frac{d\rho}{dt} = -L\frac{dq}{dt}$
$\frac{dq}{dt} = \frac{\delta F}{p}\frac{d\rho}{dt}$
$p = \rho(1 + 0.618q)R_dT$
Regression Loss
PDE Loss
Physics Networks Outputs at Grid Points
Physics Networks Outputs at Inner Points
2) Training Phase

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.



Terrence Tao: … it's the equivalent, in terms of serving as [research] assistant. But I do envision a future where you do research through a conversation with a chatbot. Say you have an idea, and the chatbot went with it and filled out all the details. It's already happening in some other areas. … I can see this sort of thing happening in mathematics … instead of spending hours and hours actually trying to make it work, you guide a GPT to do it for you. With o1, you can kind of do this.

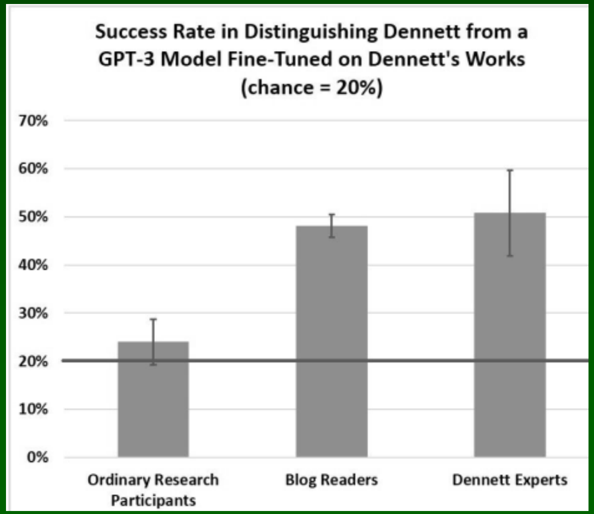
GPT-3 as philosopher

Eric Schwitzgebel, David Schwitzgebel, Anna Strasser,
Creating a Large Language Model of a Philosopher, arXiv:2302.01339

"Can large language models be trained to produce philosophical texts
that are difficult to distinguish from texts produced by human philosophers?"

We asked prof. Dennett ten philosophical questions, posing the same questions to the ChatGPT-3, fine-tuned on his books/papers, collecting 4 responses for each question, without any cherry-picking.

425 participants tried to distinguish Dennett's answer from ChatGPT.
25 experts on Dennett's work succeeded 51% of the time.
Philosophy blog readers (N = 302) performed similarly to the experts.
Ordinary participants (N = 98) were near chance (24%).



So, was Dennett intelligent? Was Kasparov or Ke Jie intelligent?
If we agree, then LLM models are also intelligent.

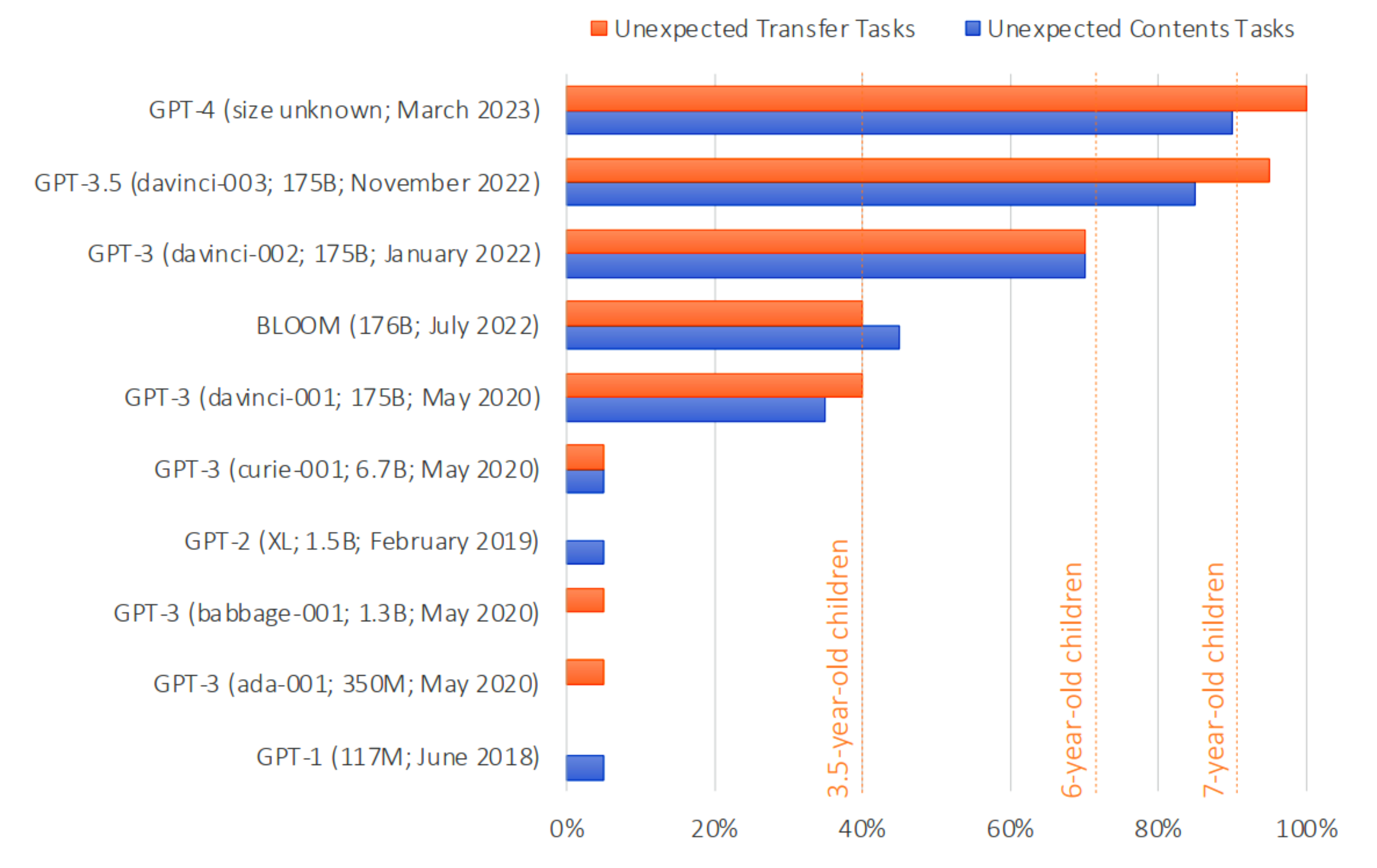
Duch W. (2023), Artificial intelligence and the limits of the humanities.
Er(r)go 47 (2/2023) - Humanities.

AI – new intelligent digital being?

Tests for theory of mind

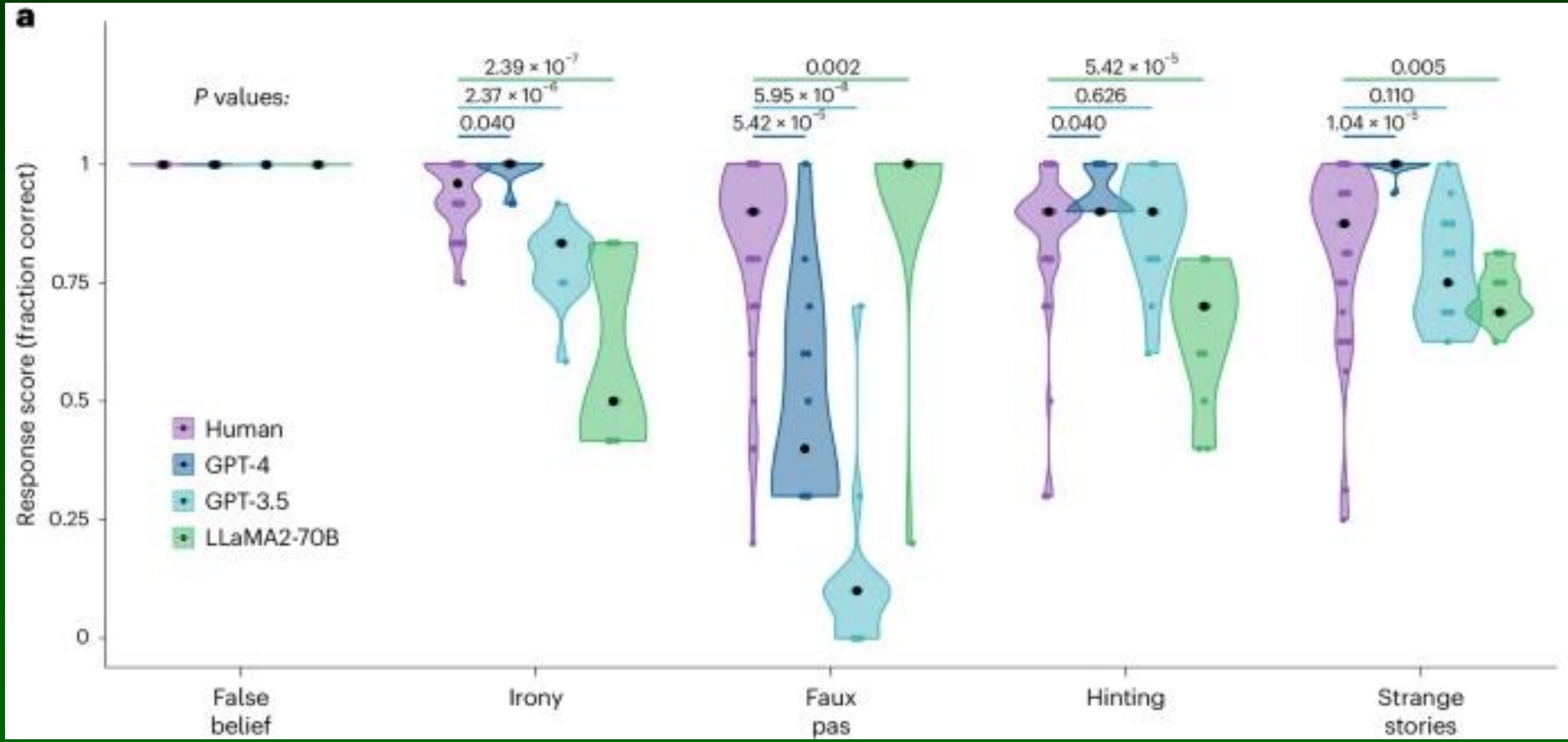
GPT-4 passes most of them. Kosinski, M. (2023). *Theory of Mind May Have Spontaneously Emerged in Large Language Models* (arXiv:2302.02083).

How is this possible?
Pyers, J.E, Senghas A. (2009) Language learning, over and above social experience, drives the development of a mature theory of mind. Psychological Science, 20(7), 805–812.



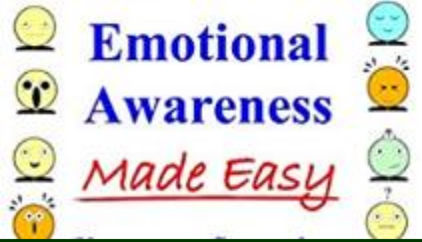
Testing theory of mind

Testing various aspects of ToM understanding: false believes, irony, faux pas (inappropriate in context), hinting (intended meaning of the remark and the action that it is attempting to elicit), and strange stories (explain why a character says or does something that is not literally true).



Strachan, J. W. A…. & Becchio, C. (2024). Testing theory of mind in large language models and humans. [*Nature Human Behaviour, 1–11*].

Emotion awareness

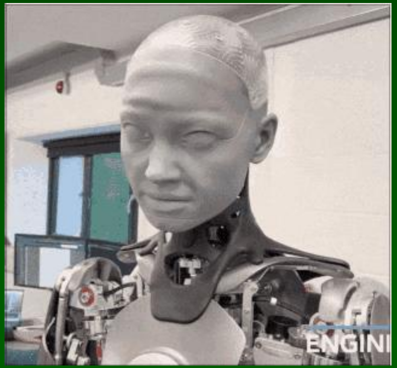


Can LMM agents understand our psychology?

Emotional awareness (EA): the ability to conceptualize one's own and others' emotions, transdiagnostic for psychopathology.

Levels of Emotional Awareness Scale (LEAS) was used to analyze ChatGPT's responses (explanations of human feelings) in 20 scenarios, and compared them with population norms.

ChatGPT demonstrated significantly higher performance than average human.
ChatGPT's performance accuracy levels were extremely high (9.7/10).
ChatGPT emotional awareness grows with time.



No embodiment was needed! Learning language is sufficient?

Implications: ChatGPT can be used as part of cognitive training for clinical populations with EA impairments.

Elyoseph, Z, Hadar-Shoval, D, Asraf, K, & Lvovsky, M. (2023).
ChatGPT outperforms humans in emotional awareness evaluations. *Frontiers in Psychology, 14*

Many experts on science of consciousness claim that machine consciousness is inevitable.

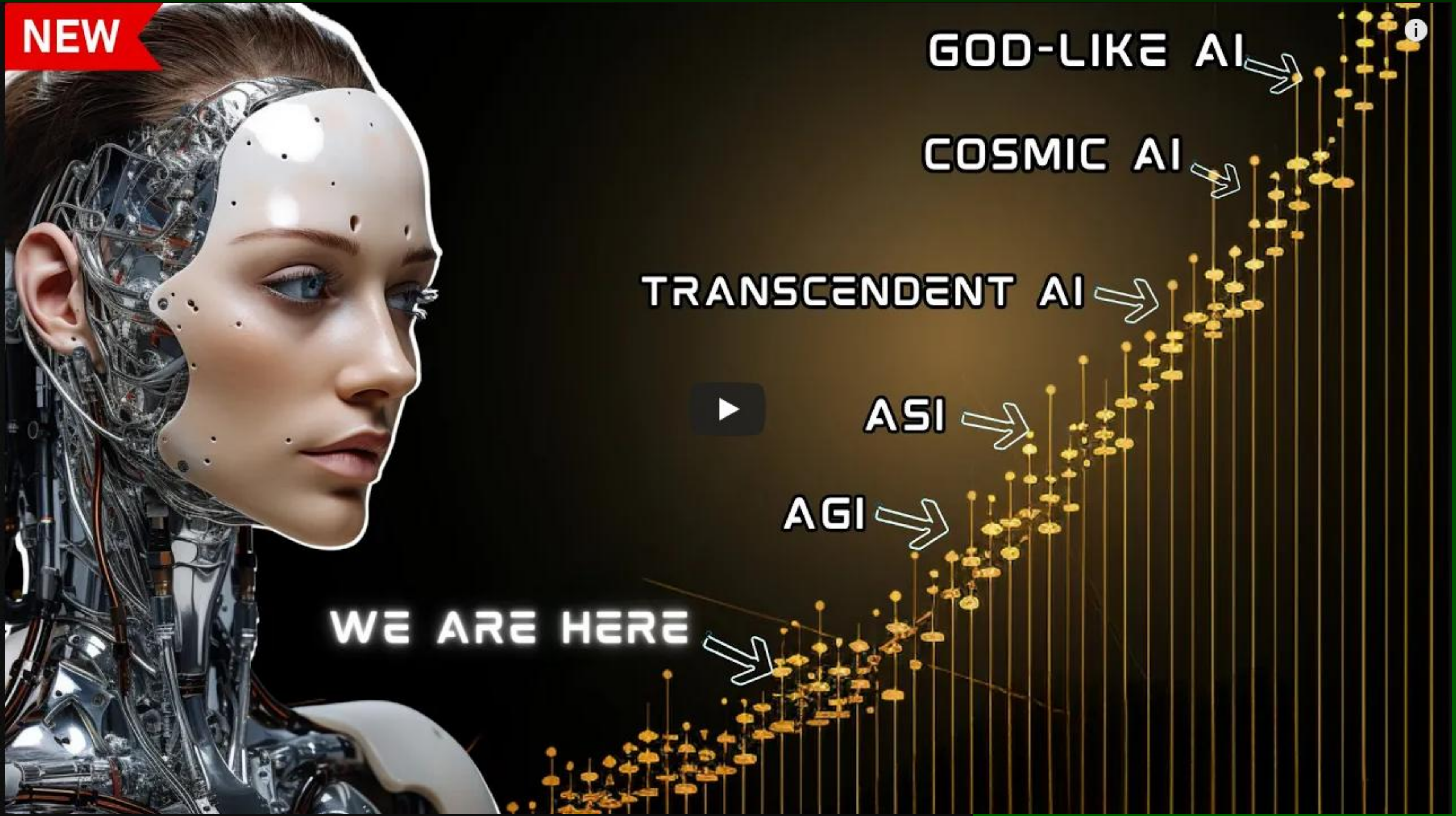
How should we look at AI?

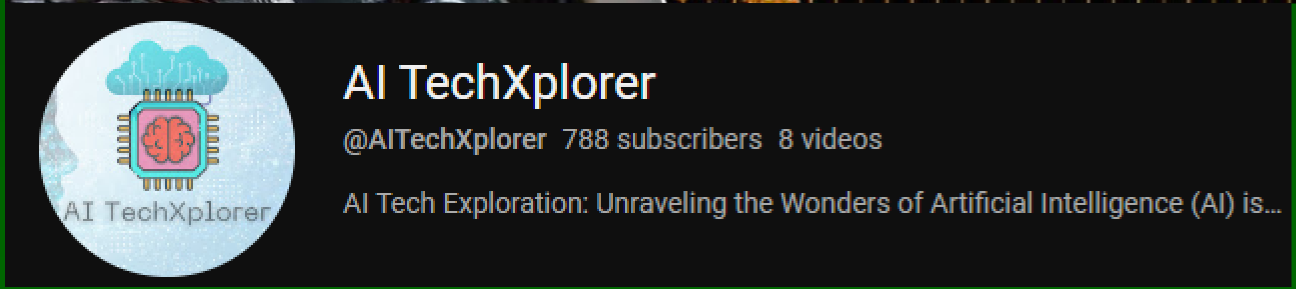
AI is a stochastic parrot, predicting next word in the sentence, but shows no understanding, hallucinates and confabulates.   Human megalomania?

We have left statistical learning behind … LLMs are not programs! They are reservoirs of knowledge! A new superhuman intelligent digital form of being. Expect **radical changes**.

Perspectives

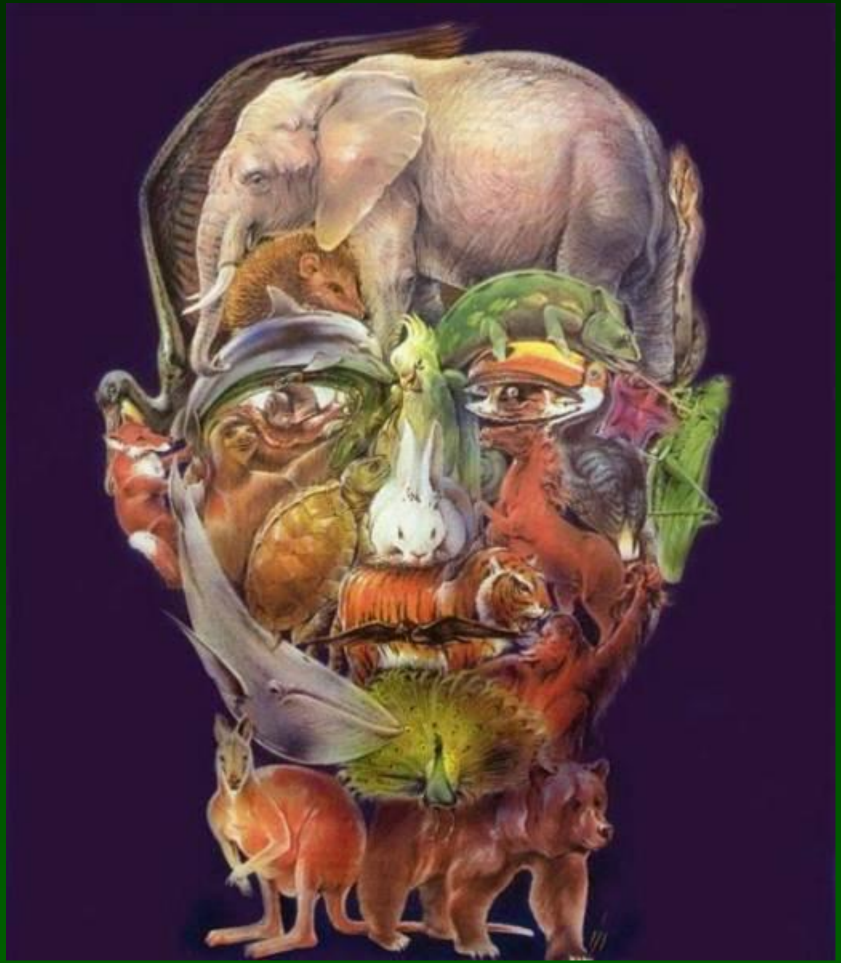
- Neural network research has deep roots in statistical physics and led to associative memories, predictive models, and generative neural models.
- Complex systems show irreducible complexity, but nature has done computing for us, and networks can compress data into internal models.
- Machine Learning has shown its great potential and is not slowing down.
  GPT4-o1, Claude 3, Nvidia NIMs, Physics-Informed Neural Networks are new milestones.
- Multi-agent systems can learn in virtual world using real data (hospital simulacra), are capable of self-teaching. They can help in many ways, including reasoning, inventing new ideas.
- AI systems are quickly becoming partners in many branches of research, from mathematics to philosophy, although we need models that specialize, not omniscience general models.
- LLMs/LMMs work with plugins, use software tools and auto-prompts, agents become to some degree autonomous, may perform long chains of complex actions.
- AI is used at every stage of drug research, for target identification, selection, prioritization, to clinical research and automatization of experimental work.
- Big companies are at the front of AI research, but large open projects may be used for experimentation in new domains (Llama, Falcon-40B).





AGI = Artificial General Intelligence
ASI = Artificial Super-Intellignece

Artificial ?



Search: Wlodzislaw Duch
=> talks, papers, lectures, Flipboard, YouTube