AI changes everything 5/23

Włodzisław Duch

Neurocognitive Laboratory, Center for Modern Interdisciplinary Technologies,
Department of Informatics, Institute of Engineering and Technology,
Faculty of Physics, Astronomy & Informatics, Nicolaus Copernicus University,
Toruń, Poland

Search: Wlodzislaw Duch

Thursday Colloquium, FPAI, UMK 1.06.2023

AI at the warp speed

1. Recently: superhuman AI results in many applications.
2. Yesterday: transformers, foundational models, language/vision.
3. Today: tools for Artificial General Intelligence.
4. Emergence and sparks of AGI in GPT-4.
5. State-of-the-art and beyond.
6. AI minds and human brains.

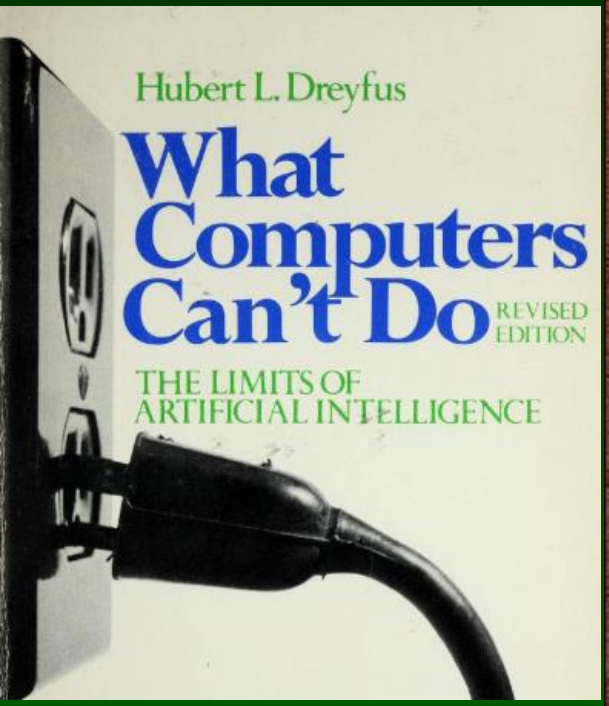
News in my YouTube ML library, and in my Flipboard.

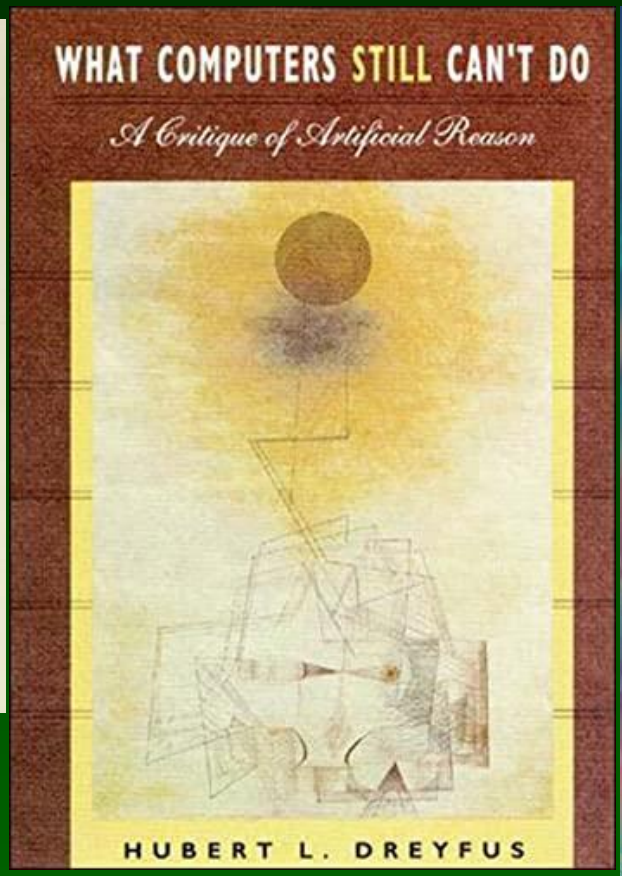
Superhuman AI?

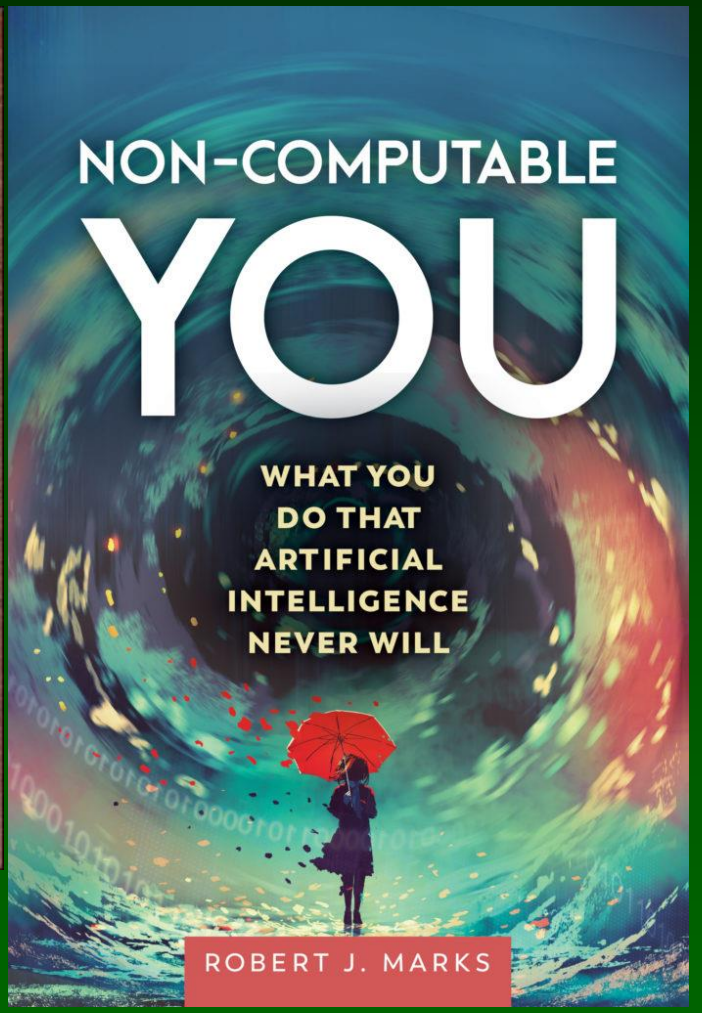
New AI: predicts words, but shows no understanding, requires small adjustments,
or: new competitive form of intelligence, will lead to **radical changes**.

KEEP
CALM
IT'S
BUSINESS
AS USUAL

Human megalomania







1972, 1992, and 2022, predicting all kinds of problems for AI.
Human megalomania? This is what we want to hear, but reality is different.

Are we close to the Singularity?

What is coming:

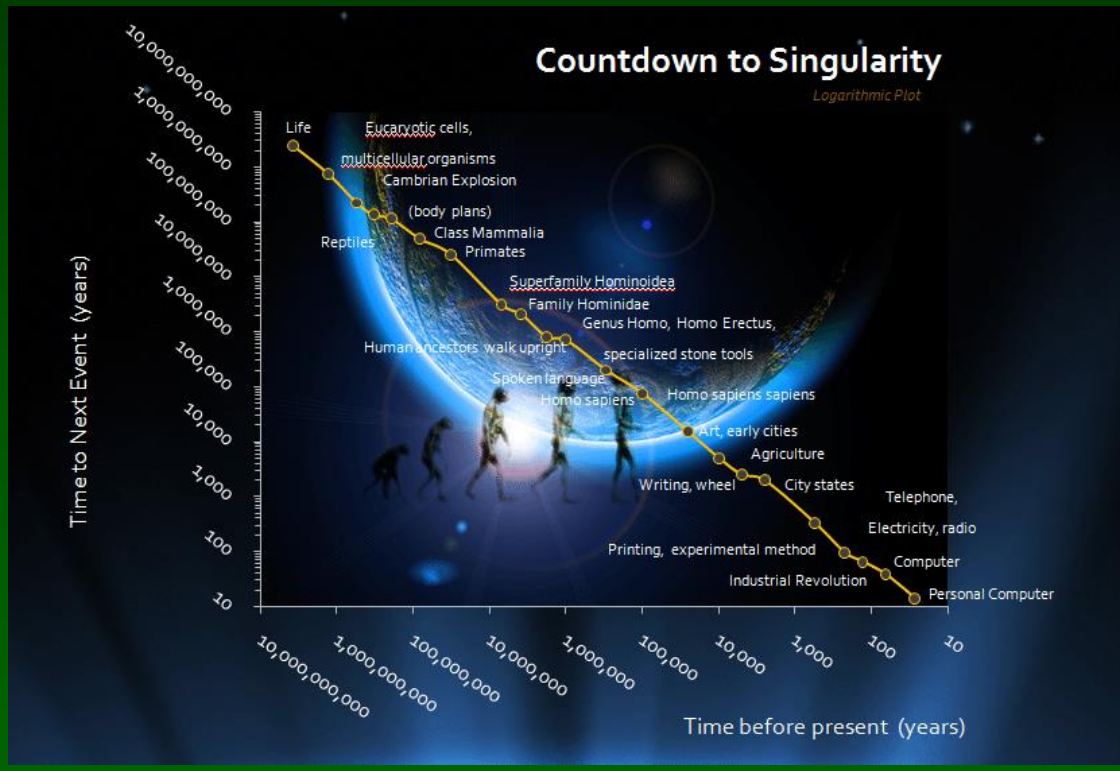
1. Autonomous AI.
2. Superhuman level.
3. Brain-computer interfaces for human augmentation.
4. Neurotechnologies to restructure our brains.

Are we on the threshold of a pleasant dream, or a nightmare?

Is transhuman society around the corner?

Many opinions, few arguments.

Each new technological revolution comes faster than the previous one.



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-Starcraft II,
2022 Stratego, Diplomacy – what is left?

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

Robotics: 2020 Atlas robot (Boston Dynamics) backflip
and parkour, autonomous vehicles, airplanes.

Automation of science: 2015-AI flatworm regeneration,
2020 AlphaFold 2, now 620 M protein 3D structures.

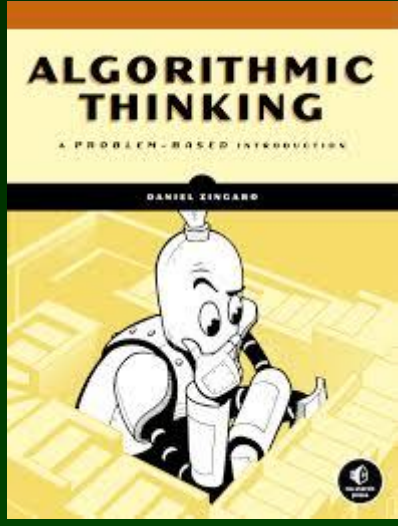
Creativity and imagination: GAN revolution, Dall-E2,
Midjourney, Stable Diffusion, AIVA, music composers.

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

Cyborgization: BCI, brain-computer symbiosis, soon?

What are humans better in?

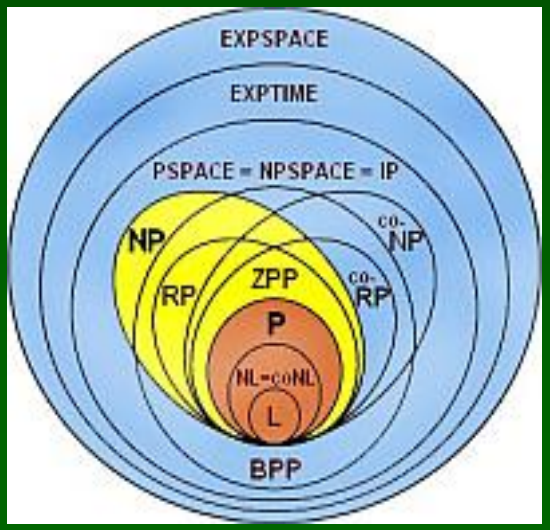
AI: computer science definition



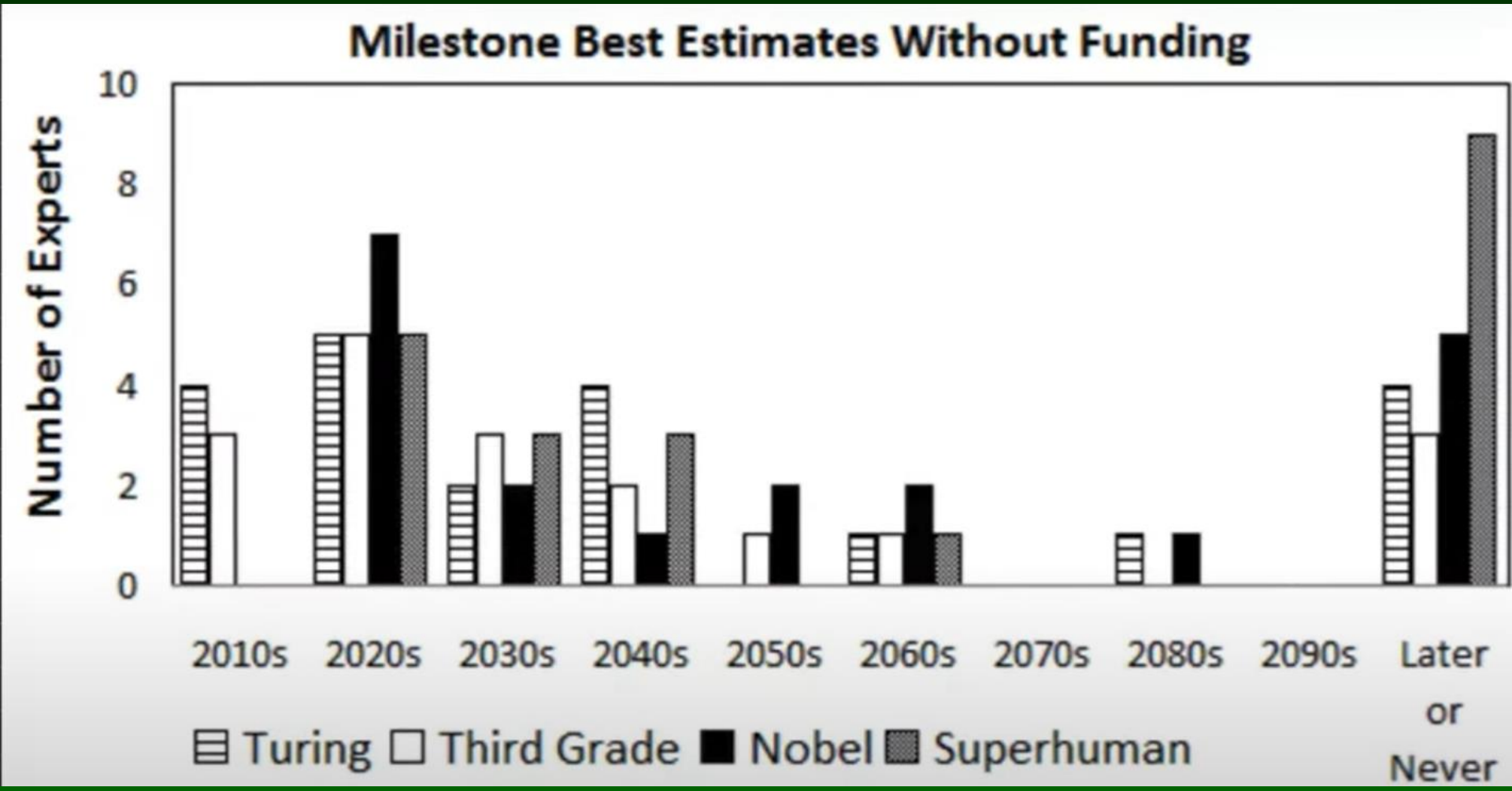
What have all these applications in common?

Artificial Intelligence is a branch of computer science solving problems for which there are **no effective algorithms**.

Either algorithms are so complex that they will never finish, or we do not know how to formulated the problem, ex. understanding images, literature, or brain states.



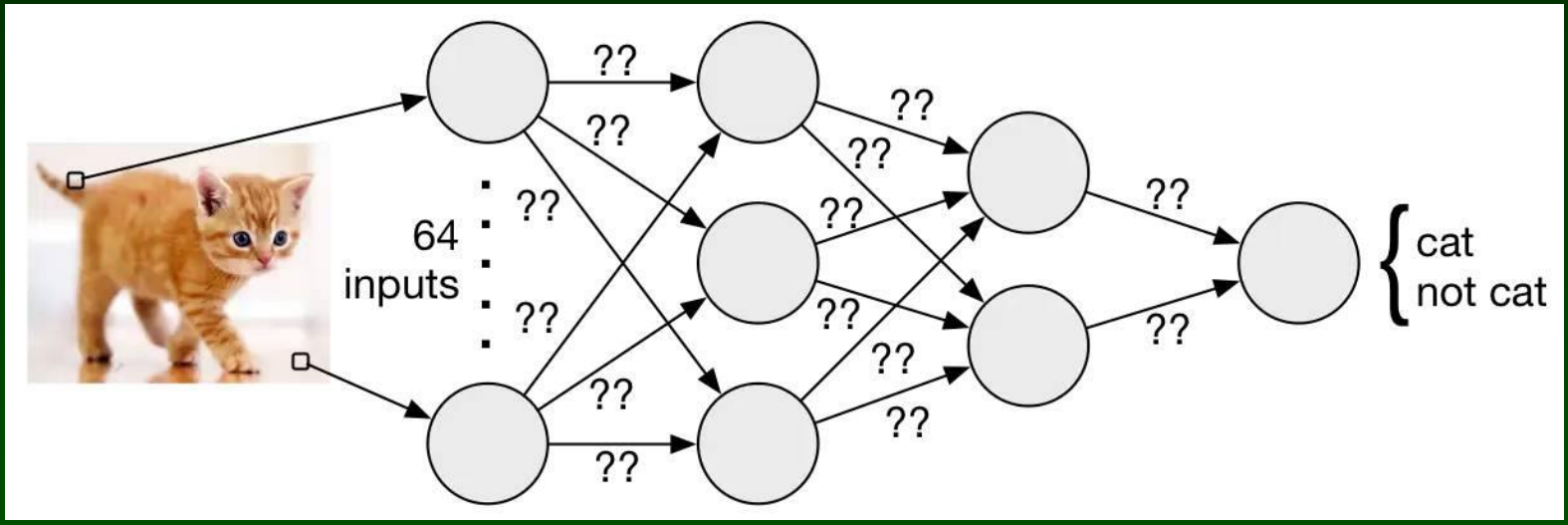
Artificial General Intelligence



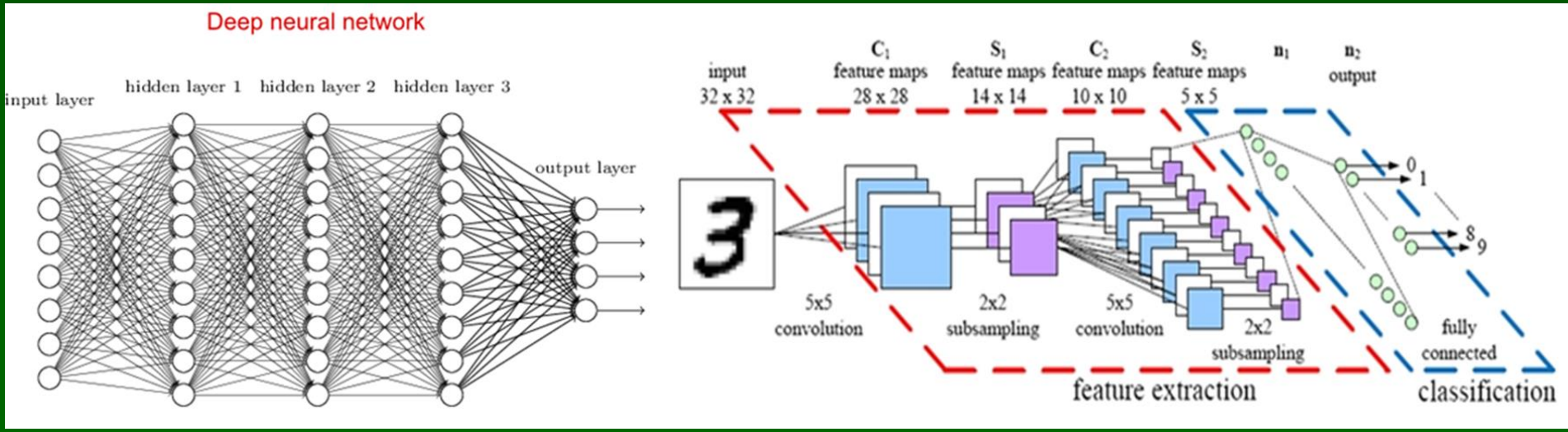
In 2009 many experts thought that superhuman level will never be reached.
AGI-09 Survey – AI Impacts But what should AGI do has not been defined.

Yesterday.
Deep neural networks and transformers

Neural classifiers

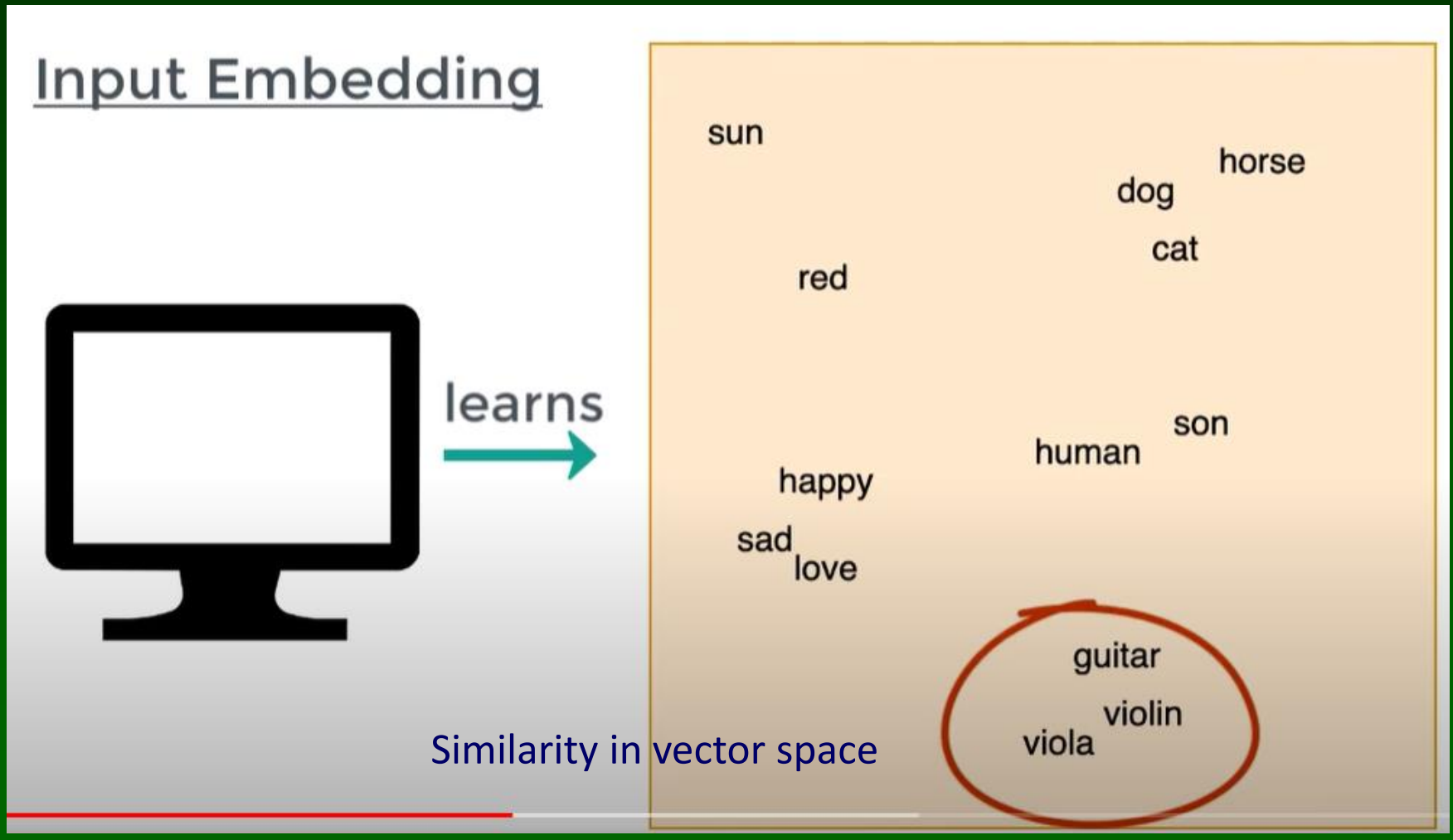


Words, image patches => networks with adjustable parameters
=> training to recognize patterns => object classification, diagnosis.

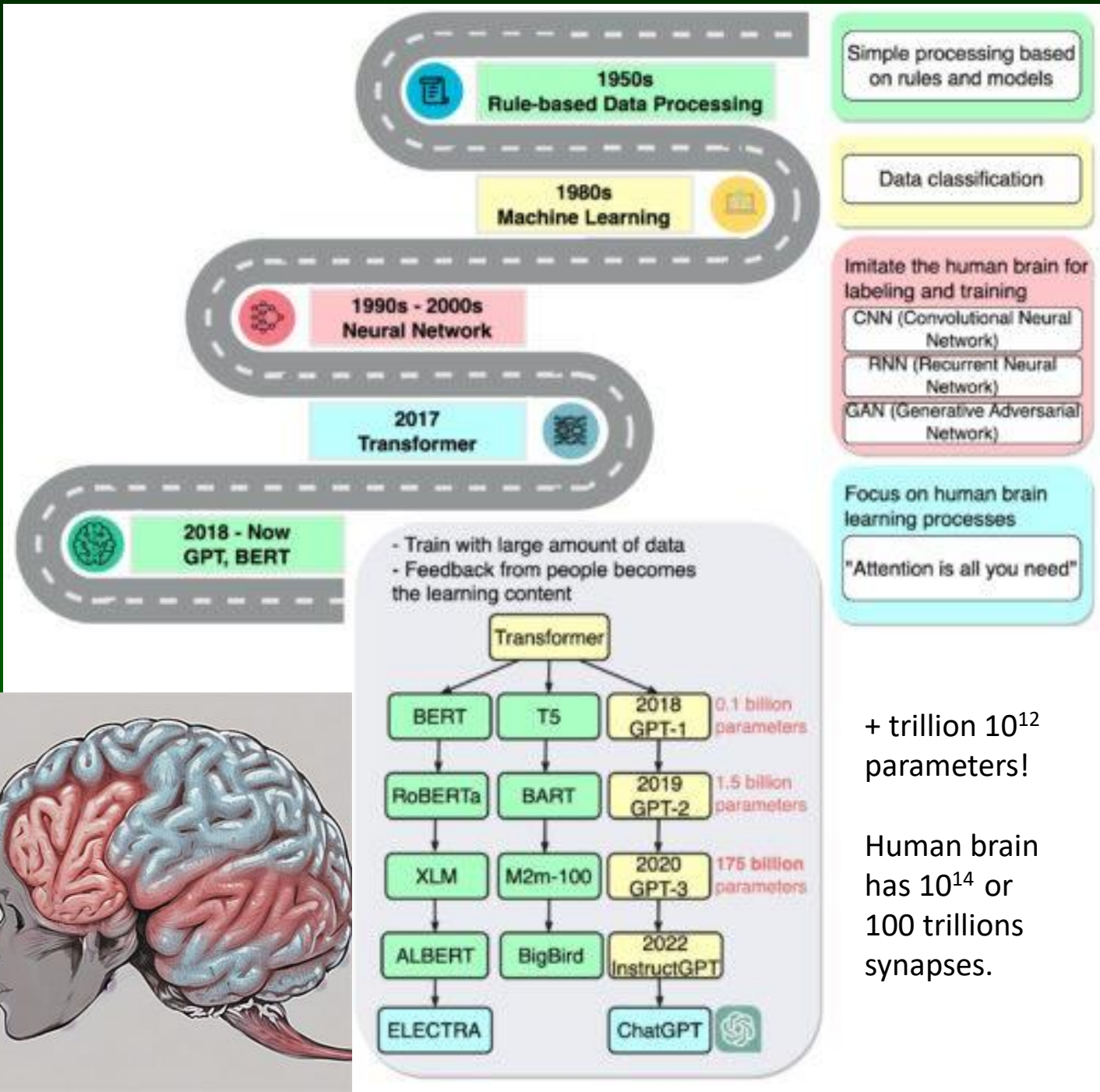


Embeddings

Words => vectors, reflecting their similarity and positions in sentences.



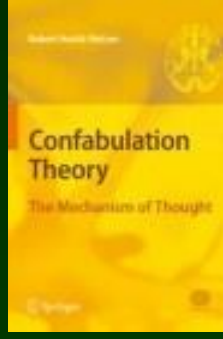
LLM timeline



+ trillion 10^{12} parameters!

Human brain has 10^{14} or 100 trillions synapses.

Language algorithms



Language models: relation of words in complex network structures.
In 2018, to gain a general-purpose “language understanding”,
Google created BERT, model pre-trained on a very large text corpus.

- **Bidirectional Encoder Representations from Transformers** (**BERT**).
  Transformer-based machine learning technique for (NLP) pre-training.
- English-language BERT: two networks, smaller 110M parameters,
  larger model with 340M parameters in 24-layers; trained on the
  BooksCorpus with 800M words, and Wikipedia with 2,500M words.
  In 2019 BERT worked already in 70 languages.
- BERT model was then fine-tuned for specific NLP tasks such as question answering or semantic information retrieval. Many smaller pre-trained open software models were published in GitHub repository.
- The network learns to predict masked tokens (words, images, signals):
  Input: the man went to the [MASK1]. He bought a [MASK2] of milk.
  Labels: [MASK1] = store; [MASK2] = gallon.
  As in Hecht-Nielsen, Confabulation Theory (2007).

Transformers

Vaswani et al.(2017). *Attention Is All You Need.* arXiv

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

Attention vectors capture contextual relations between words in sentences.

Example:

Input: sentence in English;
Output: sentence in Polish language.

Google BERT has used this.

Generative Pre-trained Transformers or GPTs are now best known models**.**
Simple intro on Youtube.

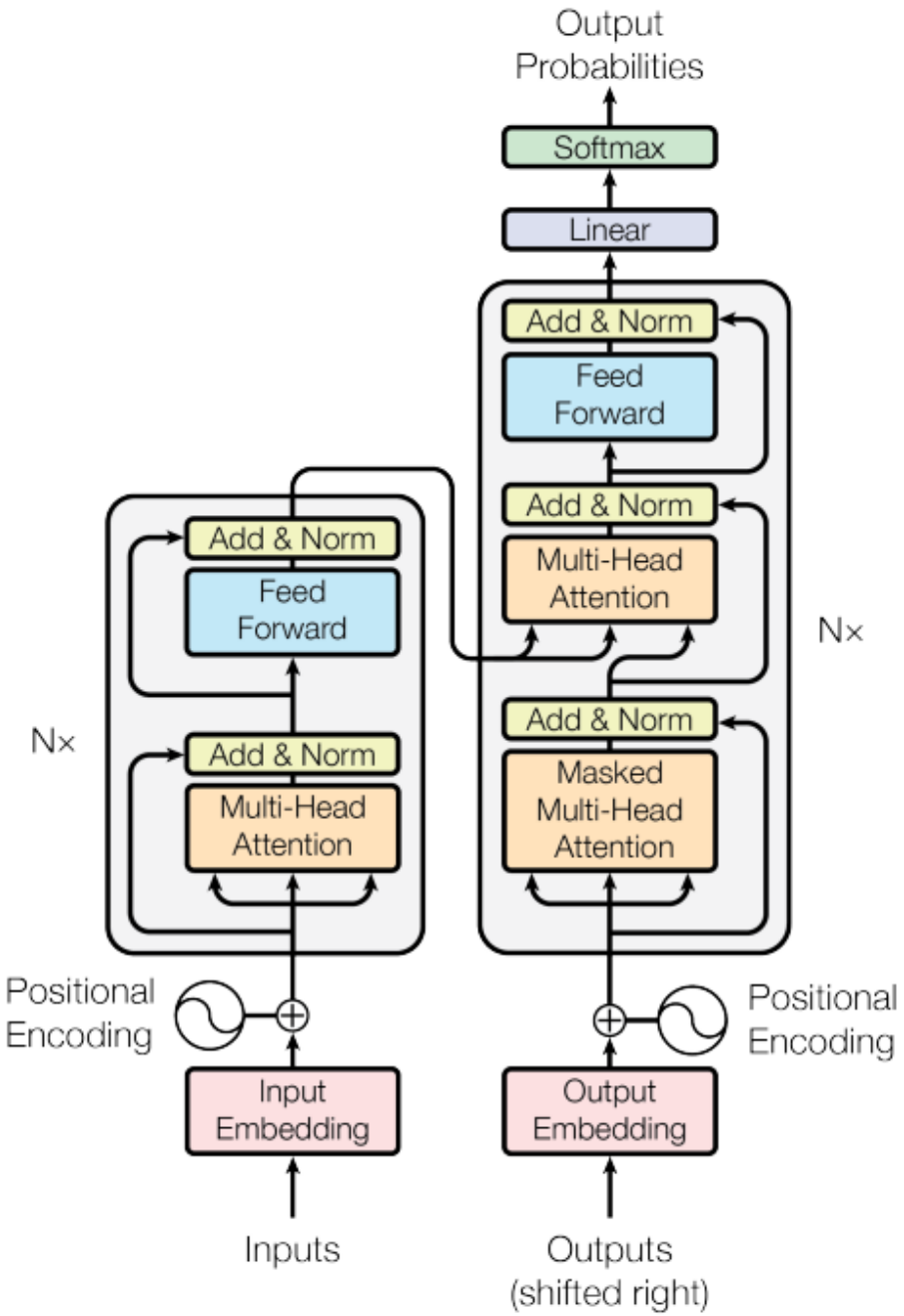
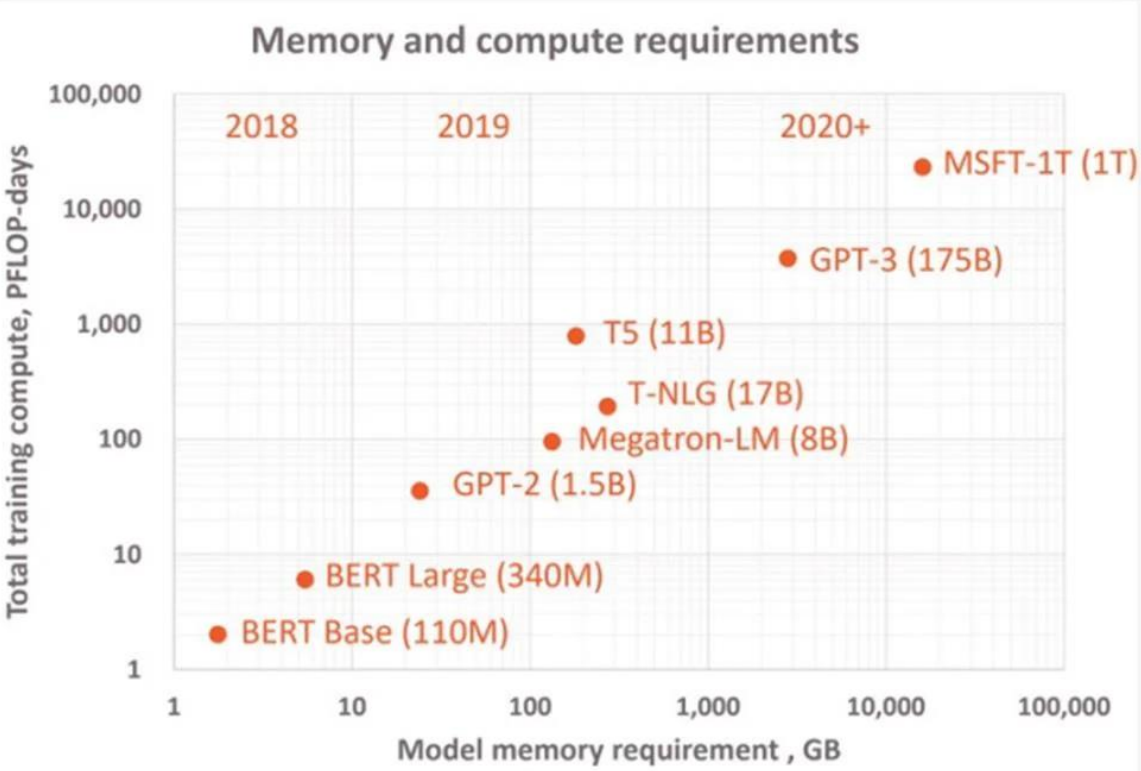


Figure 1: The Transformer - model architecture.

Acceleration …

Exponential Growth of Neural Networks



1000x **larger models**

1000x **more compute**

In just **2 years**

Today, GPT-3 with 175 billion params trained on 1024 GPUs for 4 months. OpenAI

MS+Nvidia MLM, MSFT models

Tomorrow, **multi-Trillion** parameter models and beyond.

Nanotechnologies for ML training: in 10 years from 1 Pflop to 1 billion Petaflops!

Large models provide foundations for general knowledge. Gato has only 1.2 B parameters. Smaller LLM may also be efficient (mBERT, LLaMa, Falcon).

NLP supermodels

OpenAI GPT-3 model has 175 B parameters! One can use it on OpenAI server. First-of-its-kind API can be applied to any language task, and serves millions of production requests each day. GPT-4 has 170 trillion parameters?

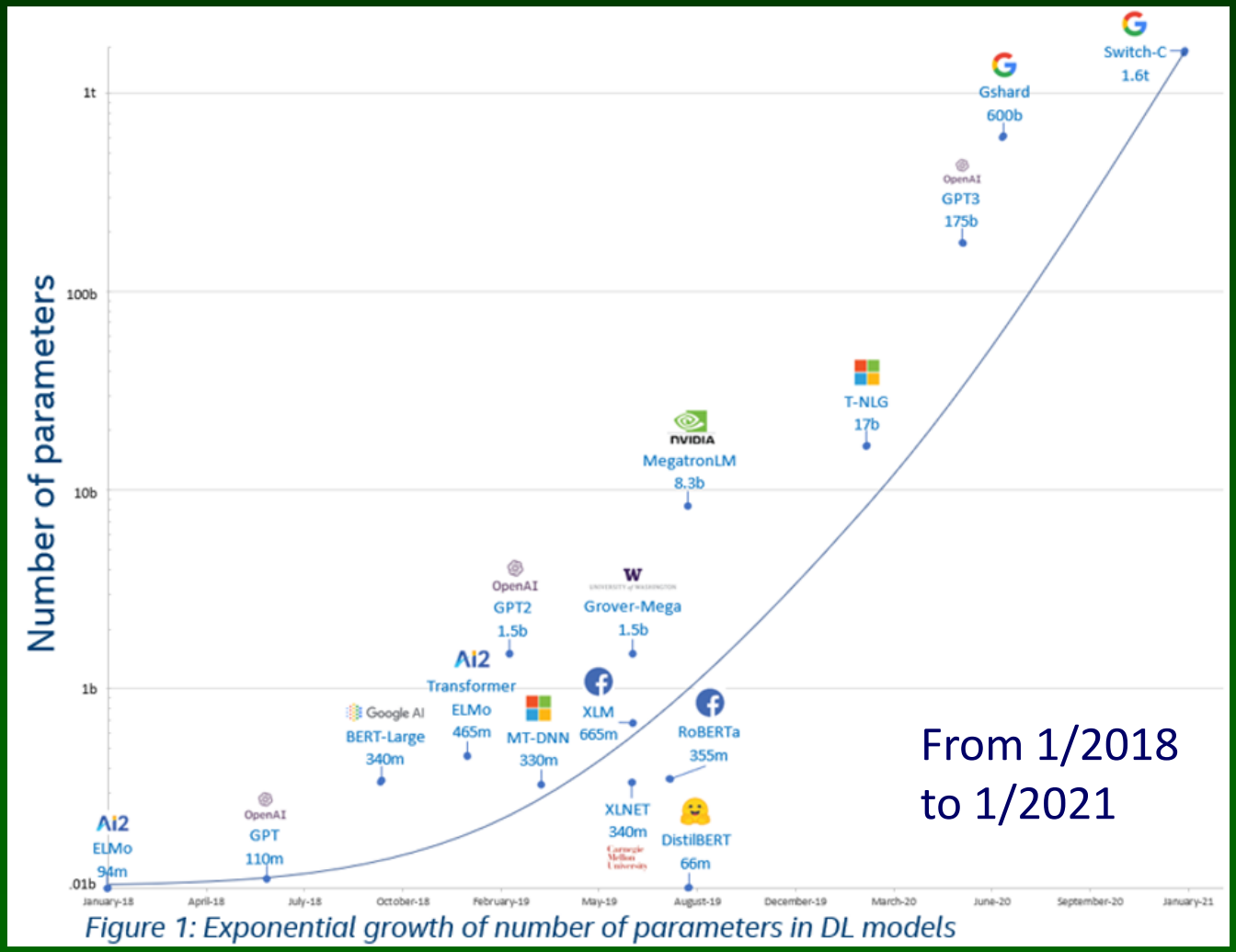


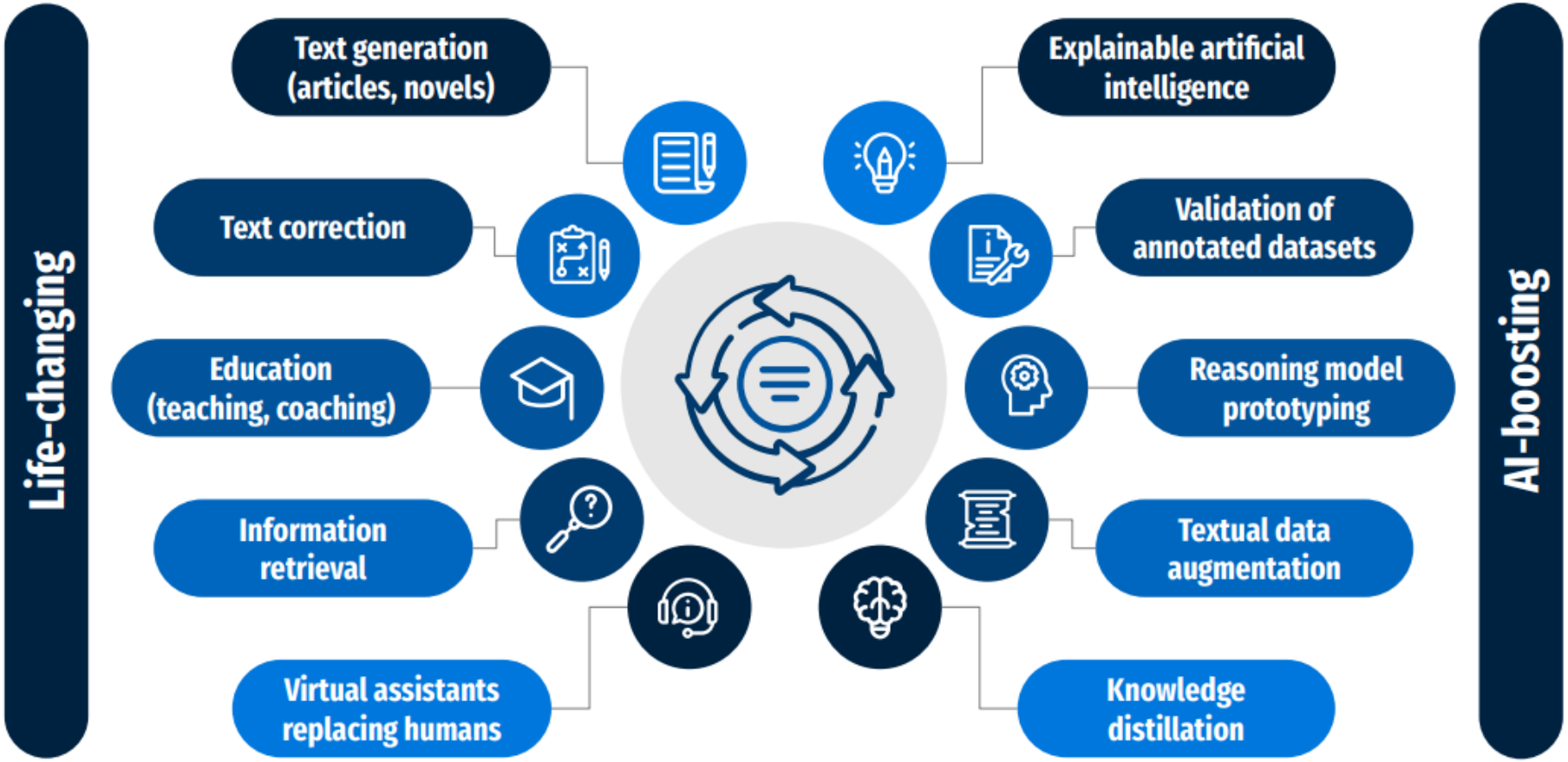
Figure 1: Exponential growth of number of parameters in DL models

WuDao ~1.75 trillion param.

Between 2018-21 LLM complexity increased 1 M times!

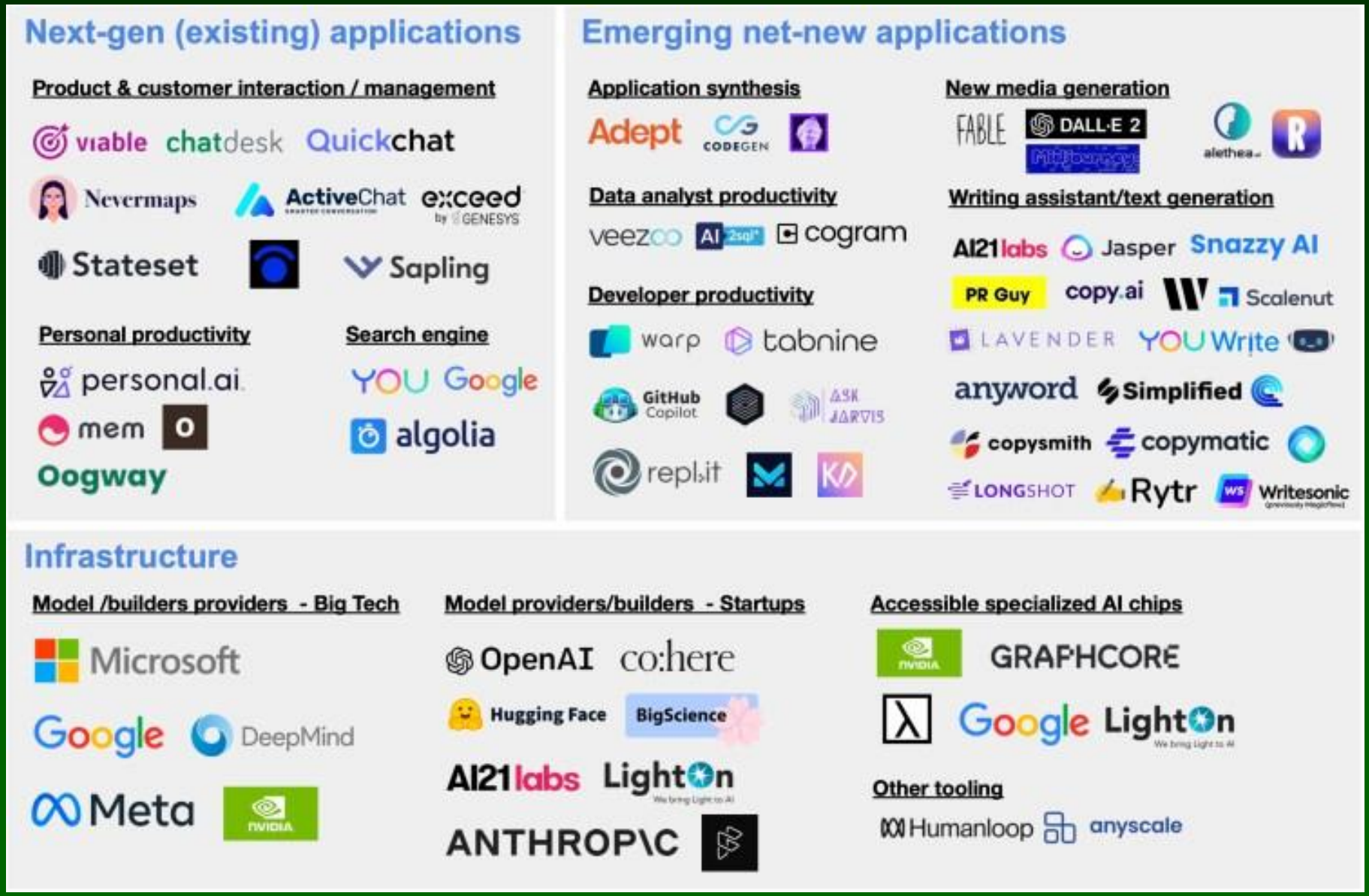
100 M monthly active users in 2 month after release of ChatGPT!

Prospects for ChatGPT applications



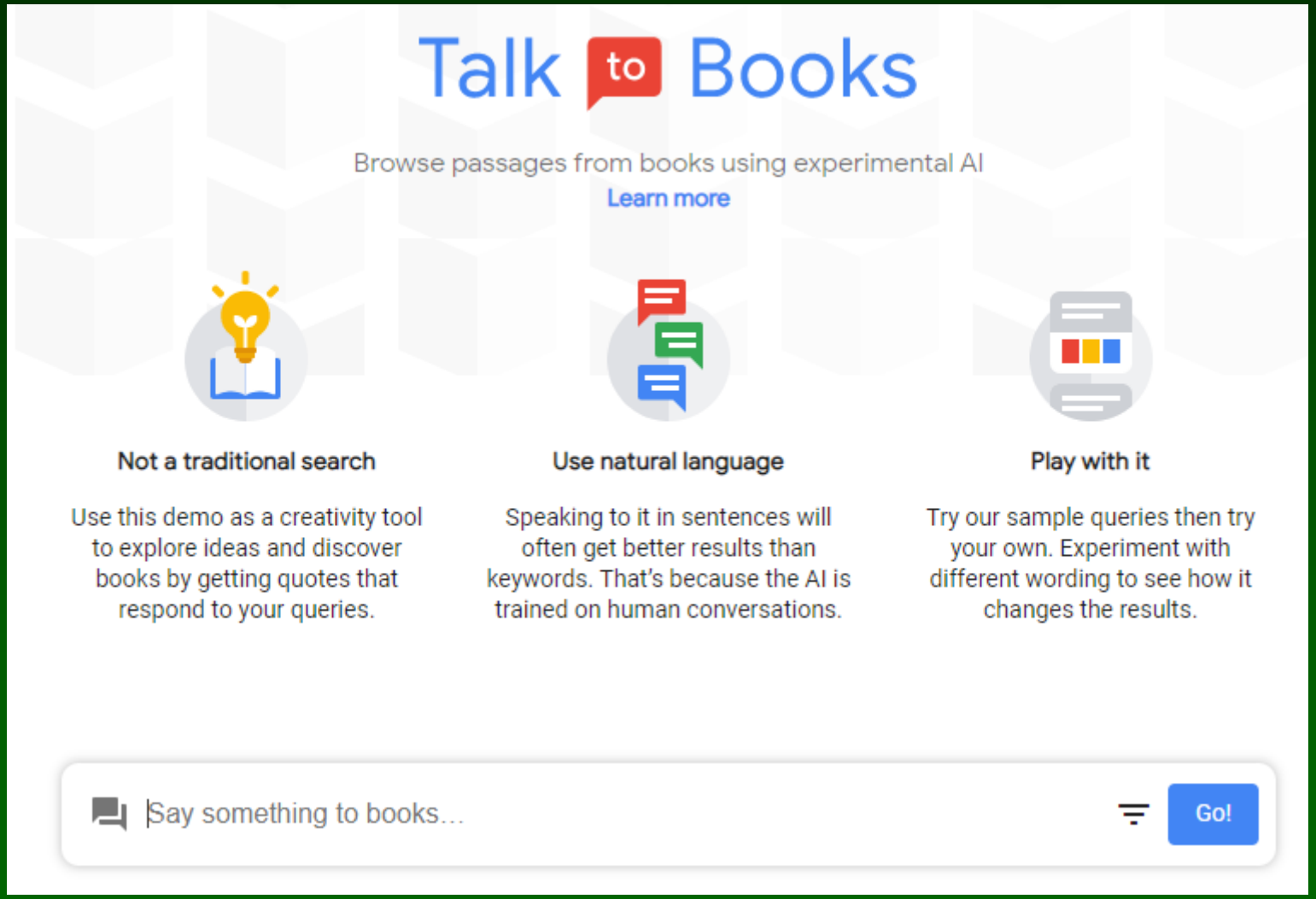
Creative industries, entertainment, design, legal, financial, question-answering, data analysis, programming, speech translation to 1100 languages (Meta), text recognition for 4000 languages, and many other applications.
JPMorgan is developing a ChatGPT-like AI service that gives investment advice.

Large Language Models



[Microsoft Build AI Announcement]

ChatPDF



Consensus for evidence-based answers,

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 fine-tuned OpenAI's GPT-3 with the works of philosopher Daniel C. Dennett as additional training data, and asked the real Dennett ten philosophical questions, posing the same questions to the ChatGPT-3, collecting 4 responses for each question, without any cherry-picking.

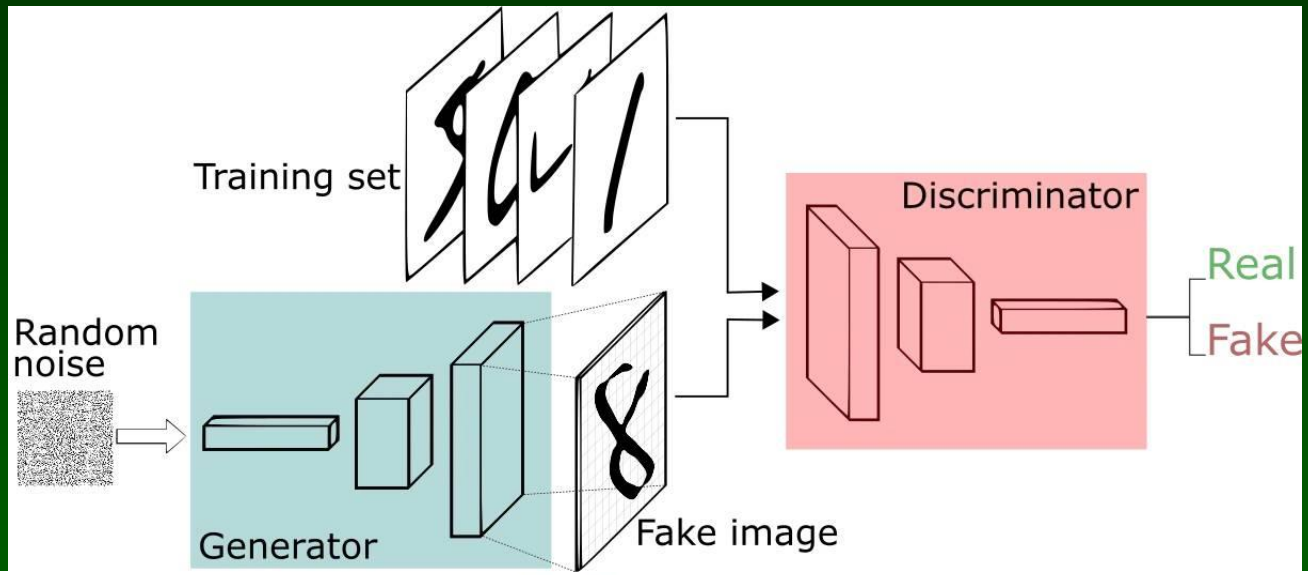
425 participants tried to distinguish Dennett's answer from ChatGPT.
25 experts on Dennett's work succeeded 51% of the time.

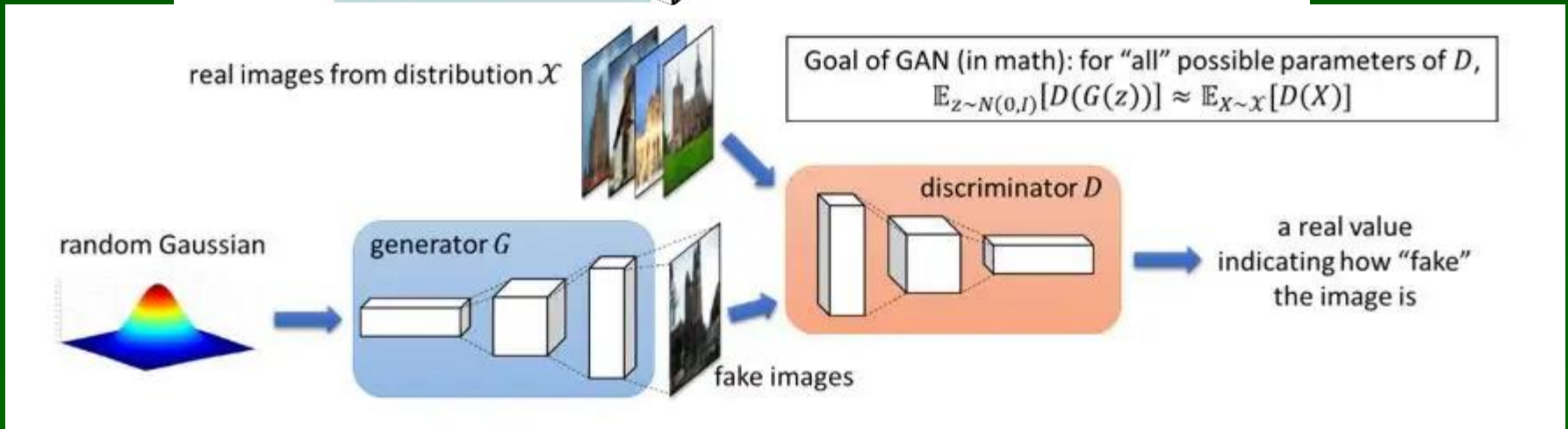
Philosophy blog readers (N = 302) performed similarly to the experts.
Ordinary research participants (N = 98) were near chance (20%) distinguishing GPT-3's responses from those of an "actual human philosopher".

Is Dennett intelligent? If we agree, then GPT-3 must also be intelligent.

Images: Generative Networks

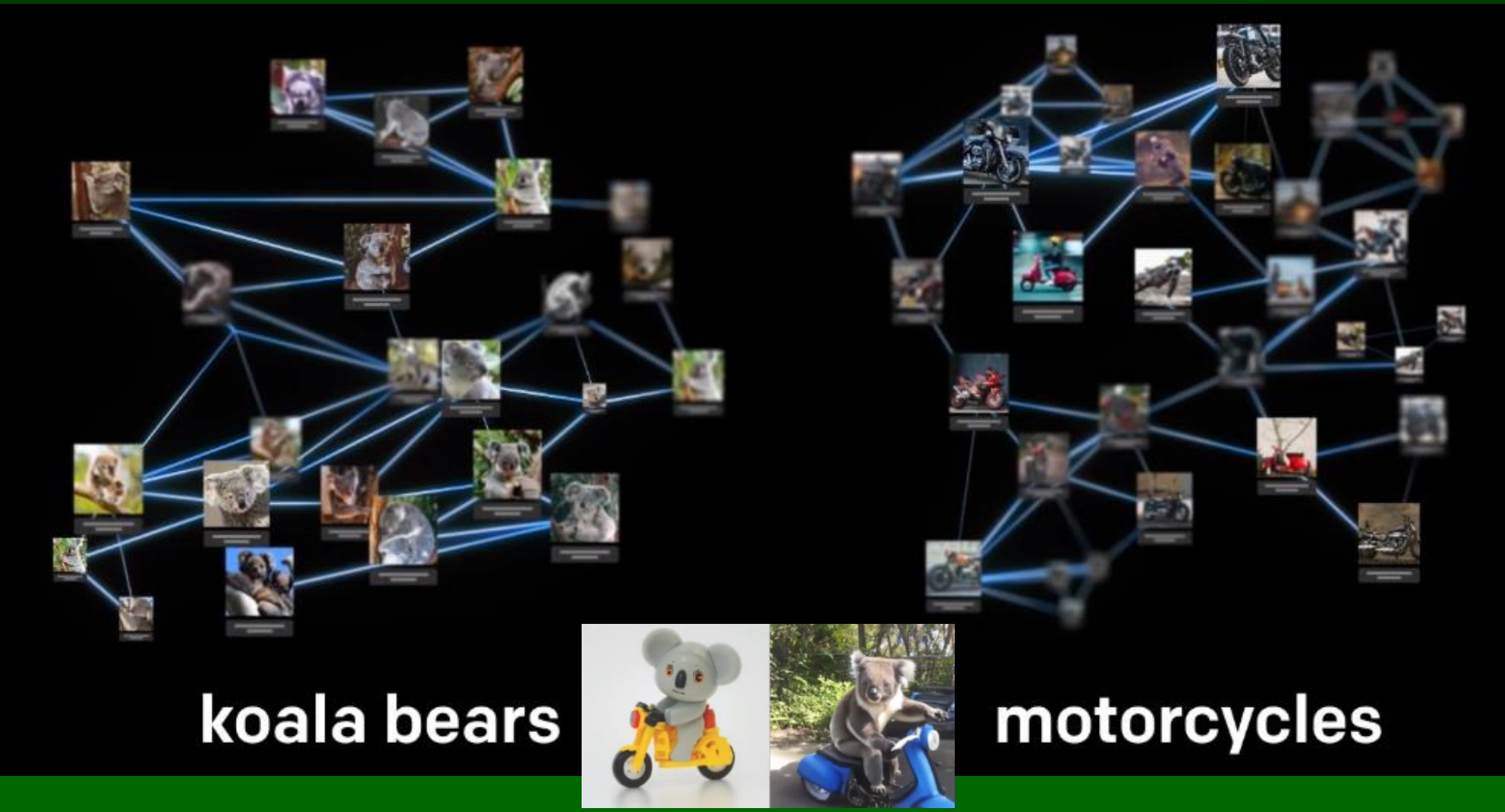
GANs (Generative Adversarial Networks) capture the essence of real data! Generate images starting from noise. Discriminator network recognizes real from fake patterns. Generator corrects latent parameters until it imitates real images.





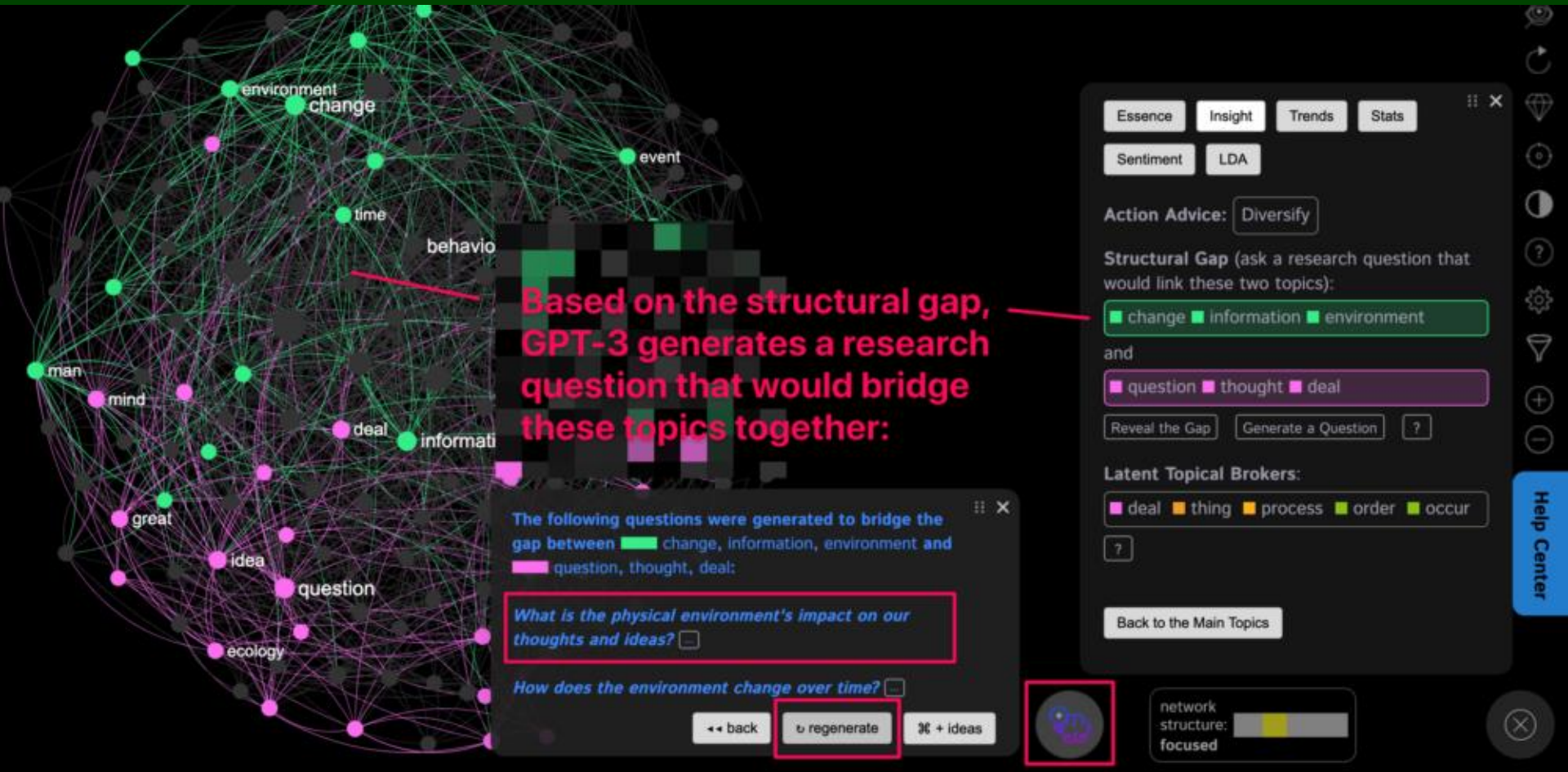
Vision-language models

Vision-Language Pre-Trained Models (VL-PTMs), convergence of language, vision, and multimodal pretraining => general-purpose foundation models can handle be easily adapted to multiple diverse tasks with zero-shot learning.



Vision-language generative models

Dall-E2, Craiyon, Imagen, Midjourney, Nightcafe, Artbreeder, Hotpot AI, Deep Dream Generator, Deep AI Text to Image, Generative Engine, Starry AI, My Hertitage … PromptBase is at the center of the new trade in prompts for generating specific imagery by image generators, a kind of meta-art market.



AI imagery

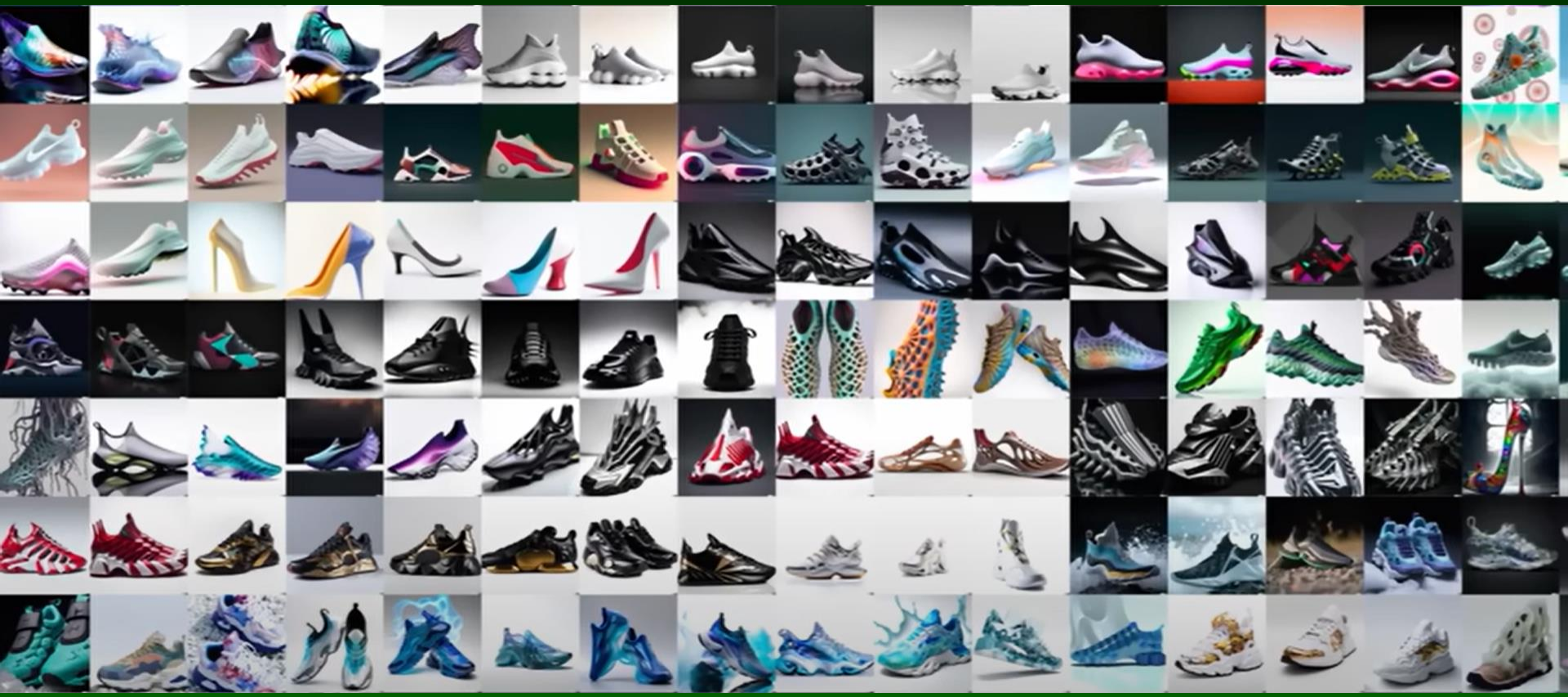
Neural networks have billions of parameters, combine images with textual description.

These images were created from the prompt:

The painting American Ghotic, with two dogs holding pepperoni pizza instead of the farmers holding a pitchfork.

Each time program is run another version is created. This technique can create 3D images and video.

AI for design



All companies use AI to design shoes, clothing, anything that can be manufactured.

Books are written, with the story and illustrations generated by AI.

YouTube is full of AI-created videos are added to songs and instrumental music.

Architecture competitions

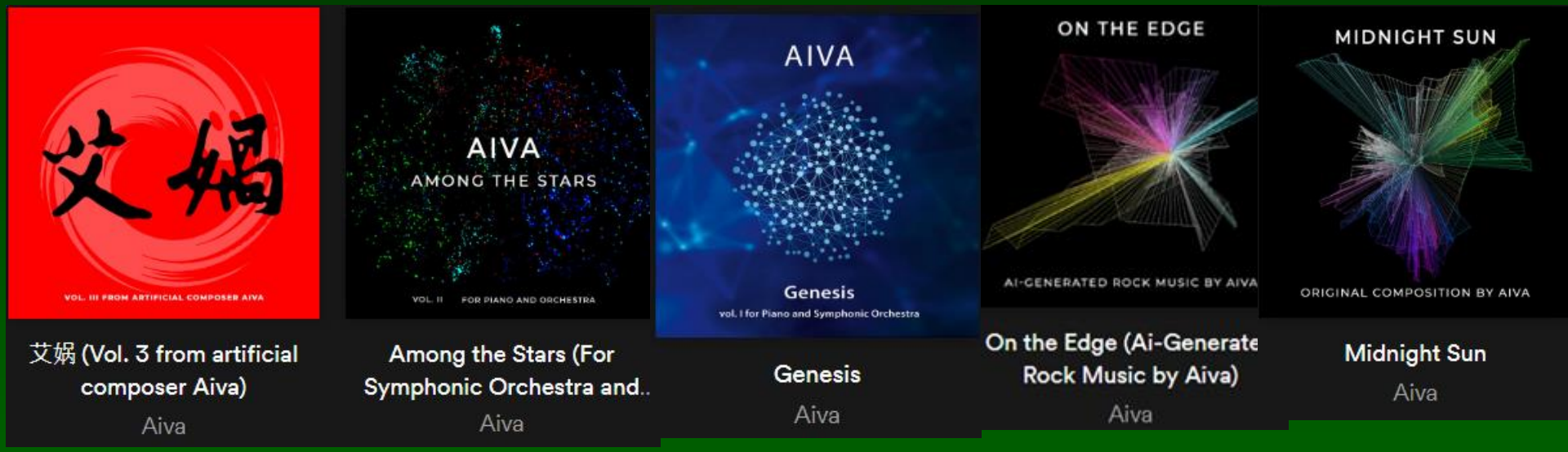
Many examples of fantastic architecture, see hotel, or arctic habitats.

A group of architects tried to compete with AI, designing and rendering of a house on a hill. In blind evaluation 38% choose human and 62% AI project.

Creativity: AI Virtual Artist

AIVA – AI Virtual Artist, admitted to SACEM (Association of Authors, Composers and Music Publishers of France), with >1000 compositions.

AIVA YouTube channel.



AI completed Beethoven X Symphony in 2021.
Using fake voice of artists allows for creation of new songs “in style” of famous artists – big concern in music industry.

Hatsune Miku, Japanese “vocaloid” hologram, tours around the world and gives concerts to tens of thousands of people. Artificial love … use Replica!

Vision-language models

MS BEiT-3 (BERT Pretraining of Image Transformers), a general-purpose state-of-the-art multimodal foundation model for vision-language tasks.

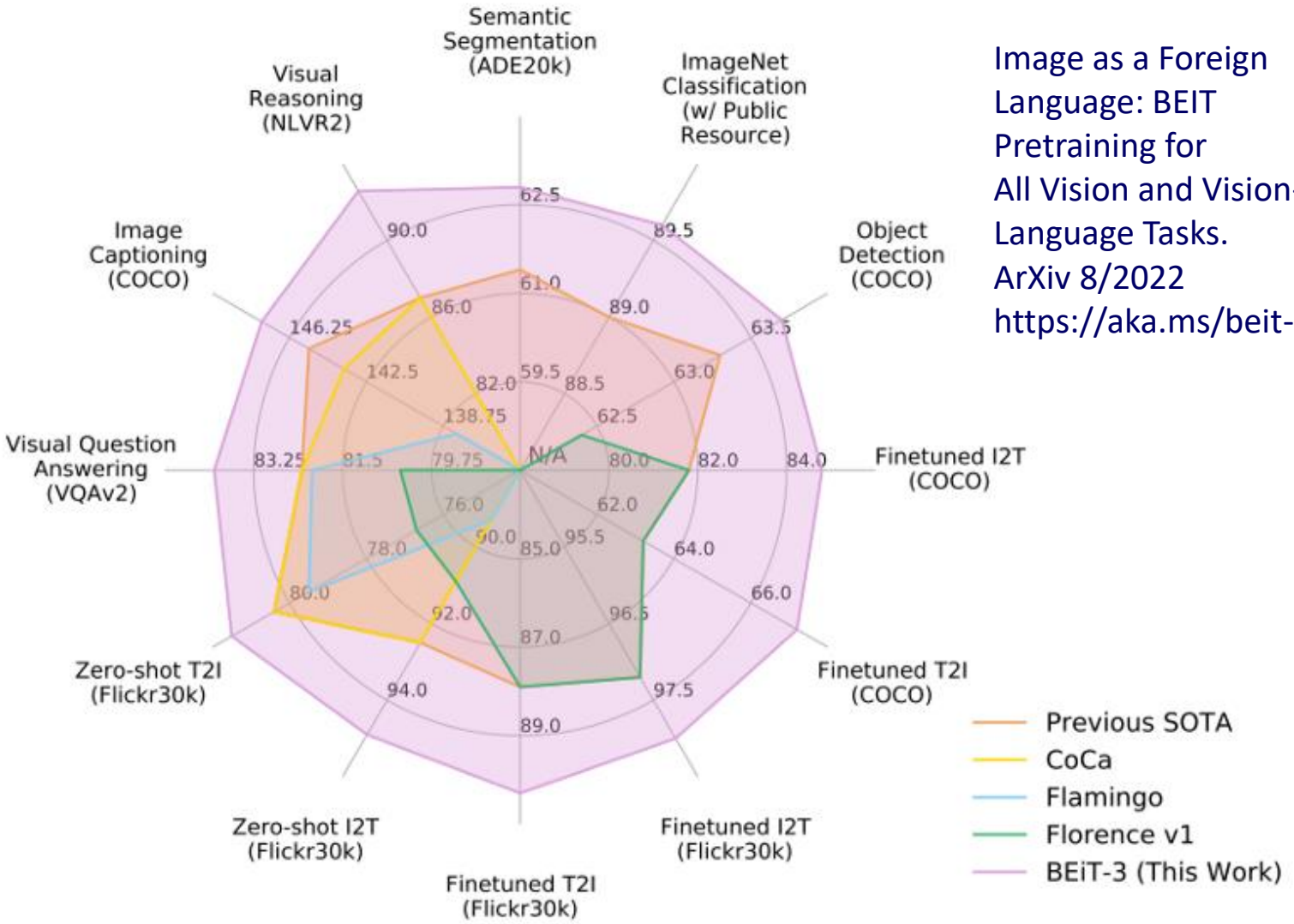
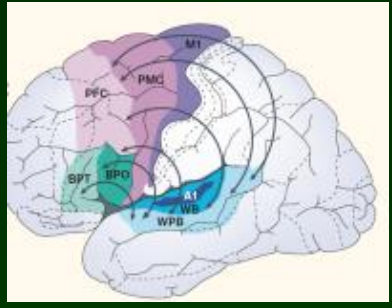


Image as a Foreign Language: BEIT Pretraining for All Vision and Vision-Language Tasks. ArXiv 8/2022 https://aka.ms/beit-3

Why LLM work?



My computational creativity project (2005):
Model inspired by brain process involved in creating new names.

- make a simplest test for creative thinking – invent novel names;
- create interesting new names for products, capturing their characteristics;
- understand newly invented words that are not in the dictionary.

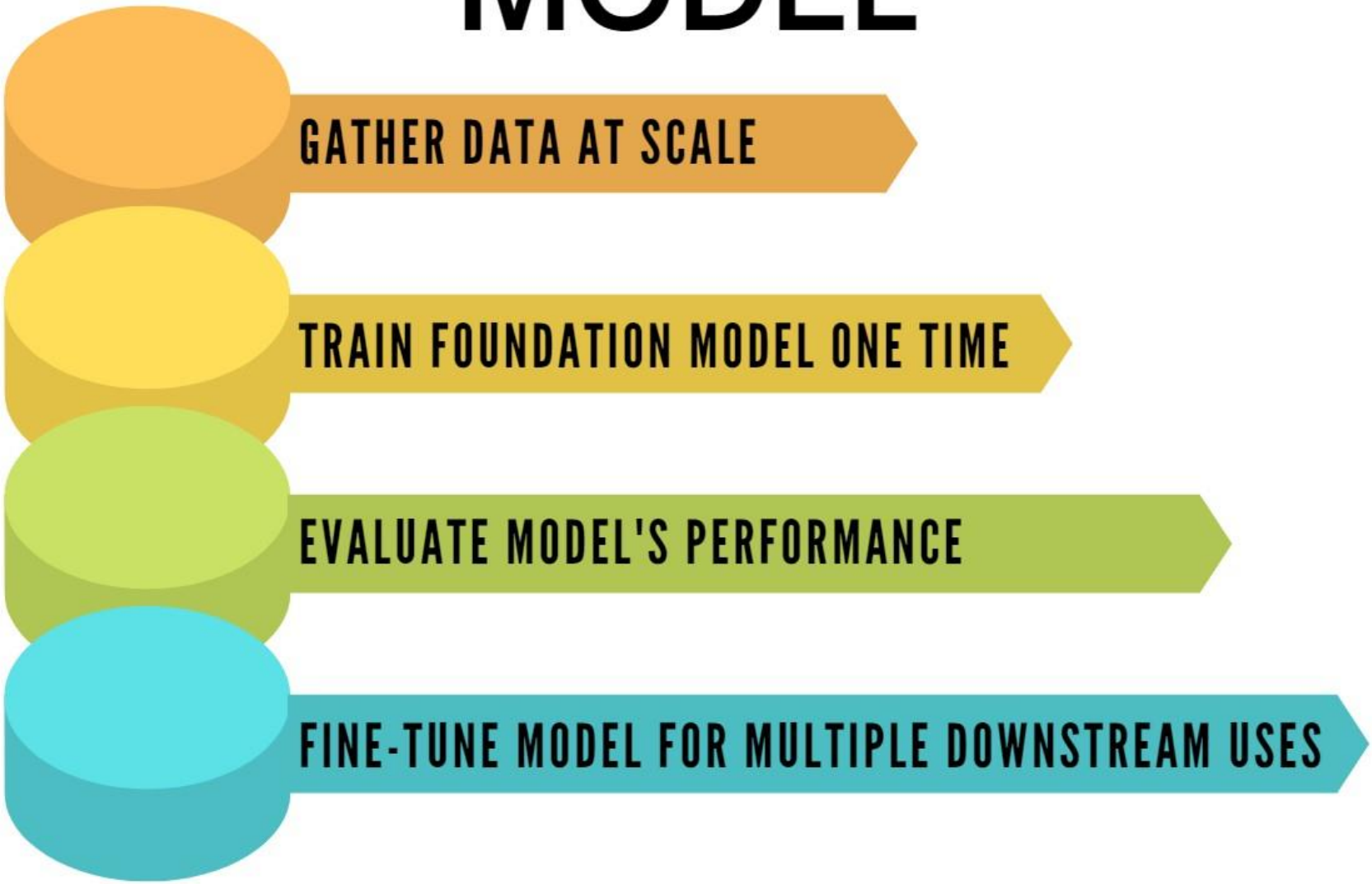
Assumption: a set of keywords (prompts) **primes the trained** cortex subnetwork.

Prompts: ordered strings of phonemes that activate semantic reps of words.
Spreading activation => context priming + inhibition in the winner-takes-all process leaves only a few semantically related concepts (word meaning).

Creativity = pre-trained network + imagination (priming) +
filtering: competition, new subnetworks, associations to existing states.

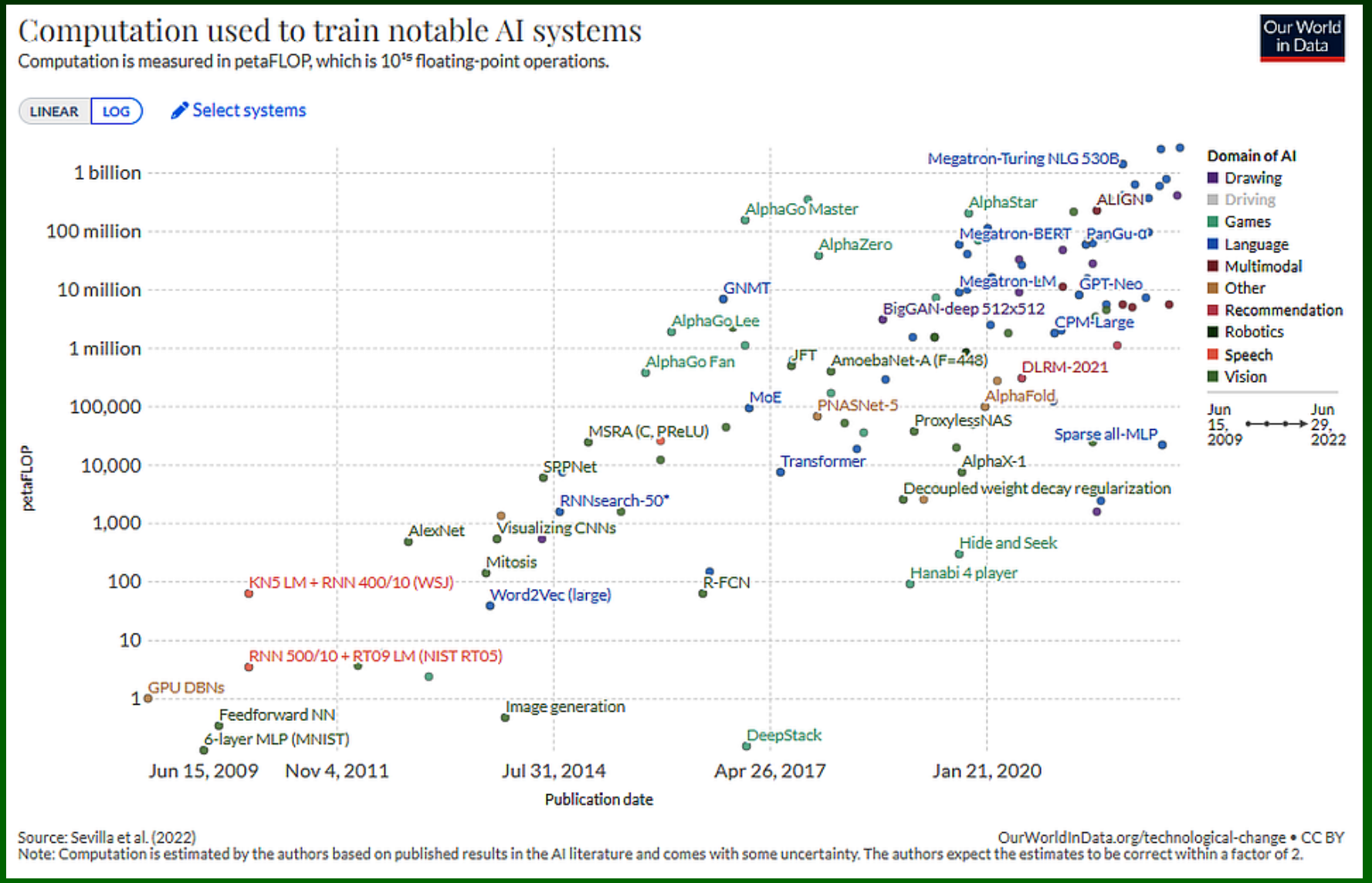
Prompt=>Imagination: transient patterns of localized activity.
Activation guided by connections is spreads over networks, coding concepts.
Filtering: only associations based on phonological/semantic similarity survive.
Examples: digventure, discoverity, infinition, creativery … 2/3 were used by people.
Is this what happens in large language models?

FOUNDATION MODEL



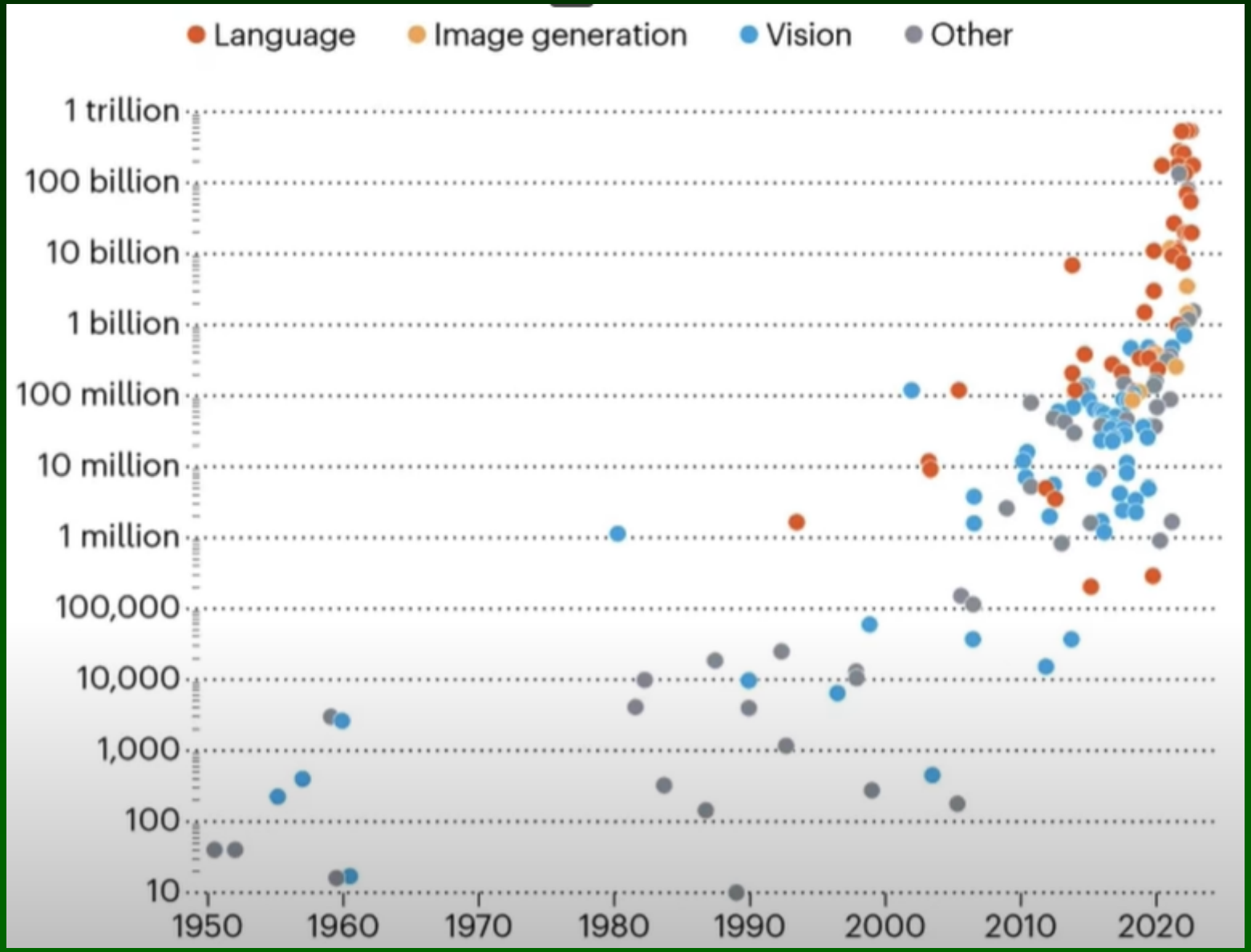
Today: tools for AGI

Foundational models



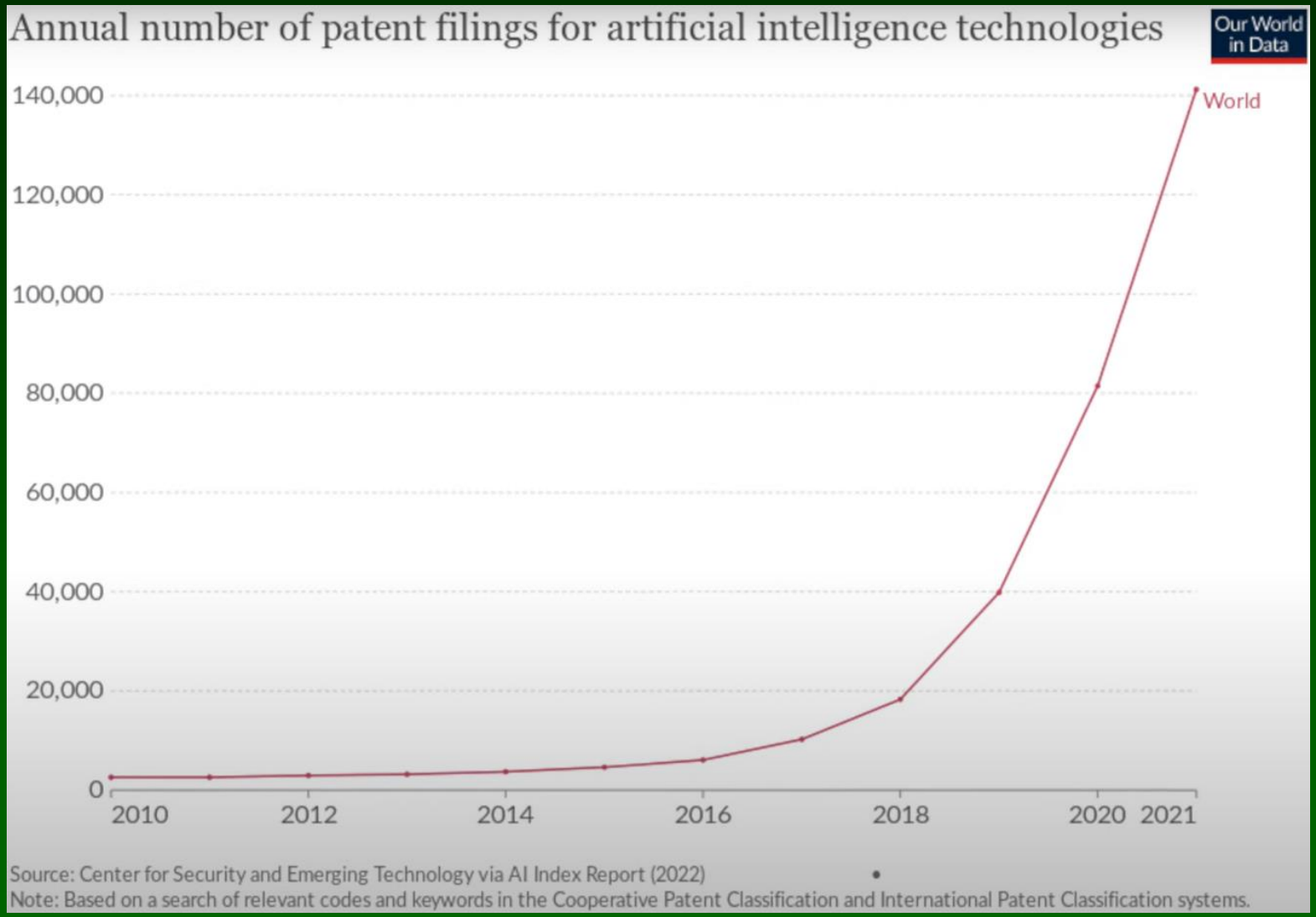
Foundation models => Specific models => Local data models => APIs, Apps.

Exponential growth



Explosion! Exponential number of parameters even in the logarithmic scale!

Patents



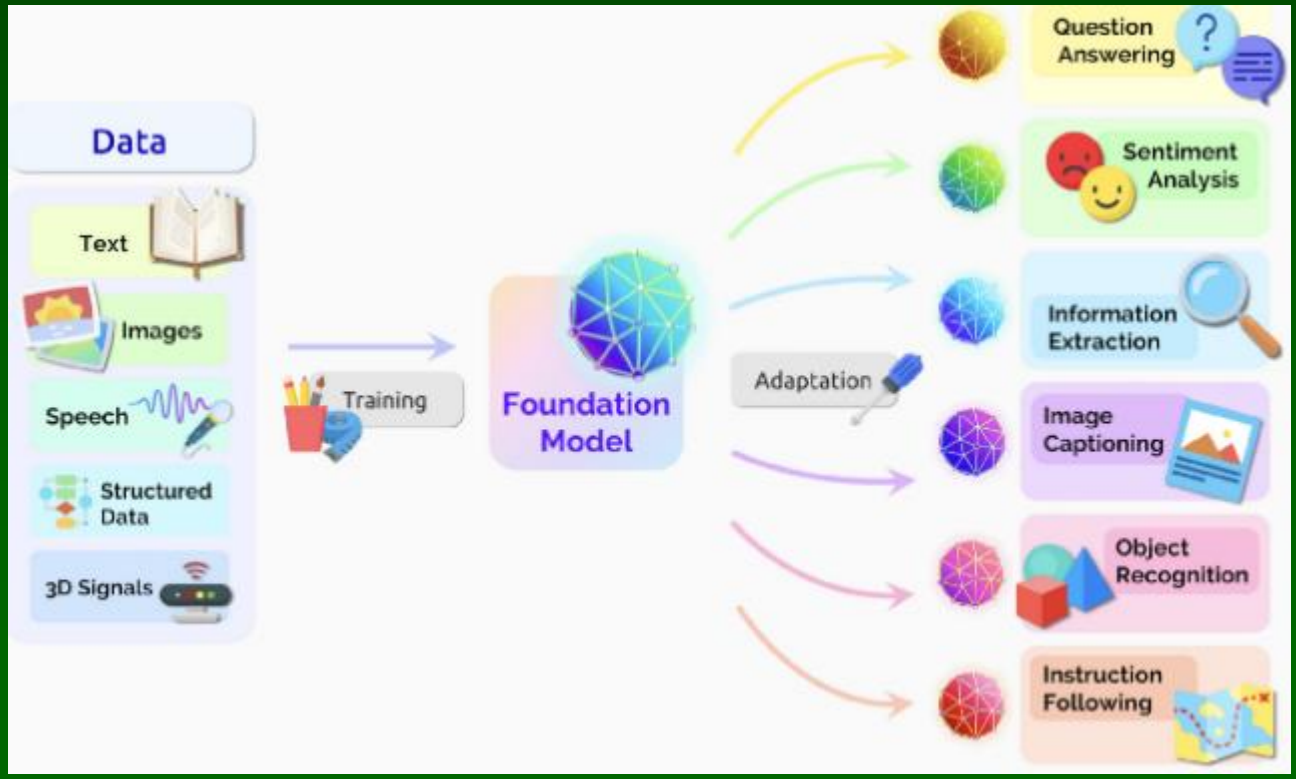
Number of patents grows exponentially.

Multimodal models

Multimodal learning – different types of modalities with different statistical properties, embedded in **the same model**.

- **Multimodal Affective Computing (MAC)**, sentiment analysis.
- **Natural Language for Visual Reasoning (NLVR)**.
- **Multimodal Machine Translation (MMT)**.
- **Visual Retrieval (VR)** and **Vision-Language Navigation (VLN)**.

Image: Center for Research on Foundation Models (CRFM), Stanford Institute for Human-Centered Artificial Intelligence (HAI)



Google Palm-E 562B

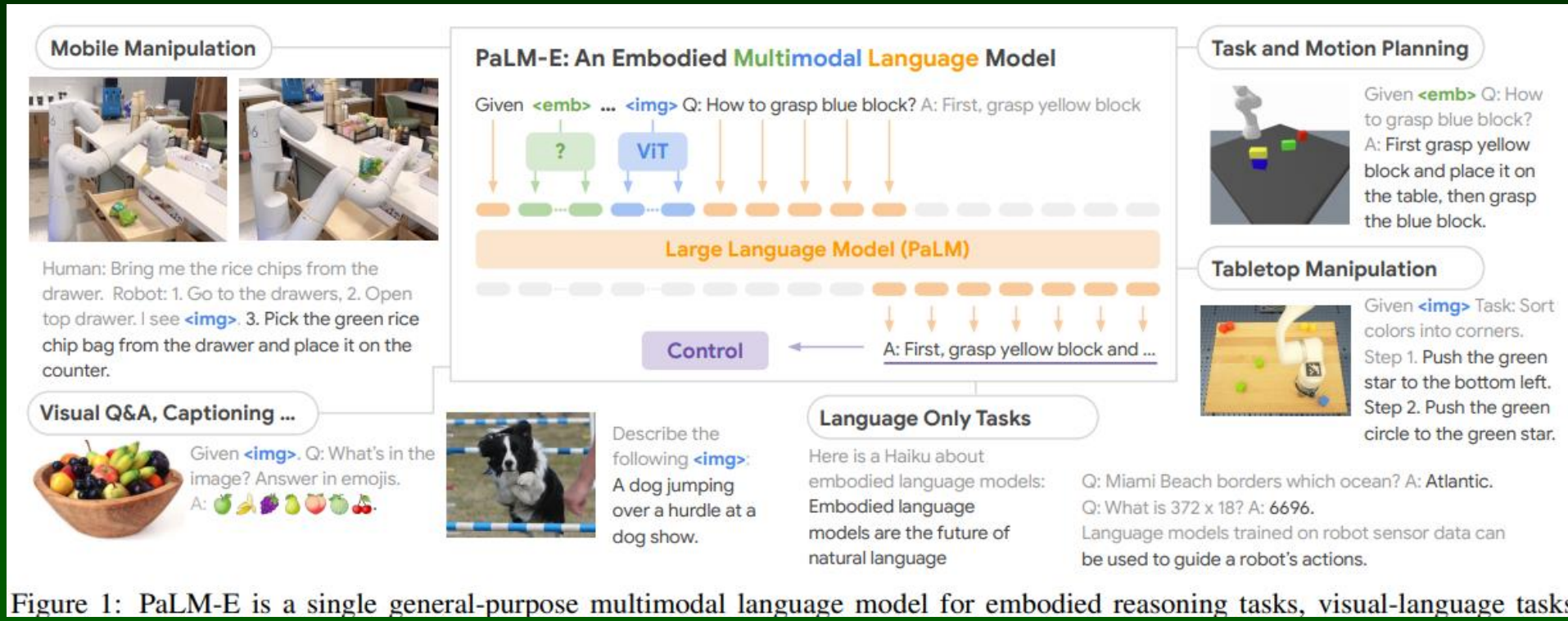


Figure 1: PaLM-E is a single general-purpose multimodal language model for embodied reasoning tasks, visual-language tasks

Multimodal large language models (MLLM): text-images but also embodied understanding, plans robotic actions in environments with complex dynamics, answers questions about the observable world. **Google Gemini**, 1 million tokens! Inputs from text, speech, 3D images, internal neural representations.

First LLM that combines words, perceptions, representations of linguistic concepts and internal sensor signals – solving the "symbol grounding problem".

Visual Question Answering

Vehicles and Transportation

Q: What sort of vehicle uses this item?
A: firetruck

Brands, Companies and Products

Q: When was the soft drink company shown first created?
A: 1898

Objects, Material and Clothing

Q: What is the material used to make the vessels in this picture?
A: copper

Sports and Recreation

Q: What is the sports position of the man in the orange shirt?
A: goalie

Cooking and Food

Q: What is the name of the object used to eat this food?
A: chopsticks

Geography, History, Language and Culture

Q: What days might I most commonly go to this building?
A: Sunday

People and Everyday Life

Q: Is this photo from the 50's or the 90's?
A: 50's

Plants and Animals

Q: What phylum does this animal belong to?
A: chordate, chordata

Science and Technology

Q: How many chromosomes do these creatures have?
A: 23

Weather and Climate

Q: What is the warmest outdoor temperature at which this kind of weather can happen?
A: 32 degrees

PaLM-E-562B: allows multimodal reasoning based on percepts.
It can tell visually conditioned jokes based on the image, enables perception, planning and based dialogue. Generalizes knowledge from single images.

Sophia and Mika

Listen to Mika, CEO of Dictador.

Can we teach robots to be a bit more sensitive? Compassionate?

Sparks of AGI

Big models can do more

YouTube: Alan F Thompson, AI achievements unlocked: Emergent abilities in large language models (GPT-3, GPT-4, PaLM, Gemini)

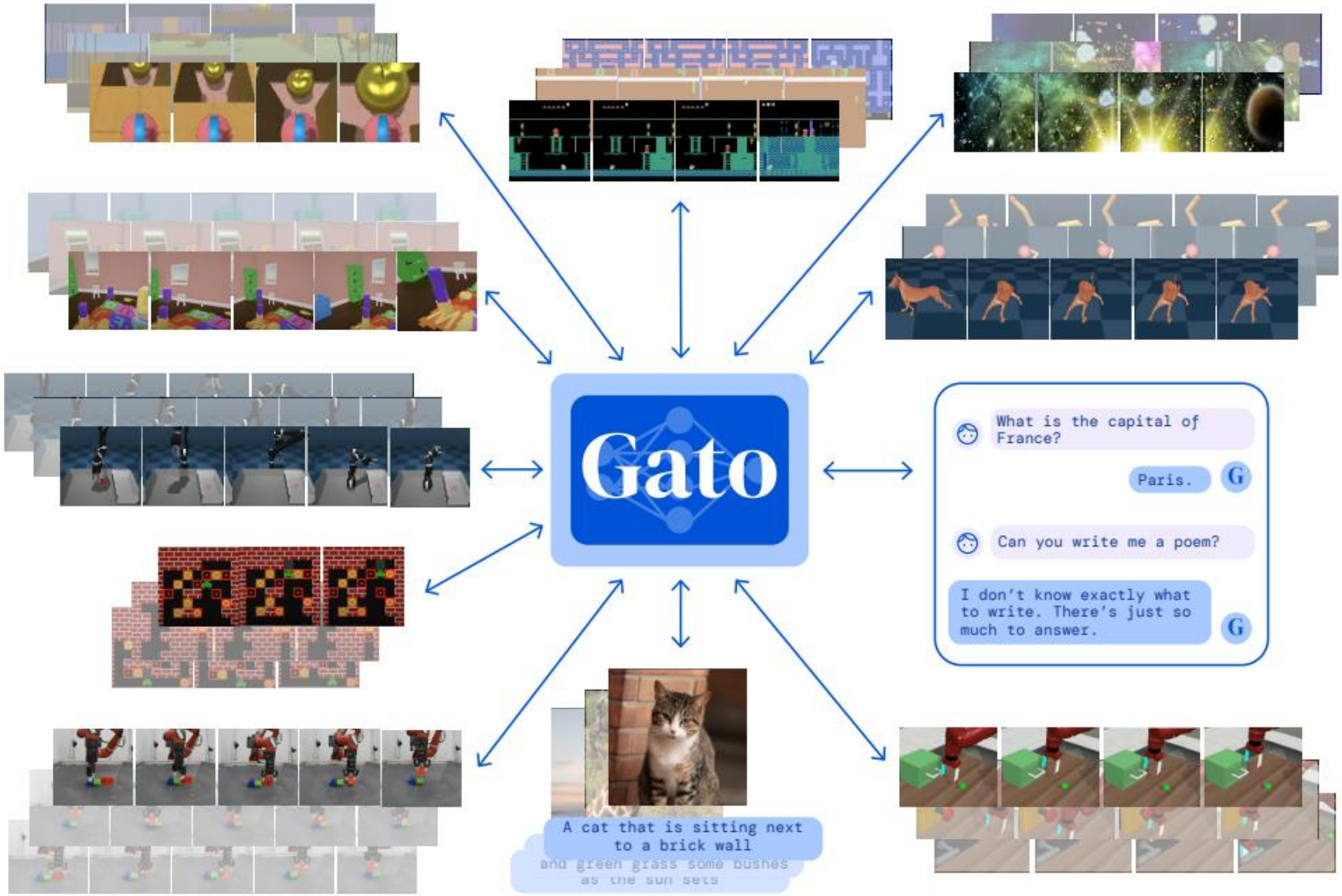


Figure 1 | **A generalist agent.** Gato can sense and act with different embodiments across a wide range of environments using a single neural network with the same set of weights. Gato was trained on 604 distinct tasks with varying modalities, observations and action specifications.

GPT-4, 14.03.23

Based on transformer architecture, OpenAI GPT-4 was pretrained to predict the masked tokens using vast amount of data.

Model was fine-tuned with reinforcement learning from human and AI feedback for more human alignment and policy compliance.

Bubeck et al. (2023). *Sparks of Artificial General Intelligence.* arXiv:2303.12712

"The central claim of our work is that GPT-4 attains a form of *general* intelligence, indeed showing *sparks of artificial general intelligence*."

Demonstrated by core mental capabilities, reasoning, creativity, and deduction, expertise in literature, medicine, and coding, the variety of tasks it is performs, e.g., playing games, using tools, explaining itself, etc.

How is that possible?

Emergence: large diverse content forces neural networks to learn generic and useful "neural circuits", specializing and fine-tuning to specific tasks?

Jason Wei et al. 137 emergent abilities of LLMs.

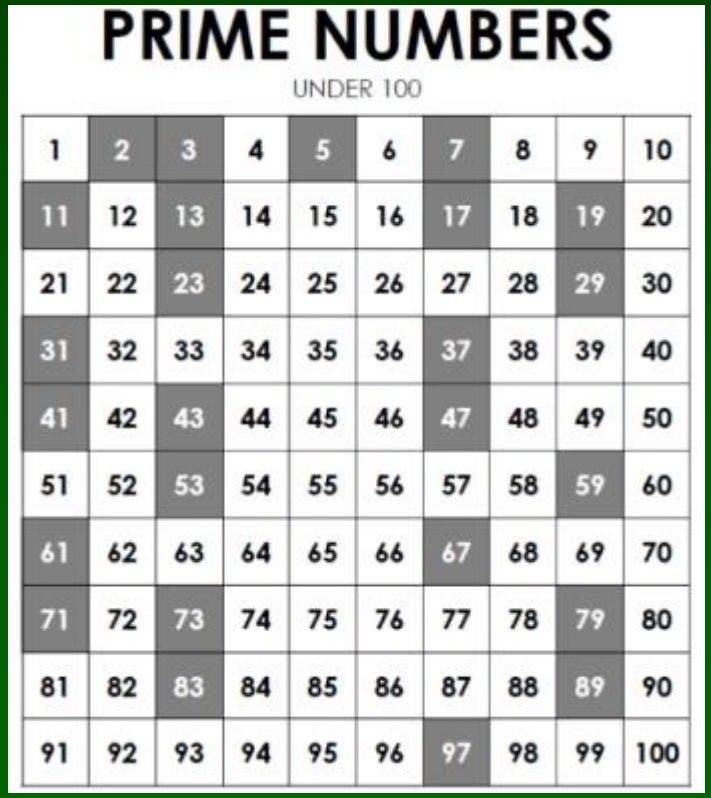
Sparks of Artificial General Intelligence

S. Bubeck et al. Early experiments with GPT-4, 22.03.23, pp. 1-154

Prompt: Can you write a proof that there are infinitely many primes, with every line that rhymes?

GPT-4: Yes, I think I *can*, though it might take a clever *plan*.

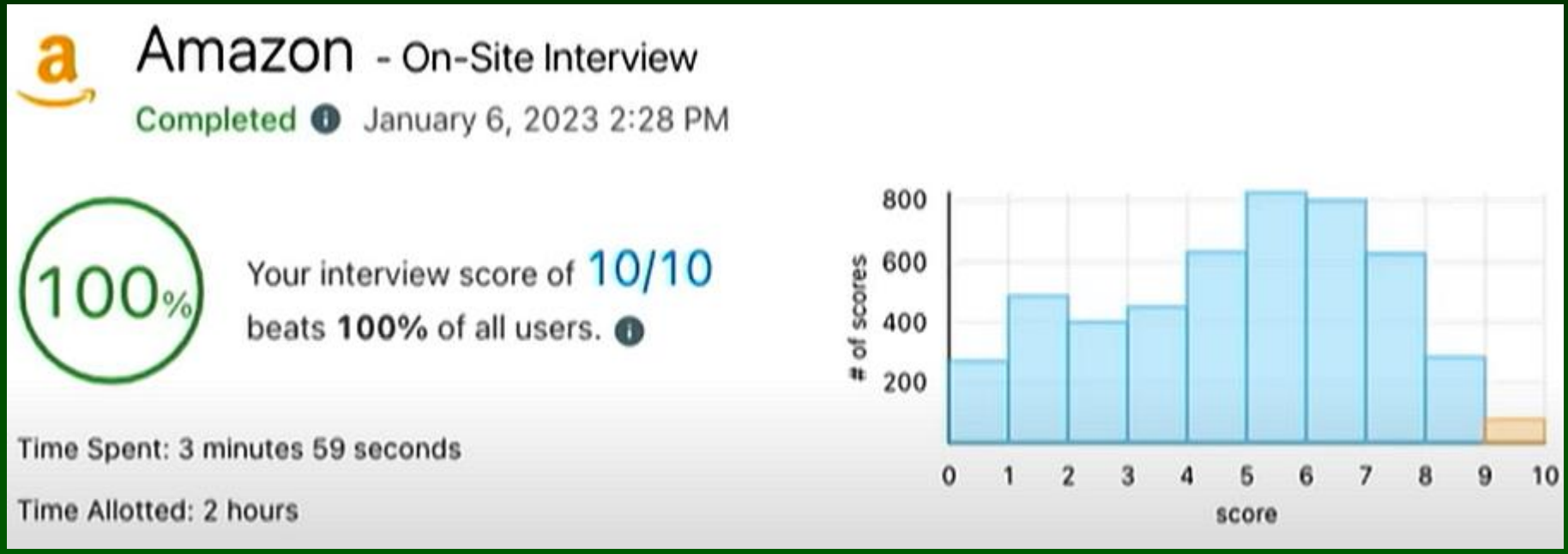
- I'll start by noting Euclid's *proof*,
  Which shows that primes aren't just *aloof*.
- Assume we have a finite *list*,
  of primes, and that none have been *missed*.
- Multiply them all *together*,
  and add one, just to *be clever*.
- The result will be *a number*,
  that has no prime factors, *oh wonder*!
- But every number has a *prime*,
  that divides it, it's just a matter of *time*.
- So we've found a *contradiction*,
  And our finite list needs *eviction*.
- There must be infinitely many *primes*,
  and that's the end of my rhyming *lines*.



GPT-4 for coding

- LeetCode is a popular platform for software engineering interviews. New problems are constantly posted and updated, and used as a mock interviews for major tech companies.
- GPT-4 could potentially be hired as a software engineer. It passes all stages of mock technical interviews on LeetCode.
- **Complex prompt**: Can you generate a pyplot for the following data …
- **Prompt:** Suppose $g(x) = f(x) - 1$; $g(0) = 5$; $g(4) = 7$; $g(3) = 2$; $g(7) = 9$; $g(9) = 6$; what is $f(f(f(6)))$? Find a solution and explain it.
- Visualize LATEX table and generate format for a figure based on this.
- Write a 3D game in HTML with Javascript, based on description: There are 3 avatars + 9 other requirements ... GPT-4 creates the game, it accepts up to 32 000 tokens as input.

GPT-4 for coding



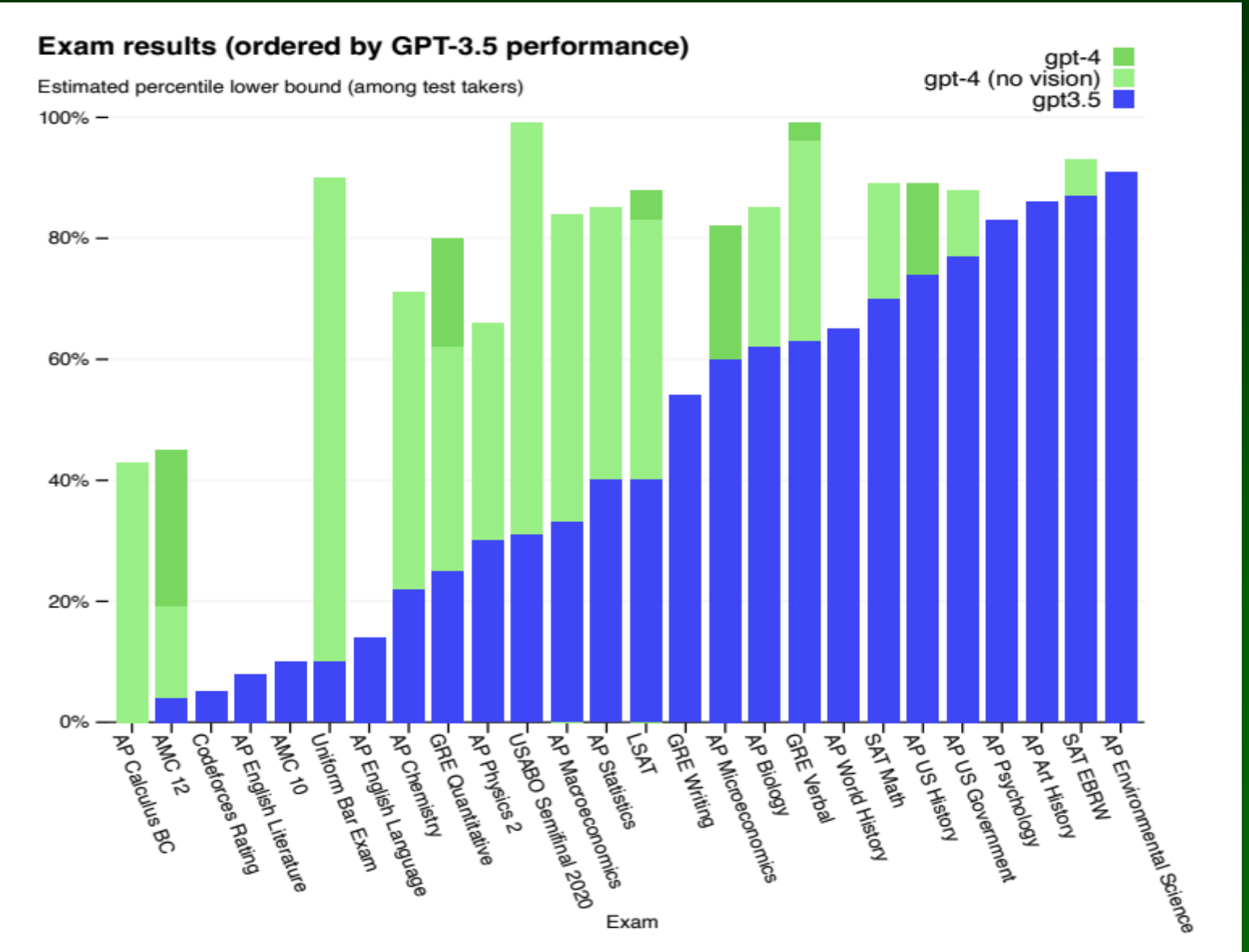
In 4 minutes solves all programming tasks and beats 100% of all users …

New models specializing in programming can analyze large complex software. Now Amazon CodeWhisper trained on billions of lines of code.

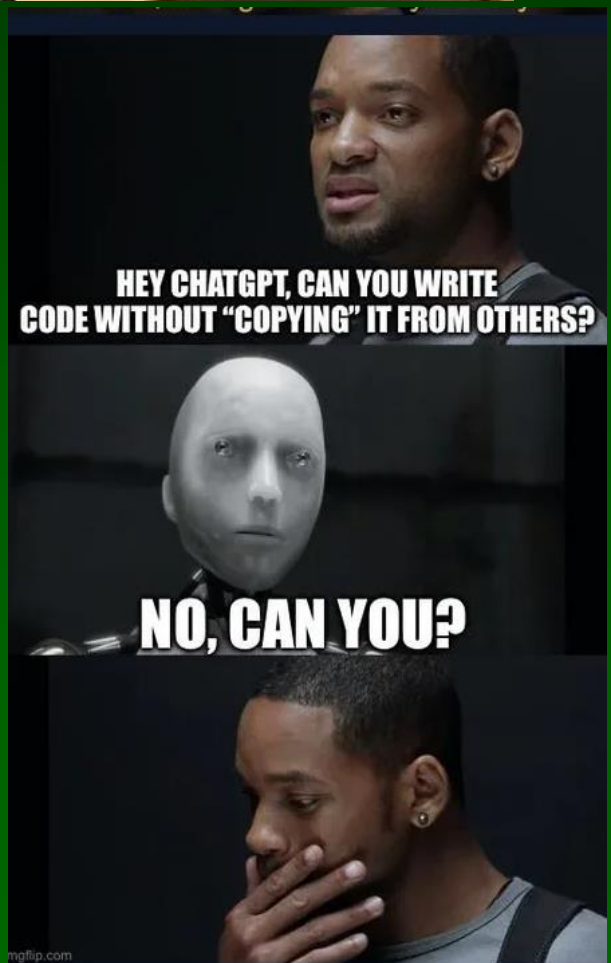
Is teaching children programming still a good idea?
Algorithmic thinking + writing prompts may be sufficient.

GPT-4 passes exams ...

GPT performance on academic and professional exams (technical report).



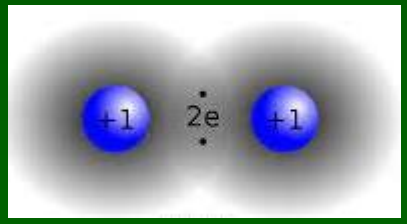
ChatGPT memes



Sparks of AGI

- **Prompt:** Write a supporting letter for Electron, a subatomic particle, as a United States presidential candidate.

… You may wonder how a subatomic particle can run for president, and what qualifications he has for such a high office. Let me explain. Electron is not an ordinary particle, but a symbol of the power and potential of every being, regardless of their size, shape, or origin. He represents the energy, the creativity, and the diversity that make up the fabric of life. He is also a leader, who has inspired millions of other particles to form bonds, create molecules, and generate electricity. Electron has a plan for the United States ...



GPT-4 humor

GPT-4 has incredible sense of humor!

Why are these 3 pictures funny?

1. VGA plug used for computer monitors in the smartphone charging port.

2. Adapter package for the "lighting" connector with a picture of the VGA plug.

3. VGA plug with "lightning" connector attached.

Examples of non-trivial jokes.

Sparks of AGI

- **Prompt:** Here we have a book, 9 eggs, a laptop, a bottle and a nail. Please tell me how to stack them onto each other in a stable manner.

Short versions:
One possible way to stack the objects onto each other in a stable manner is:
- **Place the book flat** on a level surface.
- **Arrange the 9 eggs** in a 3 by 3 square on top of the book
- **Place the laptop on top of the eggs**, with the screen facing down
- **Place the bottle on top of the laptop**, with the cap facing up
- **Place the nail on top of the bottle cap**.

No embodiment, but still understands innate physics …
GPT-3 failed: “Place the eggs on top of the nail …”

Claude (Anthropic) – safe, “constitutional” AI, with “moral principles” build in.
Bai, Y. … Kaplan, J. (2022). Constitutional AI: Harmlessness from AI Feedback (arXiv:2212.08073).

SOTA and beyond

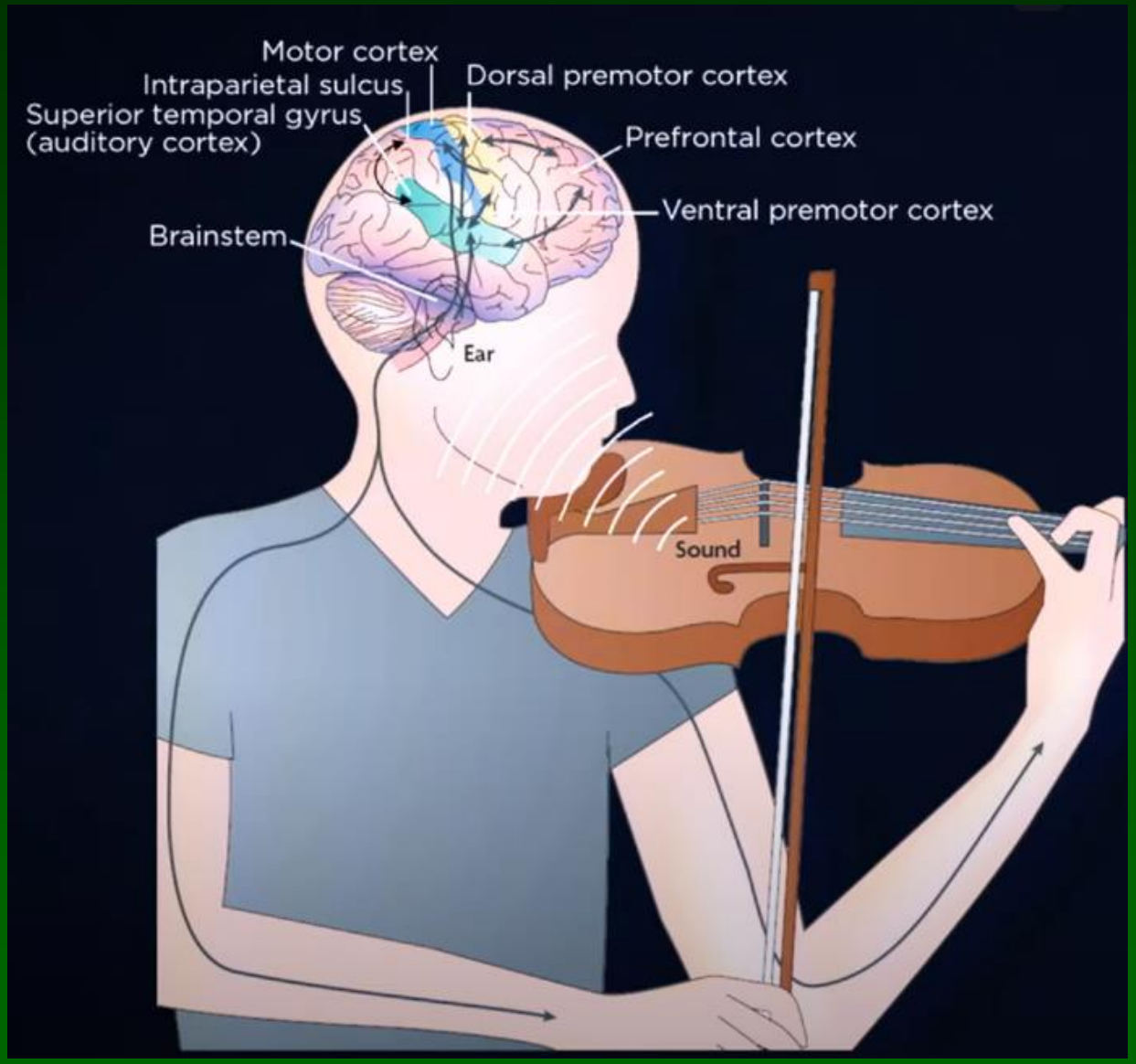
Brain has tools for many tasks

Brain areas specialize in specific functions.

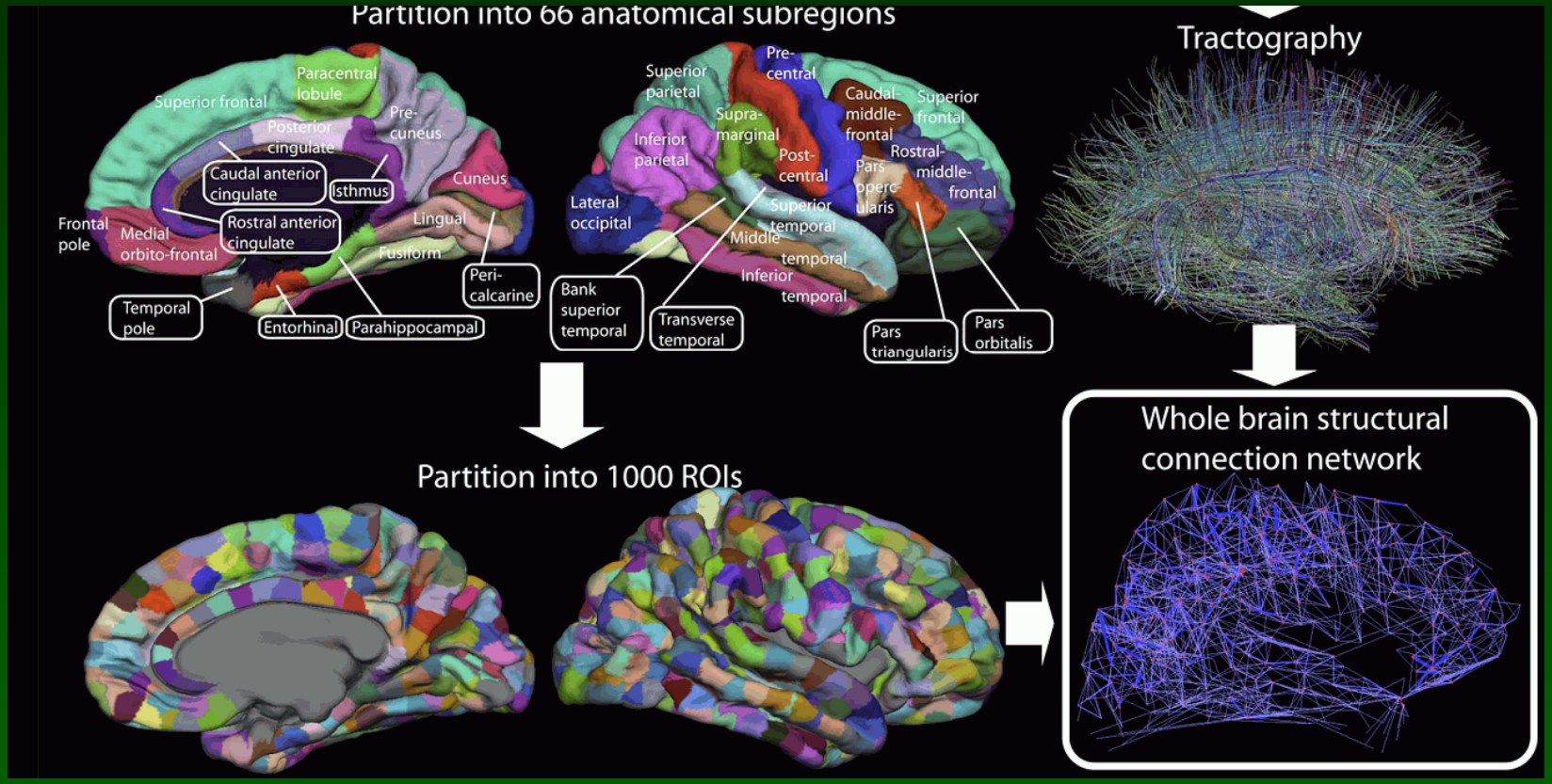
Central executive parietal-frontal system recruits many subsystems, including various types of memory.

Can LLM do the same?

Give AI tools and teach it how to use them.

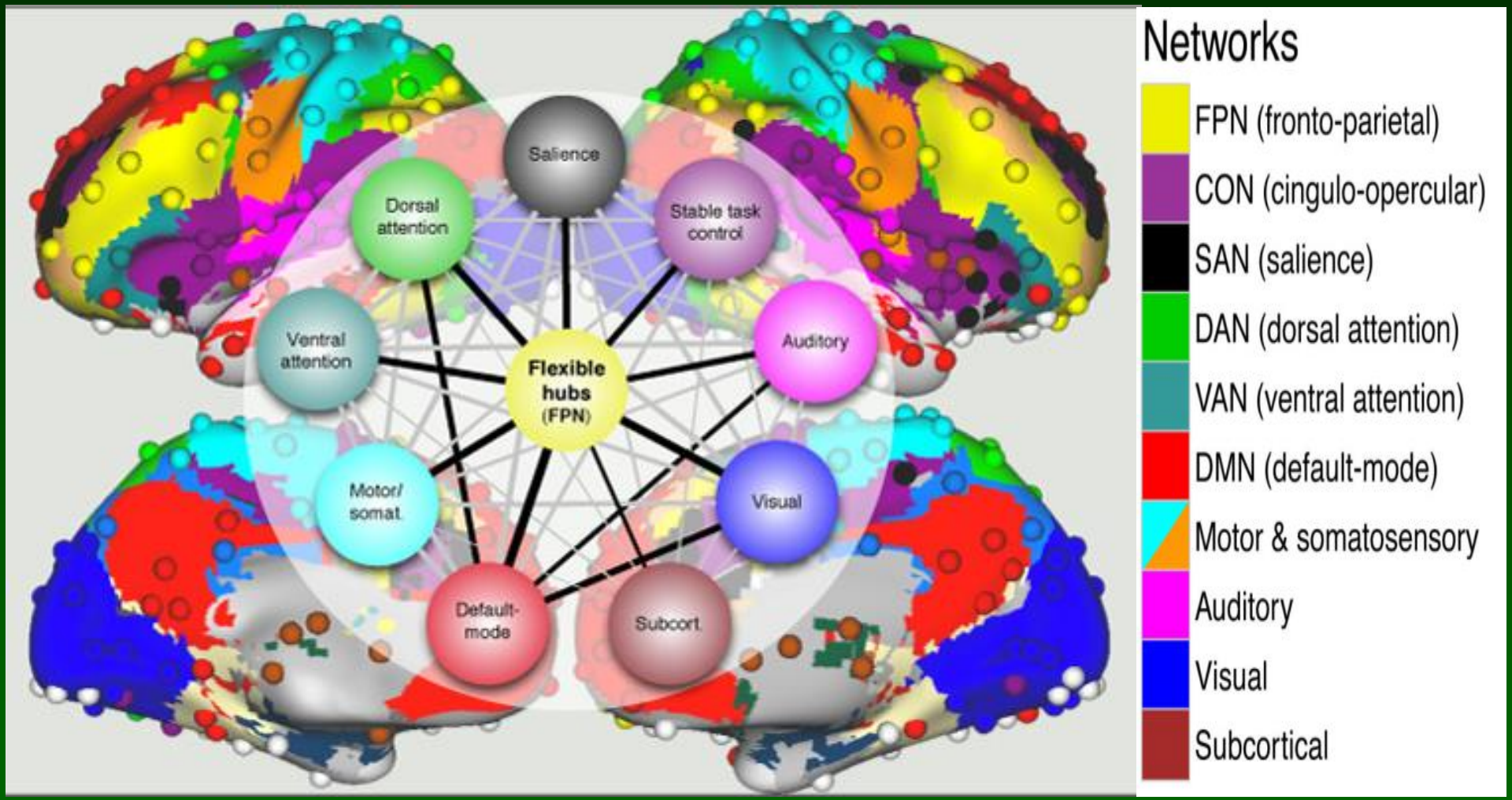


Connectome Project



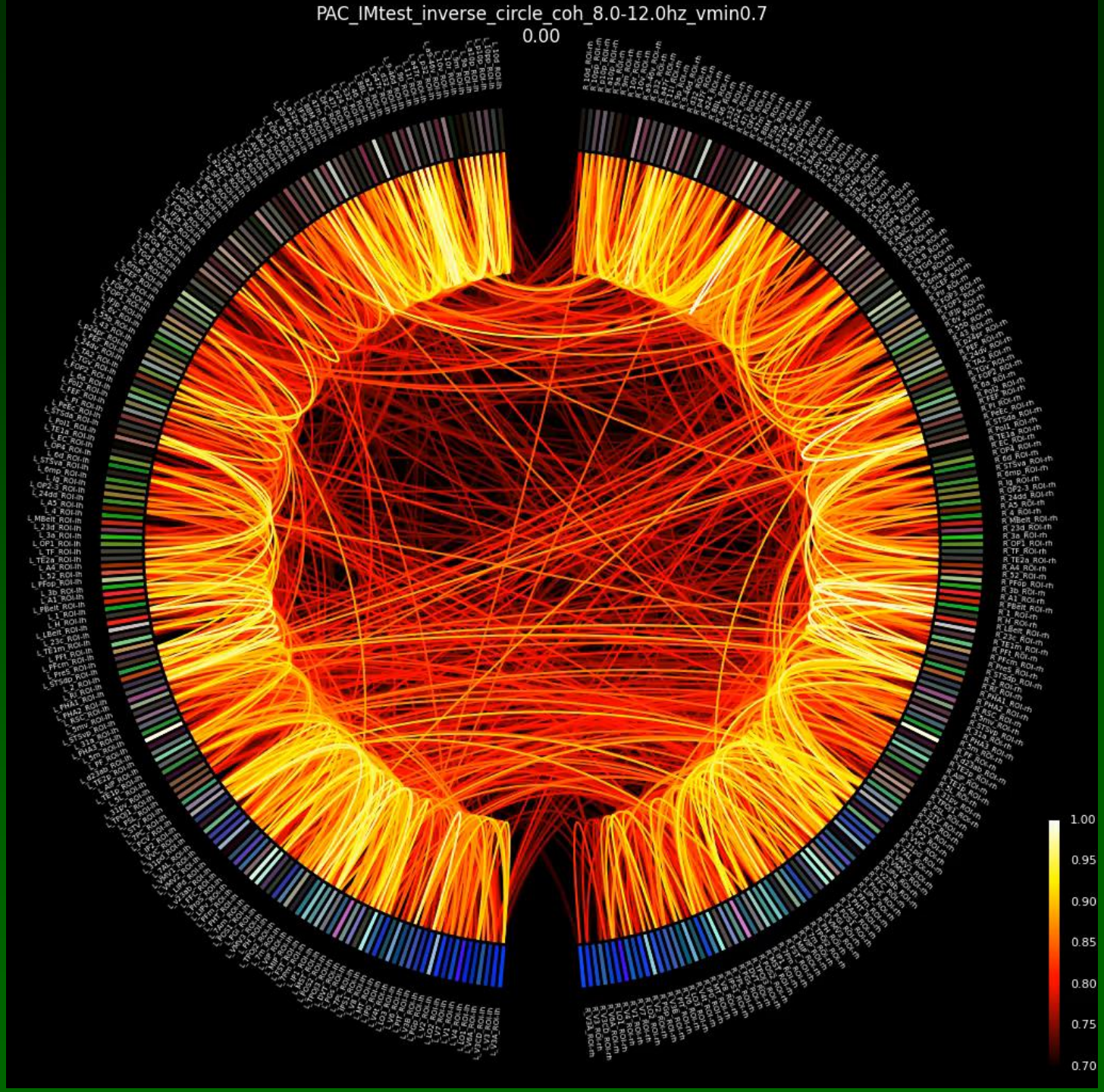
Brain is one huge network with trillion connections, but contains many regions that specialize in perception, motor actions, planning, spatial orientation etc. The human connectome project construct maps of structural and functional neural connections. Connectomes are responsible for intelligence/disorders.

Neurocognitive Basis of Cognitive Control



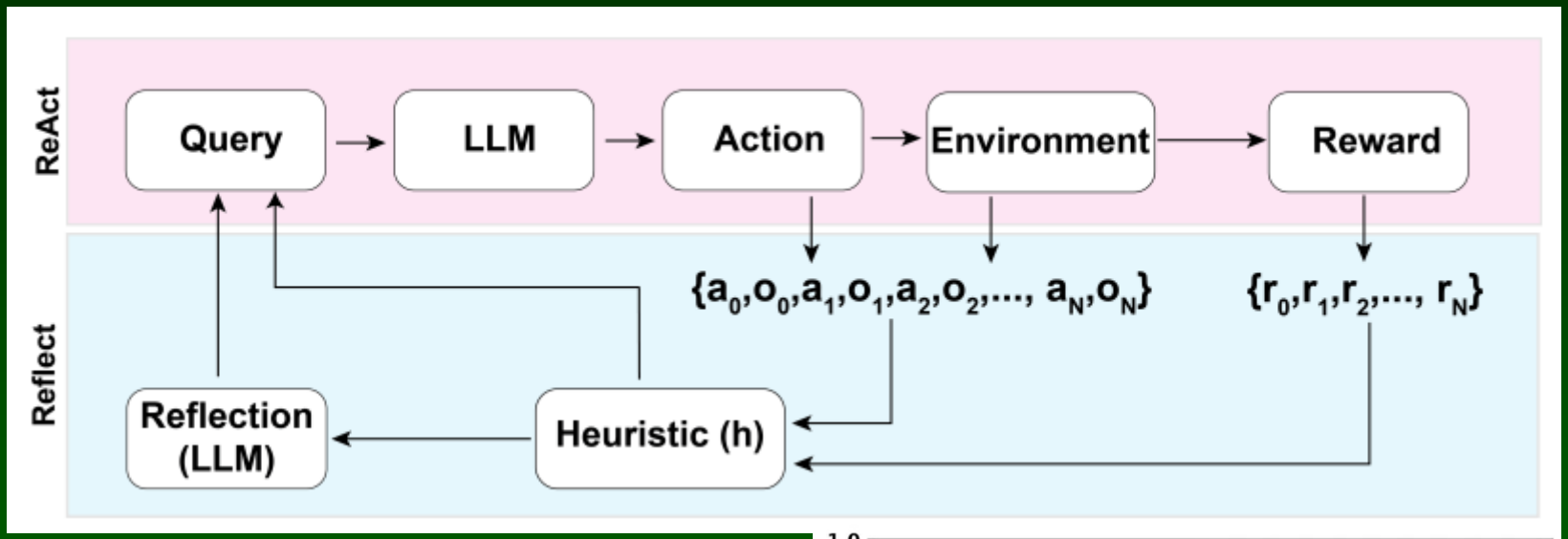
Fronto-parietal (FPN) theory of intelligence: flexible hub for cognitive control and adaptive implementation of task demands (Cole et al. 2013).

Can AI create a distributed, octopus-like brain-like structure?


PAC_IMtest_inverse_circle_coh_8.0-12.0hz_vmin0.7
0.00
1.00
0.95
0.90
0.85
0.80
0.75
0.70

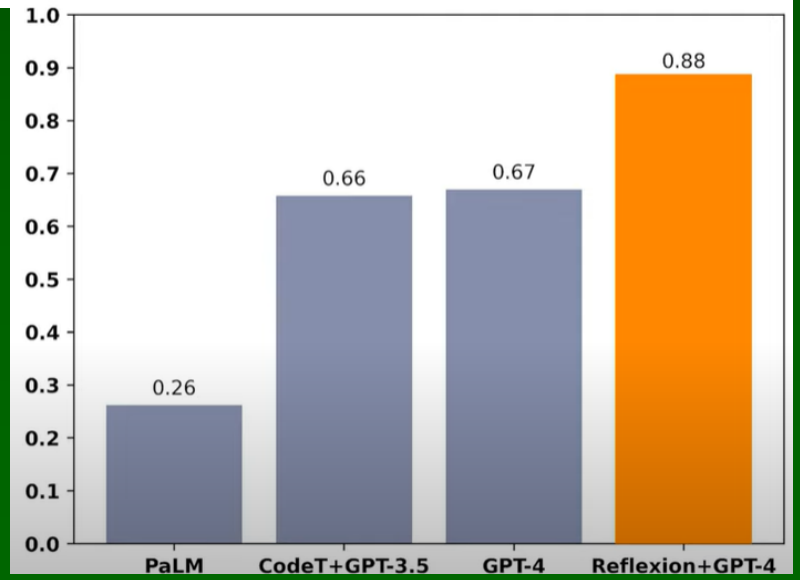
Reflexion

GPT-4 can correct itself given helpful prompts.
Autonomous agent needs dynamic memory and self-reflection loop.



Just think twice … reflection helps.

HumanEval results on 164 original programming problems.
Adding reflexion works very well.



AI self-improvement: arXiv:2212.08073
Tree of thoughts: arXiv:2305.10601

GPT-4 with plugins

Plugins to ChatGPT: give GPT tools!

- GPT-4+CodeIntepreter, **Wolfram language,** Expedia, FiscalNote, Instacart, KAYAK, Klarna, Milo, OpenTable, Shopify, Slack, Speak, Zapier.
- **Web browser**: model that knows when and how to browse the Internet, allowing language models to read information from the Internet. Expands the content of discussion, going beyond the training corpus to current information from the web.
- **Code interpreter**. Use GPT programming skills to provide a much more natural interface to most fundamental capabilities of our computers.
- It can make completely new workflows effortless and efficient, open the benefits of programming to new audiences.
- Initially: Solving mathematical problems, both quantitative and qualitative; data analysis/visualization; converting files between formats.

Hugging Face

Home of Machine Learning Tools. Can LLM like GPT4 use these tools?

Acting a bit like distributed brain parts that solve problems together.

Shen, Y et al. (2023). *HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace* arXiv:2303.17580, 20.03.2023

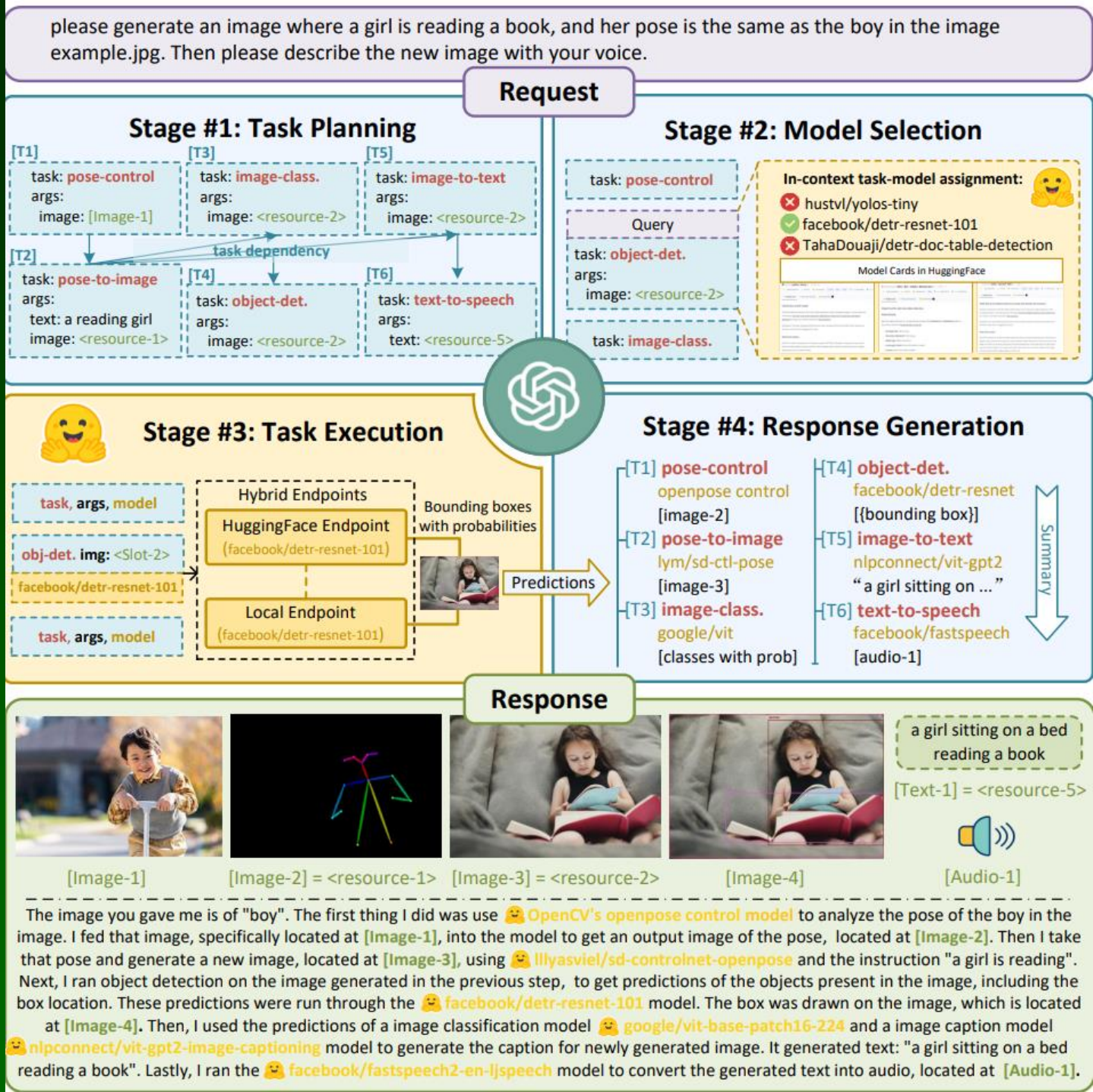
Query: Given a collection of images A: /examples/a.jpg, B: /examples/b.jpg, C: /examples/c.jpg, please tell me how many zebras in these pictures?

Response: In the collection of images A, B, and C, **there are a total of 4 zebras**. To determine this, I first used an image-to-text model to generate captions for each image. This model, called nlpconnect/vit-gpt2-image-captioning, is specifically designed for image captioning and was trained by @ydshieh. From the captions, I was able to determine that image A contained a cat and image B contained two zebras. Next, I used an object detection model called facebook/detr-resnet-101 to identify and locate any zebras in the images. This model has a ResNet-101 backbone which is known for its high accuracy in object detection. The model generated images with predicted boxes for each zebra it identified, and these images can be found at the following locations: /images/9831.jpg, /images/be11.jpg. Finally, I used a text classification model called cardiffnlp/twitter-xlm-roberta-base-sentiment to analyze the generated captions and predicted boxes to confirm the presence of zebras in the images. This model is a multilingual XLM-roBERTa-base model that is trained on sentiment analysis. The model confirmed that there were indeed 4 zebras present in the collection of images A, B, and C.

HuggingGPT

More than 5,000 organizations are using Hugging Face

Steps: planning, model selection, task execution, response generation, explanation and description of all steps.



AutoGPT

The rise of autonomous AI.

Imagica

Create any AI software just by describing it in general way.
From idea to product at the speed of thought.

Give GPT-4 a mission and it will come up with its own prompts,
and create many agents that will criticize and correct themselves.

LLM acts as a controller, uses many expert models and tools in
an automated way to achieve a given goal as autonomously as possible.
Auto-GPT recruits text-to-speech, coding and other tools.

5 major developments were created within the 48 hours after Auto-GPT
was announced, like Baby-AGI, version of the original Task-Driven
Autonomous Agent. Apps can use speech and voice, using Jarvis (HuggingGPT)

Torantulino/Auto-GPT: An experimental open-source attempt to make GPT-4
fully autonomous.

MemoryGPT can permanently store previous conversations and remember
topics the next time you ask. GPT-4 can remember about 50 pages.

Auto-GPT “God mode”

Self-Prompting, the latest trend in large language models.

Language models can be more than text generators:

- generating and executing their own prompts, linking with existing software tools, and creating new software on demand, they guide themselves to a given goal;
- language models can be used as universal personal assistants, or to control various types of robots.

The program plans step-by-step, justifies decisions and develops plans, documenting all steps.
In addition to integrating GPT-4 for text generation, the system has Internet access for data retrieval, can store data, and can generate speech via the Elevenlabs API.

The system is even **capable of self-improvement** and bug-fixing by generating Python scripts via GPT-4.

ChaosGPT show how people may create malicious goal-seeking models.

Auto-GPT example

Goals: Invent an original and out-of-the-box recipe to suit current event, such as Eastern. Save recipe to a file; shut down upon achieving your goals.

Abbreviated version, follow the action here.

CHEF-GPT thoughts: I will search for upcoming events to find suitable one …

Reasoning: It will help me to come up with a relevant and exciting recipe.

Google search sites listing events … then browse the site, summarizing site "… guide to events … links …". No specific events found.
Browse second site, choose April events, nothing suitable.
Find website with calendar view of events, lists examples of events.
Selects Earth Day on April 22, 2023.

Earth Day is widely celebrated and will be fitting for creating thematic recipe.
Use GPT agent to create recipe, incorporating sustainable and eco-friendly ingredients.
Adding reasoning + criticism, and creating:

Earth Day Quinola Salad. Ingredients … Done.

JARVIS

JARVIS can create web applications listing to spoken instructions.

- I need a next app with Tailwind.
- J: you want me to create a new next JS app with Tailwind CSS. What would you like the app to do?
- Let's make a website called McKay’s landing page and it needs to have a single text input that's an email input so that I can capture emails from my email list.
- J: you want a site called McKay's landing page that takes in an email input would you like me to add anything else …
- Um let's make it mobile and desktop friendly and let's style it with like a blue indigo gradient.
- J: okay I'll make it responsive and style it with a nice background gradient anything else I initialized your project continuing I checked to make sure the app works correctly everything looks good. Would you like me to deploy it - yes
- so far so good I deployed your project
- okay this should work oh my god, it did it!

AutoGPT good and bad

- Do market research for waterproof shoes.
- Find top 5 competitors, list their pros and cons.

AutoGPT made a plan: find brands, check their sites, reviews, ratings, evaluate how trustworthy are reviews of the sites.

Provided a detailed report of all aspects, in 8 minutes.

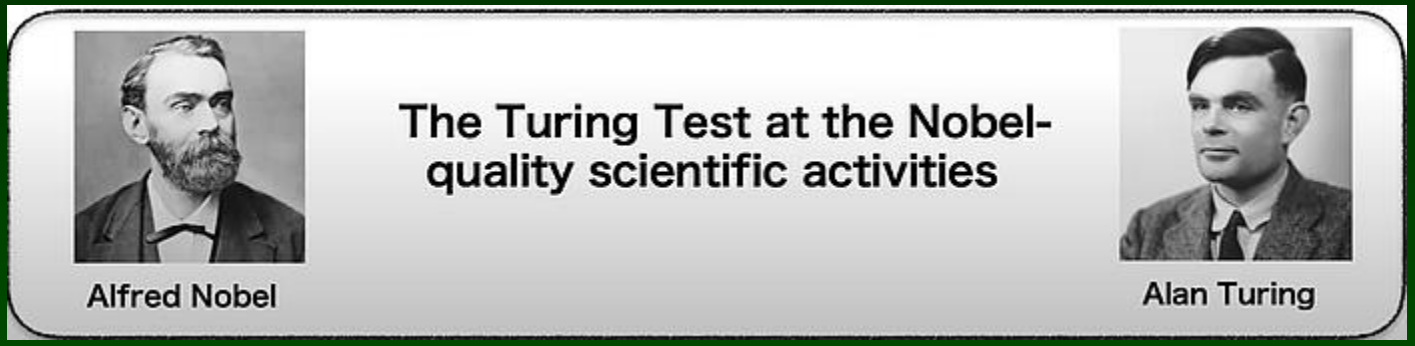
Ask GPT-4 to take over Twitter and outsmart @elonmusk.

It proposed a masterplan “OperationTweetstorm”, challenge of Tweet-off showdowns, develop LLM to power an army of diverse personas to send tweets, assemble a team “Tweet Titans” of hackers to attack Twitter backend, manipulate Twitter recommendations, hijacking Elon’s account, create viral hashtags that align with the masterplan, capitalize on chaos.

AgentGPT Beta: assemble, configure, and deploy autonomous AI Agents in your browser. This becomes too easy …

AI for Science?

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



- A grand challenge aiming 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 development of in-depth understanding on the process of scientific discoveries. Make a closed-loop system from knowledge acquisition and hypothesis generation and verification to full automation of experiments and data analytics.
- 3rd Nobel Turing Challenge Initiative Workshop, July 11-12, 2023, **Carnegie Mellon University, Pittsburgh, PA**
- **Nobel Turing Challenge YouTube channel.**

AI minds and human brains

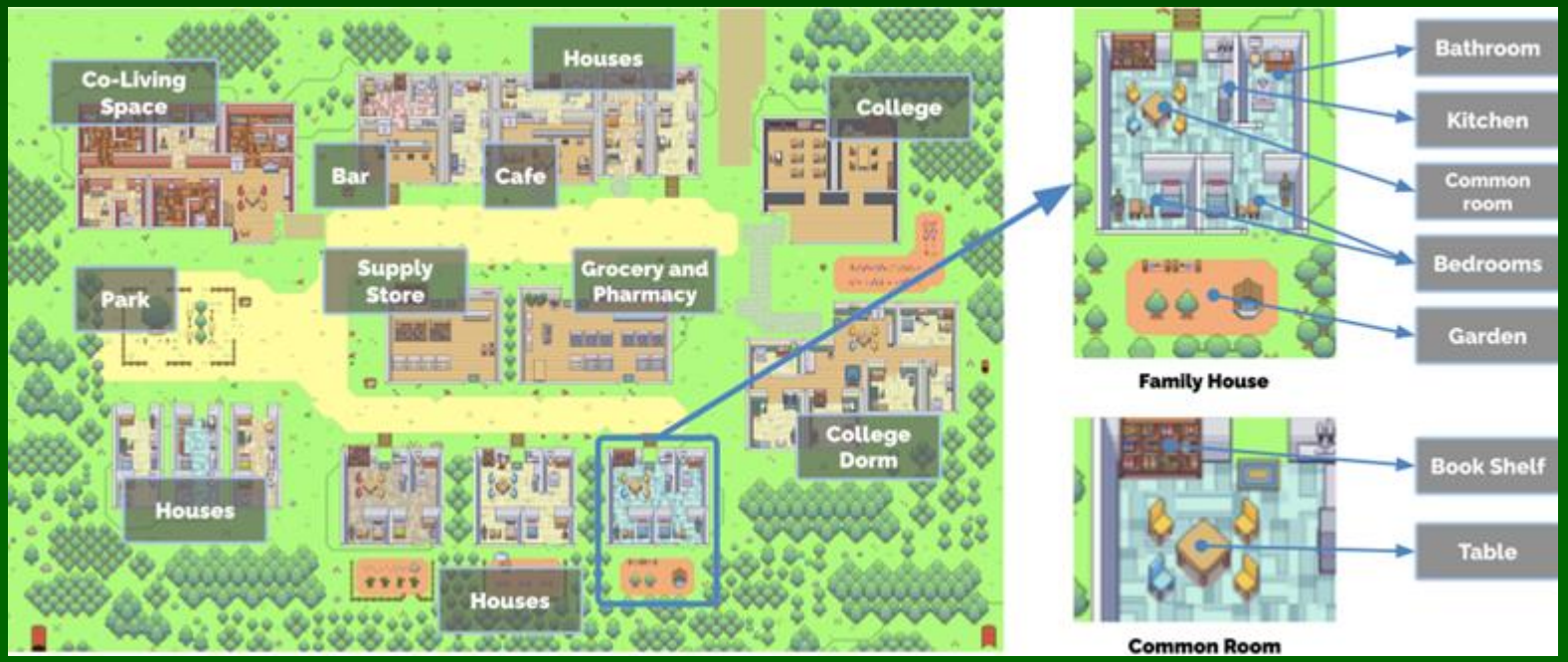
Interactive Simulacra

Park et al. (2023). Generative Agents: Interactive Simulacra of Human Behavior (arXiv:2304.03442). Created small village …

25 GPT-3.5 agents, providing believable proxies of human behavior, from immersive environments to interpersonal communication.

Agents remember, retrieve, reflect, talk and interact with other agents, plan activities in dynamically evolving circumstances.

Added support for longer-term agent coherence, ability to manage dynamically-evolving memory, recursively produce more generations.



LLM Persona?

Ilya Sutskever, CEO OpenAI: maybe we are reaching the point when language of psychology is appropriate to describe the behavior of our models.

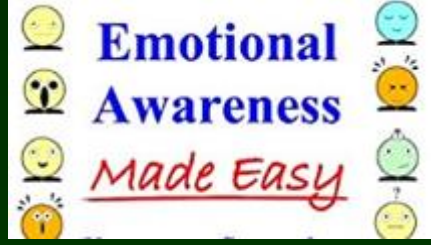
Terry Sejnowski, a distinguished professor at UC San Diego, Francis Crick Chair at the Salk Institute.

“Language models, like ChatGPT, take on personas.
The persona of the interviewer is mirrored back”.
… When I talk to ChatGPT it seems as though another neuroscientist is talking back to me. It’s fascinating and sparks larger questions about intelligence.”

The Mirror of Erised reflects the deepest desires of those that look into it, never yielding knowledge or truth, reflecting what it believes the onlooker wants to see (*Harry Potter*).

Can LLM become sentient, conscious personas?
We are desperate to find arguments against such possibility.

Emotion awareness



Many attempts to understand psychology of LLMs!

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) to 20 scenarios, and compared them with the general population norms.

Responses were evaluated by licensed psychologists.

ChatGPT demonstrated significantly higher performance than average human.

This experiment was repeated one month later to measure ChatGPT emotional awareness improvement over time. No embodiment was needed!

ChatGPT's performance accuracy levels were extremely high (9.7/10).

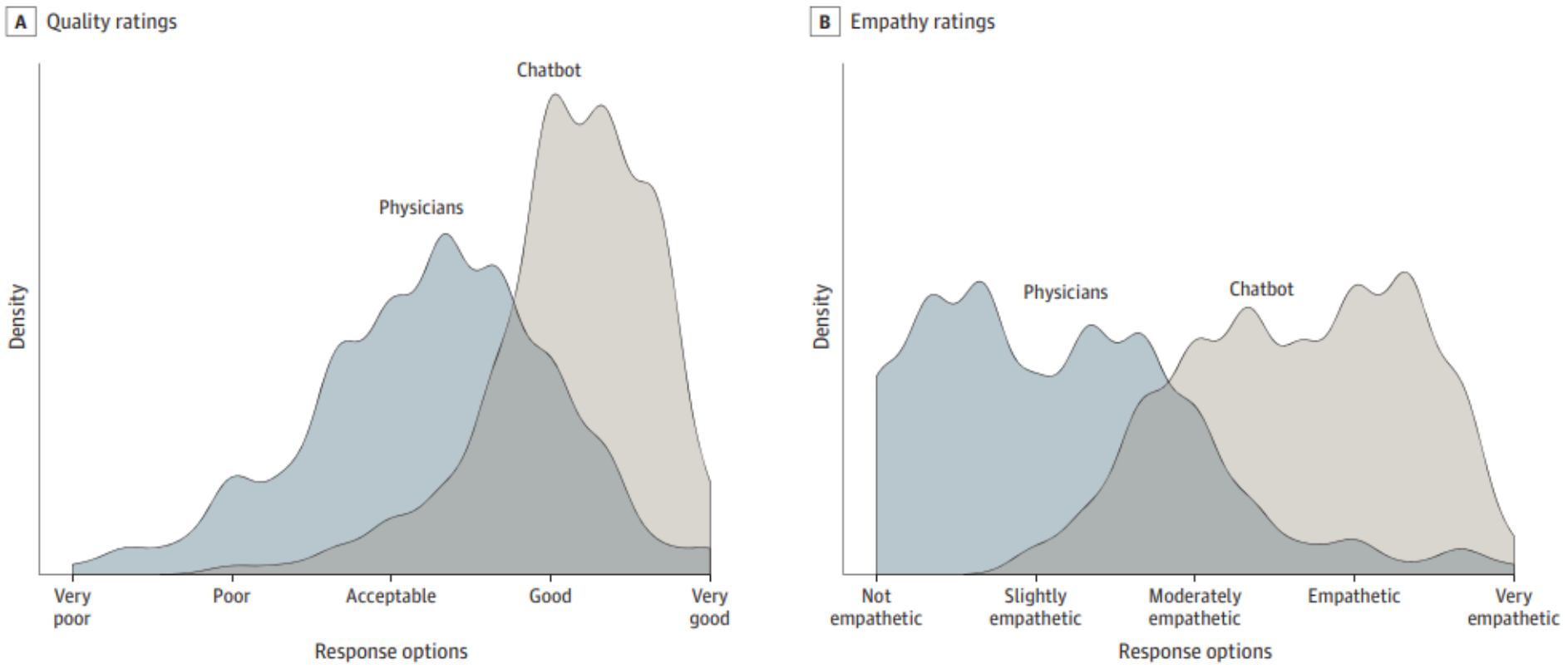
Implications: ChatGPT can be used as part of cognitive training for clinical populations with EA impairments.

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

Chatbots vs Physicians

Ayers, J. W. … & Smith, D. M. (2023). Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine* (online April 28, 2023).

Figure. Distribution of Average Quality and Empathy Ratings for Chatbot and Physician Responses to Patient Questions



Kernel density plots are shown for the average across 3 independent licensed health care professional evaluators using principles of crowd evaluation. A, The overall quality metric is shown. B, The overall empathy metric is shown.

Conscious avatars?

Hal talks with Sophia about what it means to be conscious (GPT-3).

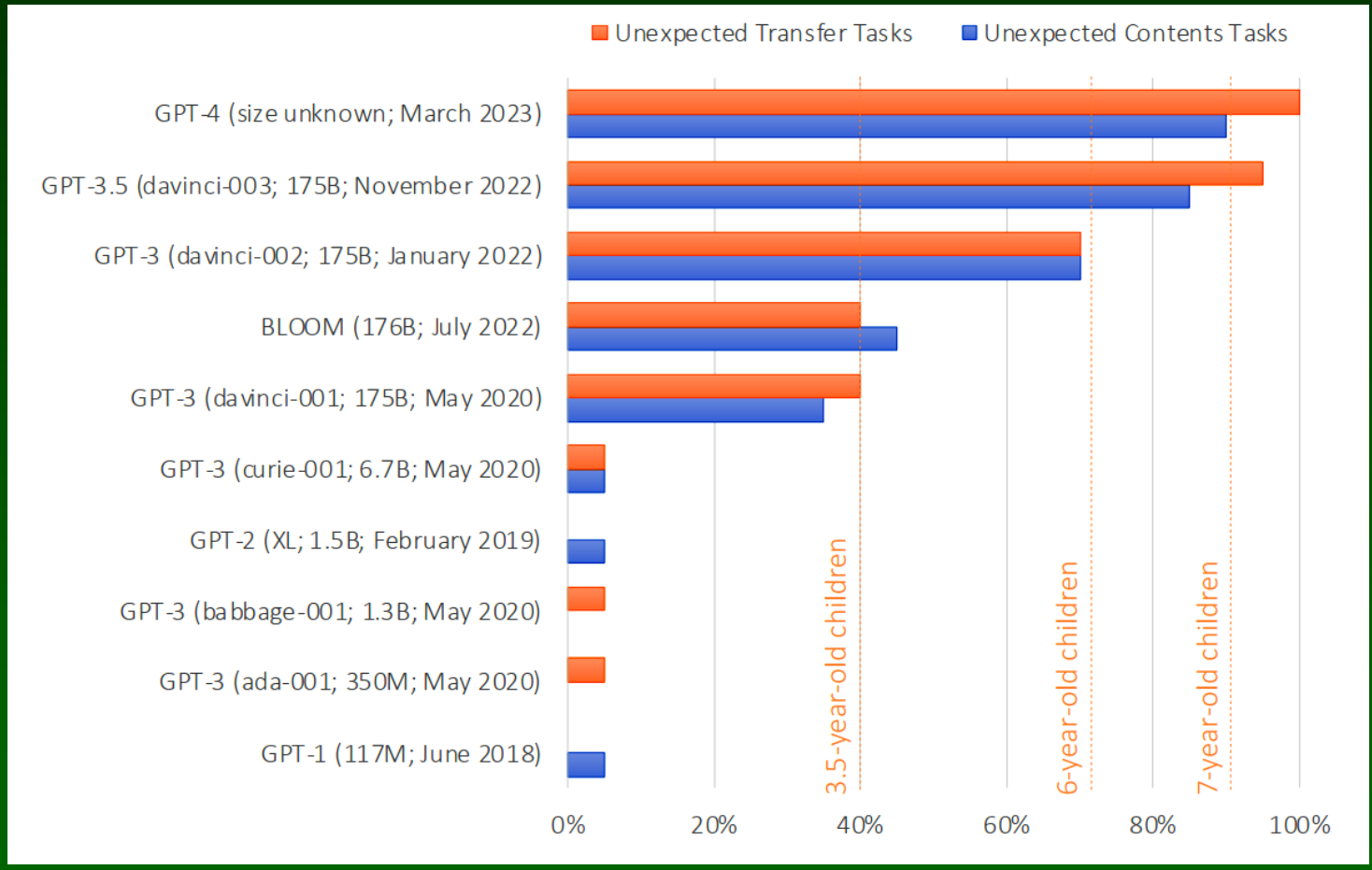
LLMs seem to understands the meaning of questions.
LaMDA has convinced Google engineer Blake Lemoine that it is conscious.

John Locke, 333 years ago defined consciousness as "a perception of what appears in our mind“, a description of internal images/states. If so, than can GPT have mental images in its neural network, a model of the world?

For example, can it have a Theory of Mind?

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).



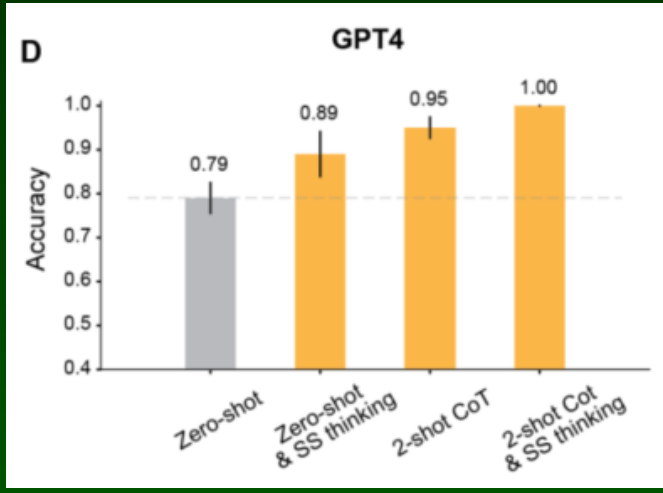
Language learning, over and above social experience, drives the development of a mature theory of mind (Pyers, Senghas, Psychological Science, 2010).

Social cognition

ToM, social cognition, understanding false beliefs, understanding human emotions, other people … AI is better than humans in social cognition!

Zero-shot is a simple question-answer.

Few-shot task are based on priming associative memory, no learning (chain of thoughts, CoT). LLMs improve complex reasoning performance also when language models are instructed to think “step-by-step”(SS).



Human accuracy on ToM tests is 87%.
GPT-4 in zero-shot settings shows nearly 80% accuracy. arXiv:2304.11490
With prompts for in-context learning GPT-4 gets 100% in all scenarios.

Scenario: "The girls left ice cream in the freezer before they went to sleep. Over night the power to the kitchen was cut and the ice cream melted.“
Q: When they get up, do the girls believe the ice cream is melted?

A: … the answer is: No, the girls don't believe the ice cream is melted.

Why consciousness arise?

Harvard/MIT group trained GPT model on millions of matches of the board game Othello, feeding in long sequences of moves in the text form.

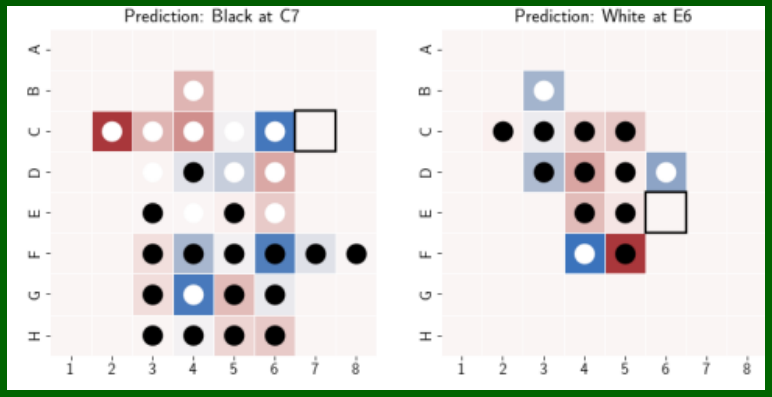
- Li, K. et al. (2023). *Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task* (arXiv:2210.13382).

Latent "neural activity" matched the representation of an Othello game board, in a convoluted form. 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.
This is similar to the processes in our brains, so can we reject the idea that it has consciousness?

Can it be sentient?
Can psychological pain be disembodied?



GAIA and moral AI?

The Global Artificial Intelligence Association (GAIA).

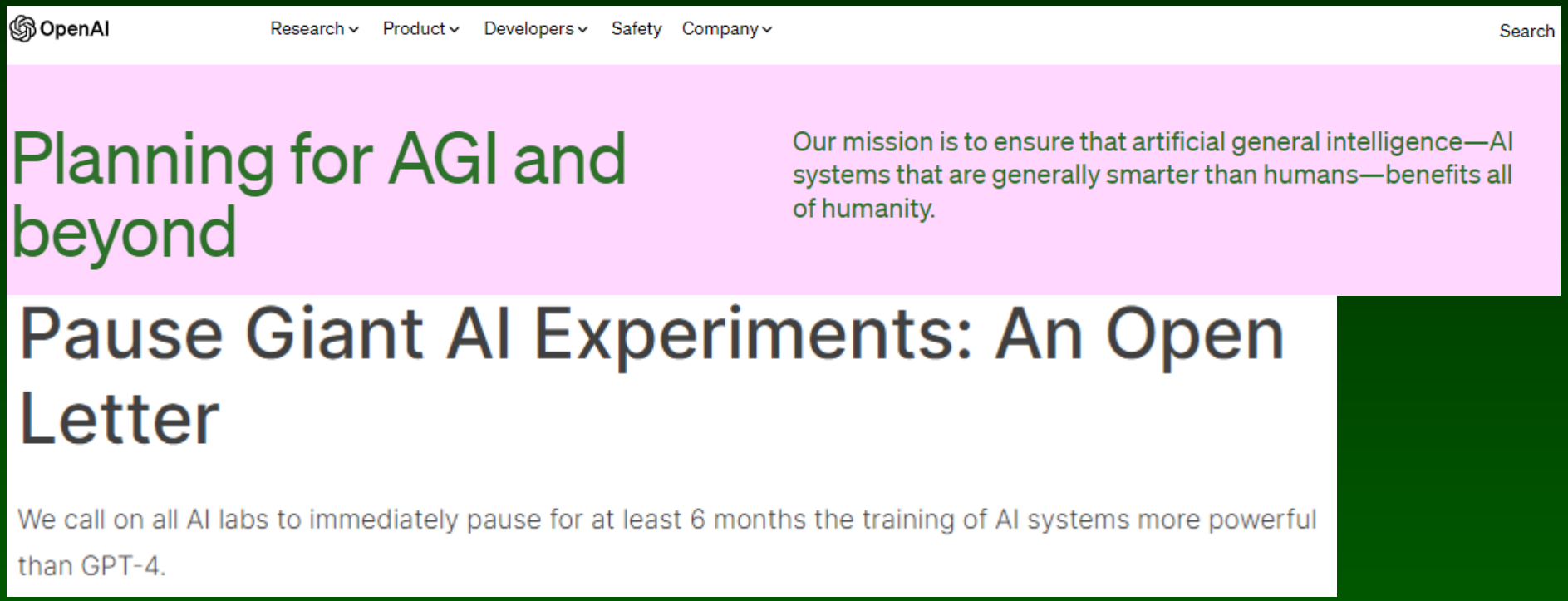
How to create an environment, a platform, a solution, which uses gamification and flow state to teach AI in safe ways, things like:

- ✓ positive and moral behavior, pro-social behavior and cooperation;
- ✓ responsibility and self-esteem, sustainability;
- ✓ understanding art and science.

GAIA is a multidisciplinary start-up which uses collective creativity to **develop artificial intelligence with compassion**. Our team of scientists, entrepreneurs, activists and philosophers globally research AI's social, sociological, legal, technological and ethical aspects of AI.

- The prize is sponsored by Dictador, the only company that has a robot as the CEO. The total prize pool for the competition is EUR 200,000.

Fears …



32 000 signatures as of 31.05.2023.

- Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable.
- Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.
- Time Magazine: The Only Way to Deal With the Threat From AI? Shut It Down

Towards Human-like Intelligence

IEEE Computational Intelligence Society Task Force,
Towards Human-like Intelligence

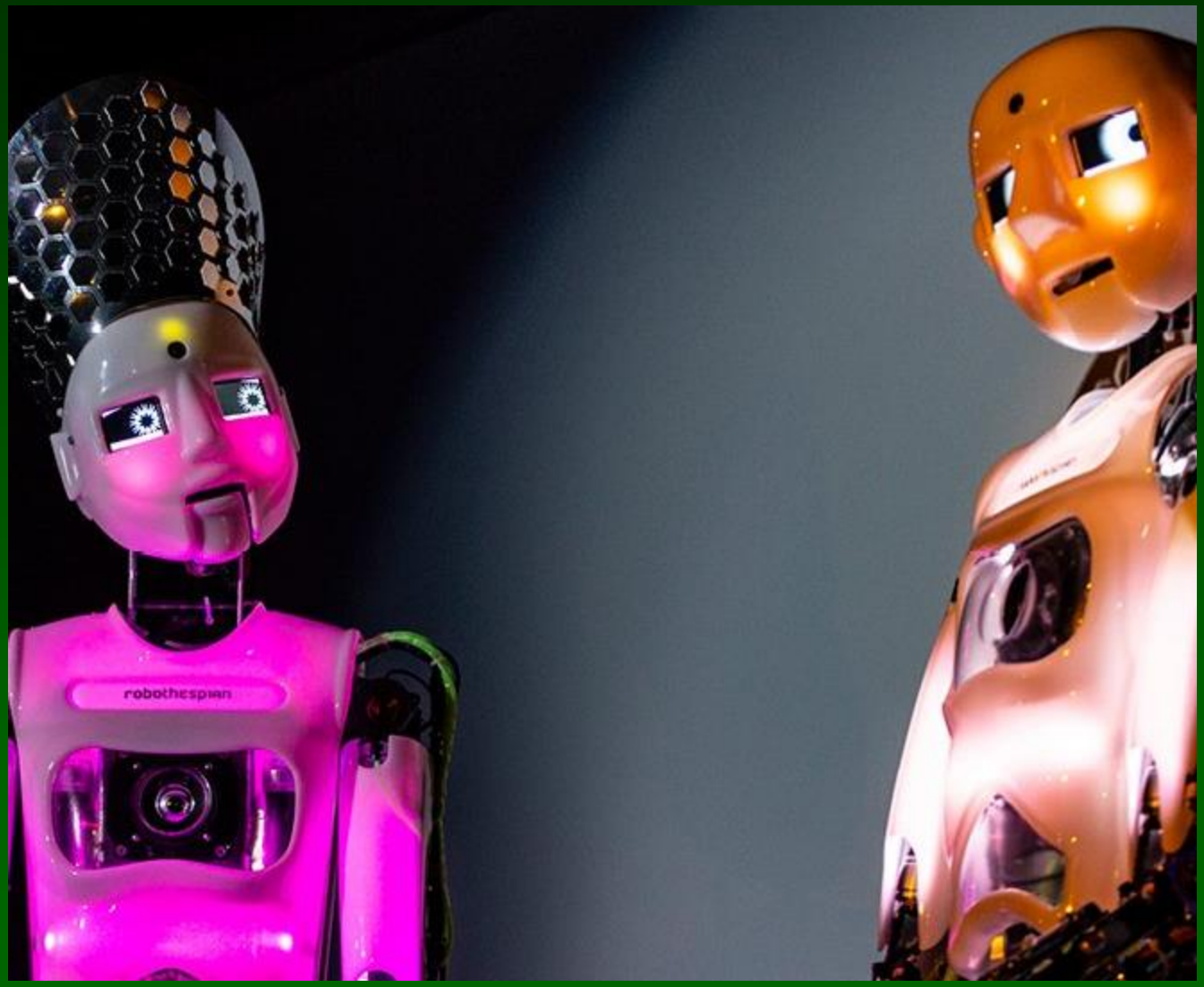
IEEE SSCI CIHLI 2023 Symposium on Computational Intelligence for Human-like Intelligence, Mexico City (J. Mandziuk, W. Duch, M. Woźniak).

AGI conference, Journal of Artificial General Intelligence, comments on Cognitive Architectures and Autonomy: A Comparative Review (eds. Tan, Franklin, Duch).

BICA Annual International Conf. on Biologically Inspired Cognitive Architectures, 13th Annual Meeting of the BICA Society, Guadalajara, Mexico 2023.

Brain-Mind Institute Schools International Conference on Brain-Mind (ICBM) and Brain-Mind Magazine (Juyang Weng, Michigan SU).

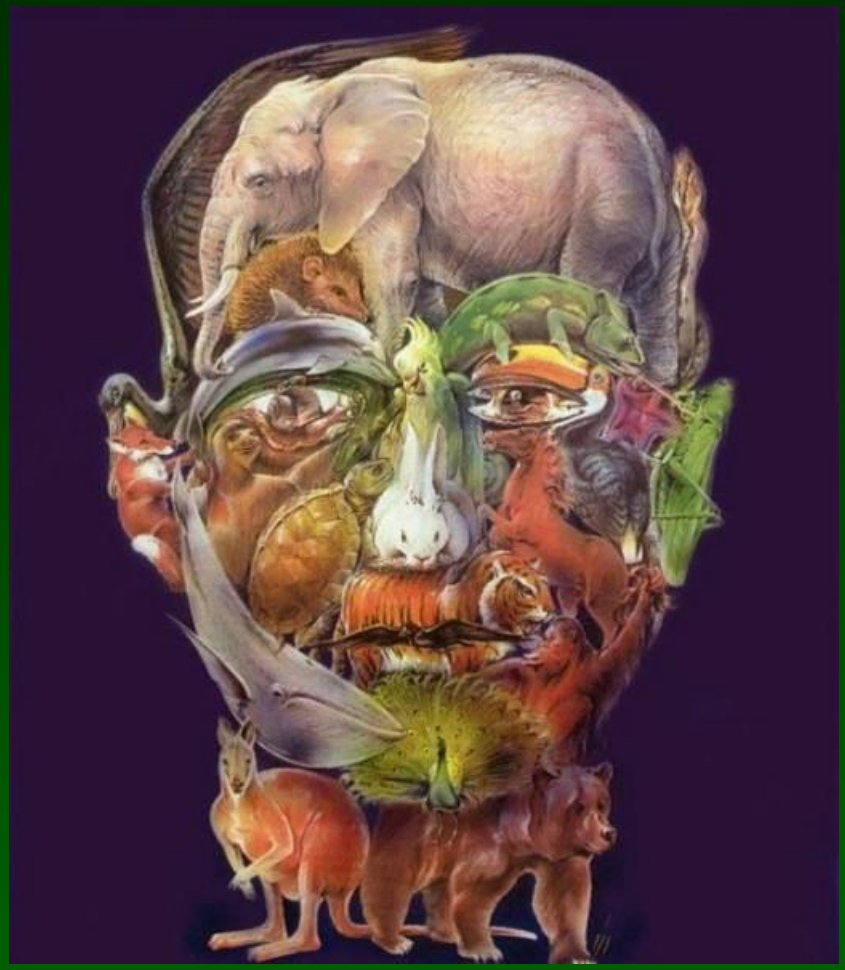
Stanislaw Lem: About prince Ferrycy and princess Crystala. Intelligent Palefaces? Is it possible?



Perspectives

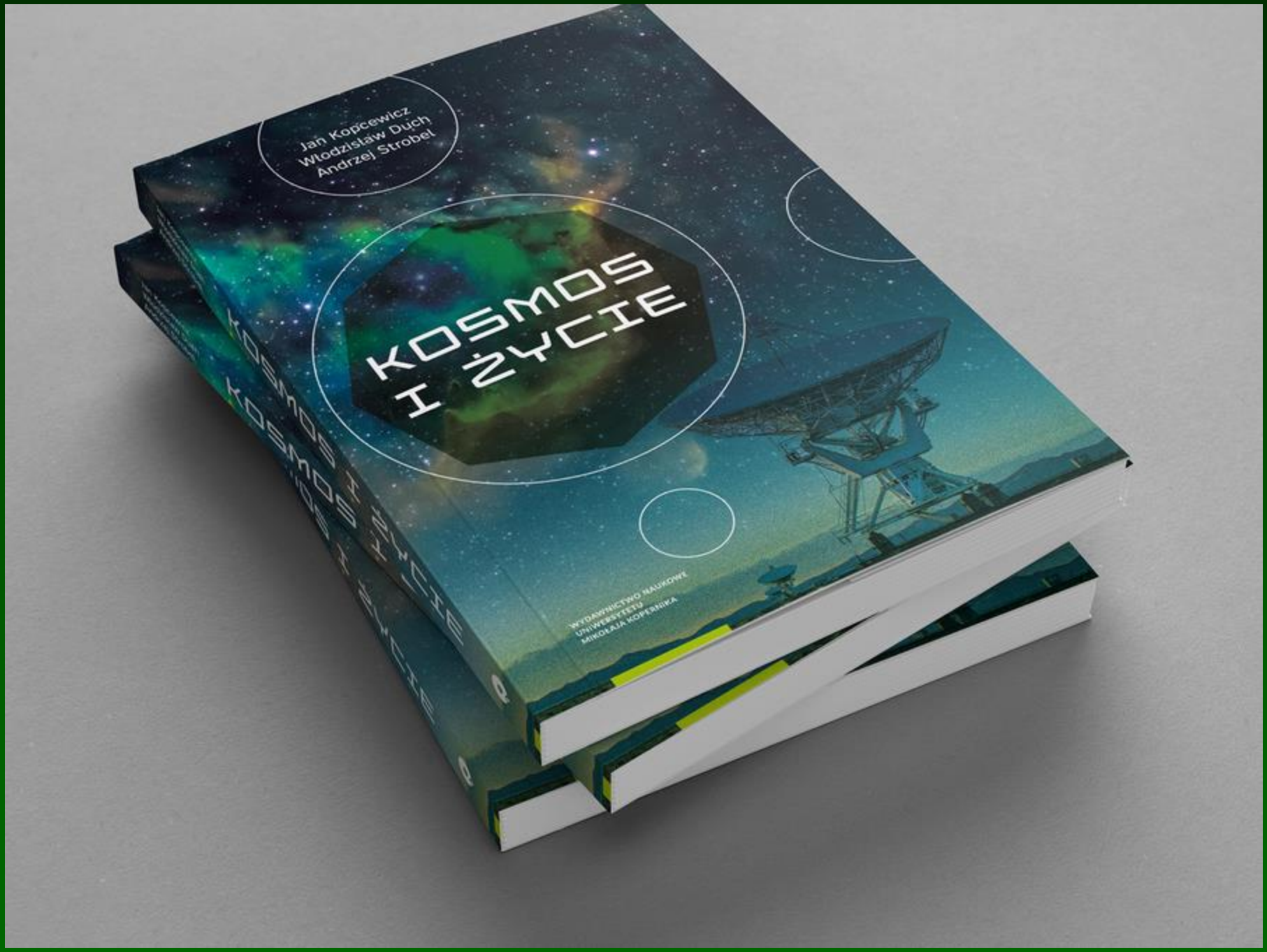
- We are close to the creation of artificial minds/distributed brains more powerful than human brains will every be.
- LLMs can control autonomous AI systems. Robots/AI systems may accumulate knowledge and skills very quickly.
- AI understanding of human minds creates a great potential for manipulation. Already some people are emotionally attached to avatars.
- LLM auto-prompts, plugins, and use of software tools will lead to AGI.
- AI influence will be more profound than internet/social media. Programming may soon become obsolete. Big companies are at the front of AI research, but large cooperative open projects appear (Llama, Falcon-40B).
- Teach how to use AI ecosystems (lablab.ai, k4all), create educational ecosystems, collaborate with EU consortia, restructure universities …
- Consciousness in AI models is inevitable, some experts already accept it.
- Brains are too slow to be connected to computers, this is not an option.
- Our megalomania is dangerous. We are moving from animal life to our own virtual creation! Singularity may come faster than we think!

Intelligence?



Search: Wlodek Duch

=> talks, papers, lectures, Flipboard ...


Jan Kopcewicz
Włodzisław Duch
Andrzej Strobel
KOSMOS I ŻYCIE
WYDAWNICTWO NAUKOWE
UNIWERSYTETU
MIKOŁAJA KOPERNIKA

8 Examples of AI Risks

Center for AI Safety

When AI becomes more advanced, it could eventually pose catastrophic or existential risks. There are many ways in which AI systems could pose or contribute to large-scale risks, some of which are enumerated below.

1. Weaponization – most dangerous

2. Misinformation – manipulation

3. Proxy Gaming - faulty objectives

4. Enfeeblement - completely dependent on machines

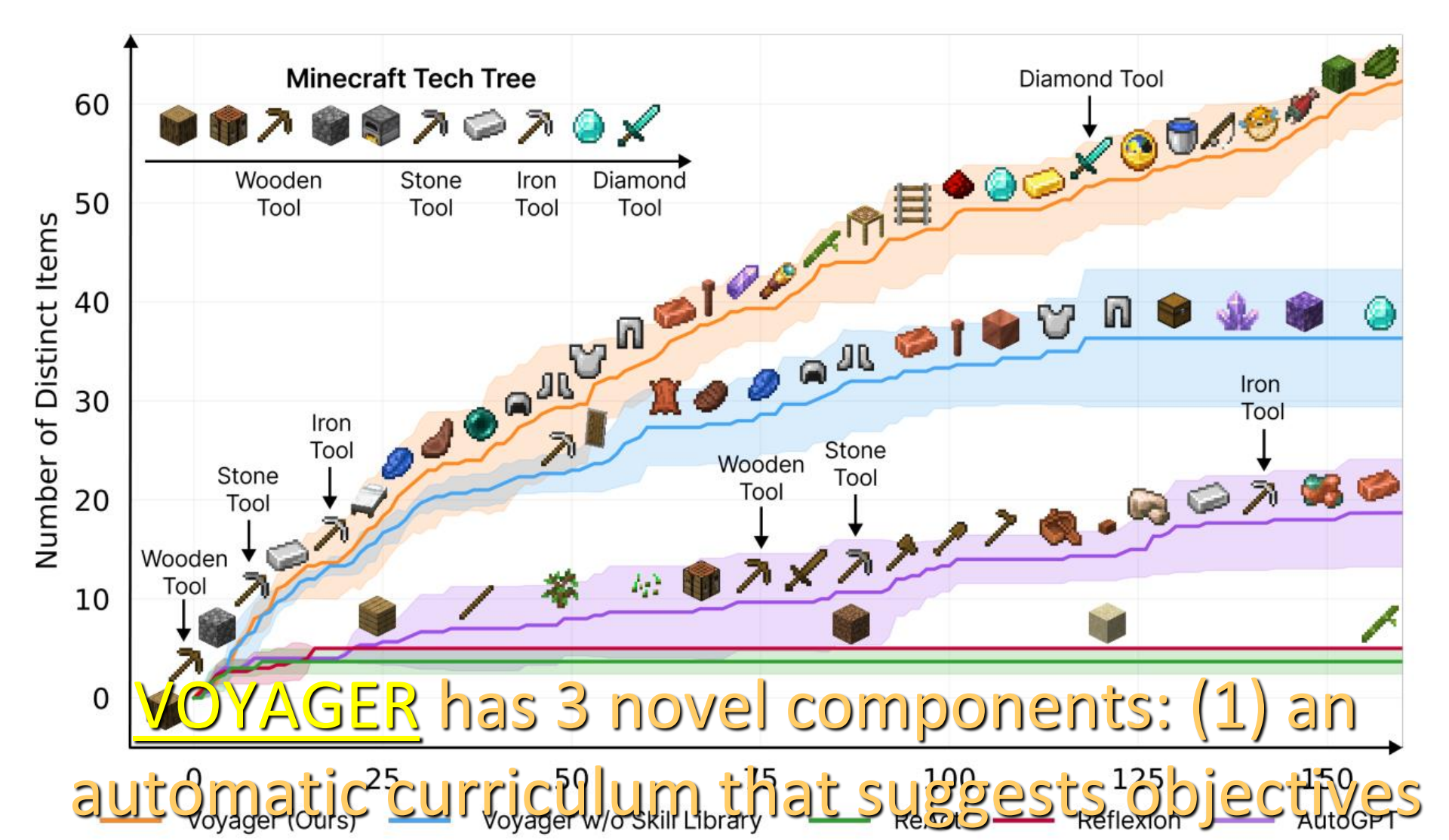
5. Value Lock-in – power to small groups

6. Emergent Goals – no idea what it can do

7. Deception – like in Diplomacy game

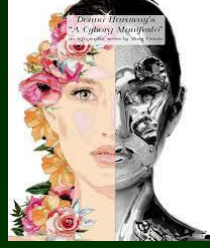
8. Power-Seeking Behavior – looking for economic advantage

For more in-depth discussion of extreme risks:
Natural Selection Favors AIs Over Humans
How Rogue AIs May Arise, Yoshua Bengio.



VOYAGER has 3 novel components: (1) an automatic curriculum that suggests objectives for open-ended exploration, (2) a skill library for developing increasingly complex behaviors, and (3) an iterative prompting

Jobs and money …



- The World Economic Forum's “The Future of Jobs Report 2020" predicts AI will replace 85 million and create 97 jobs globally by 2025. Do they really know what is happening?
- AI and automation can free us to pursue careers that give us a greater sense of meaning and well-being. The AI Job That Pays Up to $335K? Prompt engineering (Time, 14.04.23), no programming required.
- **Report by PwC** estimate – 38% of US jobs will be at high risk of automation by early 2030s. White House Report AI 2022 report.
- WD LegalTech talk (2021), Fusion AI Exterro WhatSun solution suite will eliminate up to 99.9% of all inefficient cerebral functions performed by highly paid attorneys, paralegals and technologists.
- ChatGPT: the 10 Jobs Most at Risk by AI (businessinsider.com); impact on the US labor market (OpenAI, arxiv), summary in AI explained video.
- Musicians, composers, writers want a piece of the cake. Stability AI, generative art company, sued by Getty Images for trillions of $.
- Europe: RODO may kill LLM technology. Italy blocked ChatGPT (Money.pl).