Part 2/4
AI Optimizations
AI Designing AI
AI Mind-Reading

# → Hallucinations are (almost) all you Need

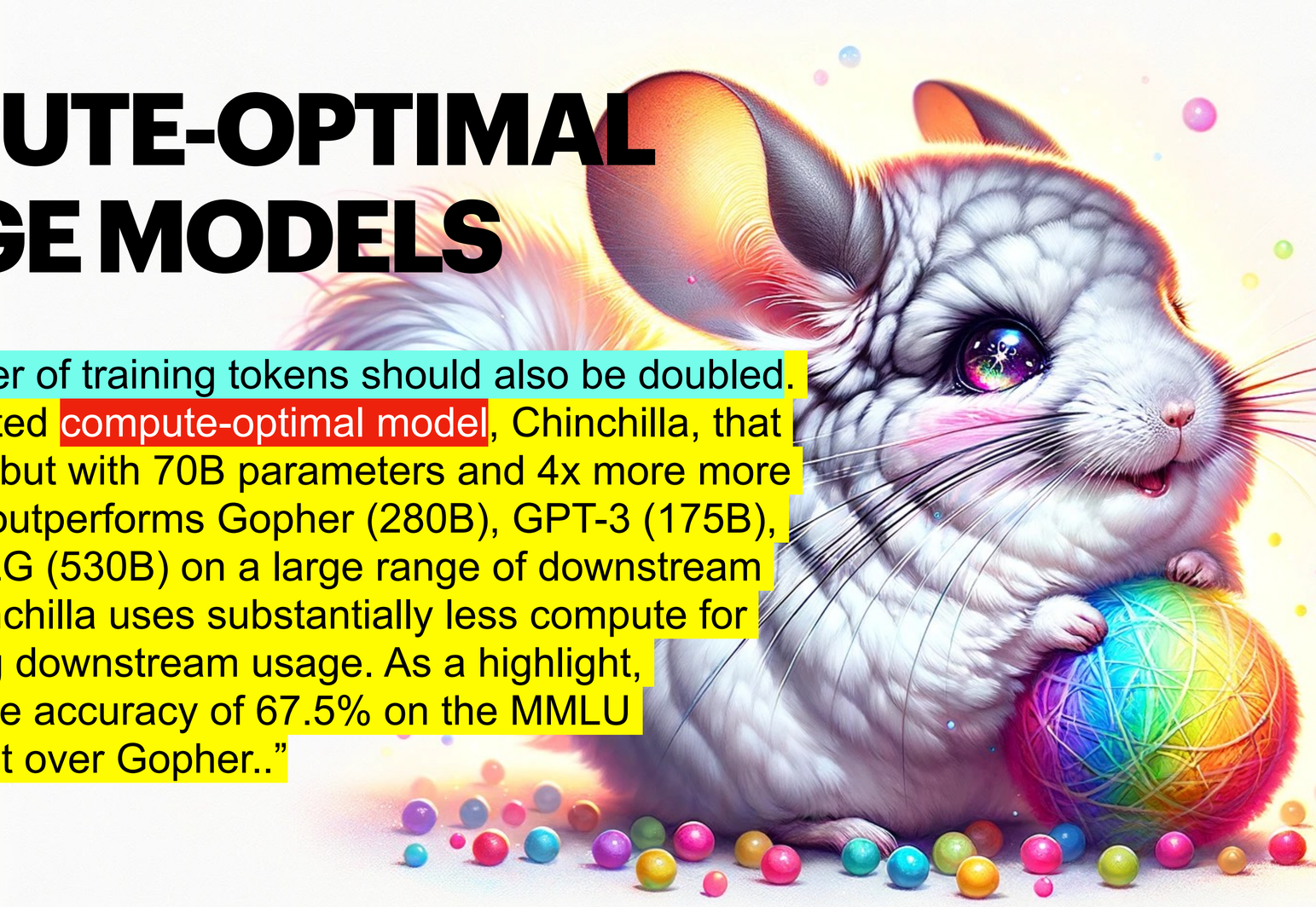
JHAVE @ GLIA.CA

CENTRE FOR DIGITAL NARRATIVE
UNIVERSITY OF BERGEN

This rapid *artistic* **overview of key scientific AI examples that covers a year** (loosely defined as starting with GPT-4 on March 14th, 2023) is framed by **the hypothesis that fundamental research in science is being transformed by a practice predominantly associated with the arts: namely hallucinations**.


→ AI Optimizations
Hallucinated matrix-math, preferences, prompts, benchmarks, corpuses.

# CHINCHILLA: TRAINING COMPUTE-OPTIMAL LARGE LANGUAGE MODELS

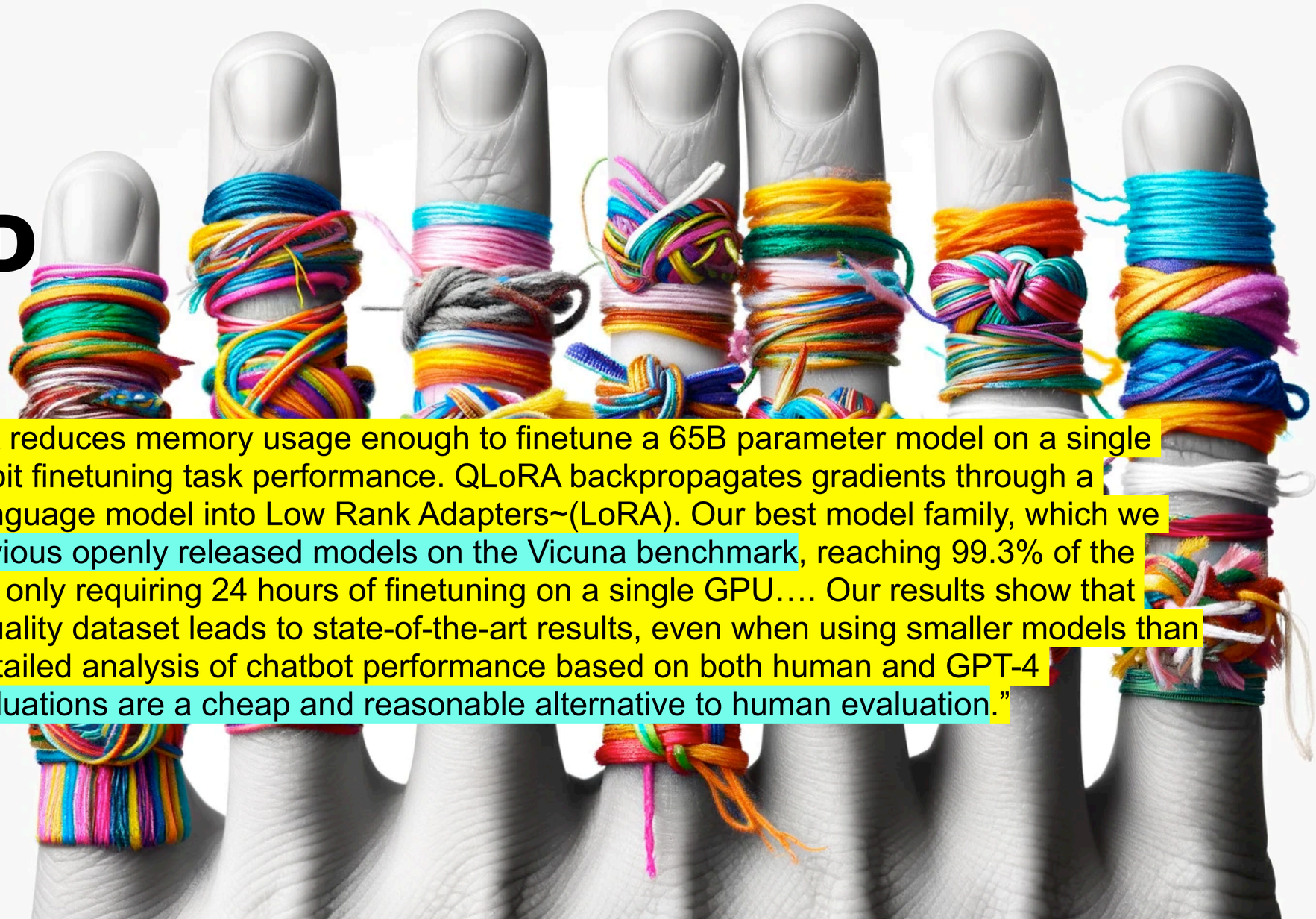


“for every doubling of model size the number of training tokens should also be doubled. We test this hypothesis by training a predicted compute-optimal model, Chinchilla, that uses the same compute budget as Gopher but with 70B parameters and 4x more more data. Chinchilla uniformly and significantly outperforms Gopher (280B), GPT-3 (175B), Jurassic-1 (178B), and Megatron-Turing NLG (530B) on a large range of downstream evaluation tasks. This also means that Chinchilla uses substantially less compute for fine-tuning and inference, greatly facilitating downstream usage. As a highlight, Chinchilla reaches a state-of-the-art average accuracy of 67.5% on the MMLU benchmark, greater than a 7% improvement over Gopher..”

→ AI Optimizations

**DeepMind**. 29 Mar 2022

# QLORA: EFFICIENT FINETUNING OF QUANTIZED LLMS



“an efficient finetuning approach that reduces memory usage enough to finetune a 65B parameter model on a single 48GB GPU while preserving full 16-bit finetuning task performance. QLoRA backpropagates gradients through a frozen, 4-bit quantized pretrained language model into Low Rank Adapters~(LoRA). Our best model family, which we name Guanaco, outperforms all previous openly released models on the Vicuna benchmark, reaching 99.3% of the performance level of ChatGPT while only requiring 24 hours of finetuning on a single GPU…. Our results show that QLoRA finetuning on a small high-quality dataset leads to state-of-the-art results, even when using smaller models than the previous SoTA. We provide a detailed analysis of chatbot performance based on both human and GPT-4 evaluations showing that GPT-4 evaluations are a cheap and reasonable alternative to human evaluation.”

→ AI Optimizations

**Arxiv**. 23 May, 2023


# QLoRA: Efficient Finetuning of Quantized LLMs

**Tim Dettmers*** **Artidoro Pagnoni*** **Ari Holtzman**

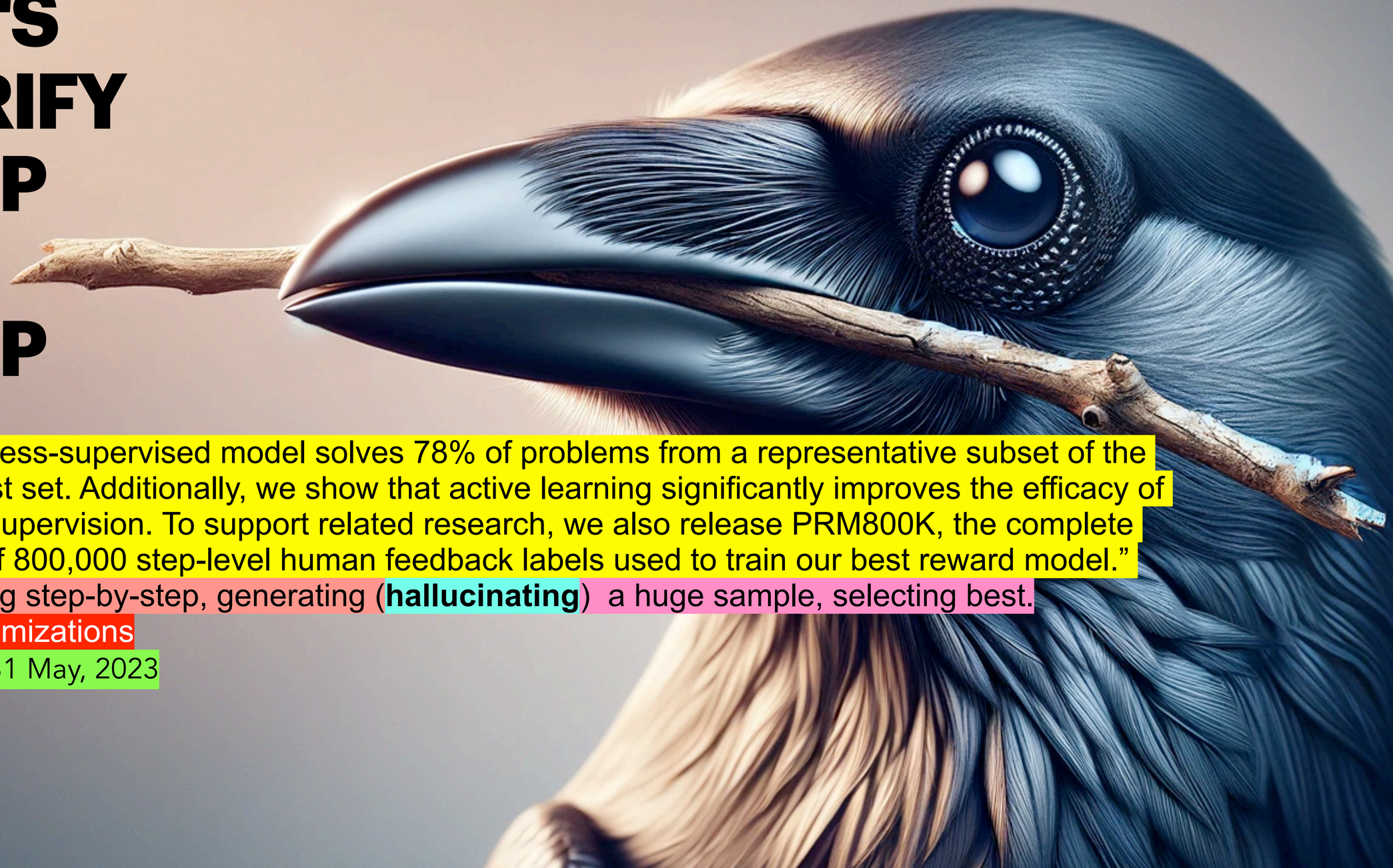
**Luke Zettlemoyer**

University of Washington
{dettmers,artidoro,ahai,lsz}@cs.washington.edu

## Abstract

We present QLoRA, an efficient finetuning approach that reduces memory usage enough to finetune a 65B parameter model on a single 48GB GPU while preserving full 16-bit finetuning task performance. QLoRA backpropagates gradients through a frozen, 4-bit quantized pretrained language model into Low Rank Adapters (LoRA). Our best model family, which we name **Guanaco**, outperforms all previous openly released models on the Vicuna benchmark, reaching 99.3% of the performance level of ChatGPT while only requiring 24 hours of finetuning on a single GPU. QLoRA introduces a number of innovations to save memory without sacrificing performance: (a) 4-bit NormalFloat (NF4), a new data type that is information theoretically optimal for normally distributed weights (b) Double Quantization to reduce the average memory footprint by quantizing the quantization constants, and (c) Paged Optimizers to manage memory spikes. We use QLoRA to finetune more than 1,000 models, providing a detailed analysis of instruction following and chatbot performance across 8 instruction datasets, multiple model types (LLaMA, T5), and model scales that would be infeasible to run with regular finetuning (e.g. 33B and 65B parameter models). Our results show that QLoRA finetuning on a small high-quality dataset leads to state-of-the-art results, even when using smaller models than the previous SoTA. We provide a detailed analysis of chatbot performance based on both human and GPT-4 evaluations showing that GPT-4 evaluations are a cheap and reasonable alternative to human evaluation. Furthermore, we find that current chatbot benchmarks are not trustworthy to accurately evaluate the performance levels of chatbots. A lemon-picked analysis demonstrates where **Guanaco** fails compared to ChatGPT. We release all of our models and code,

# LET'S VERIFY STEP BY STEP

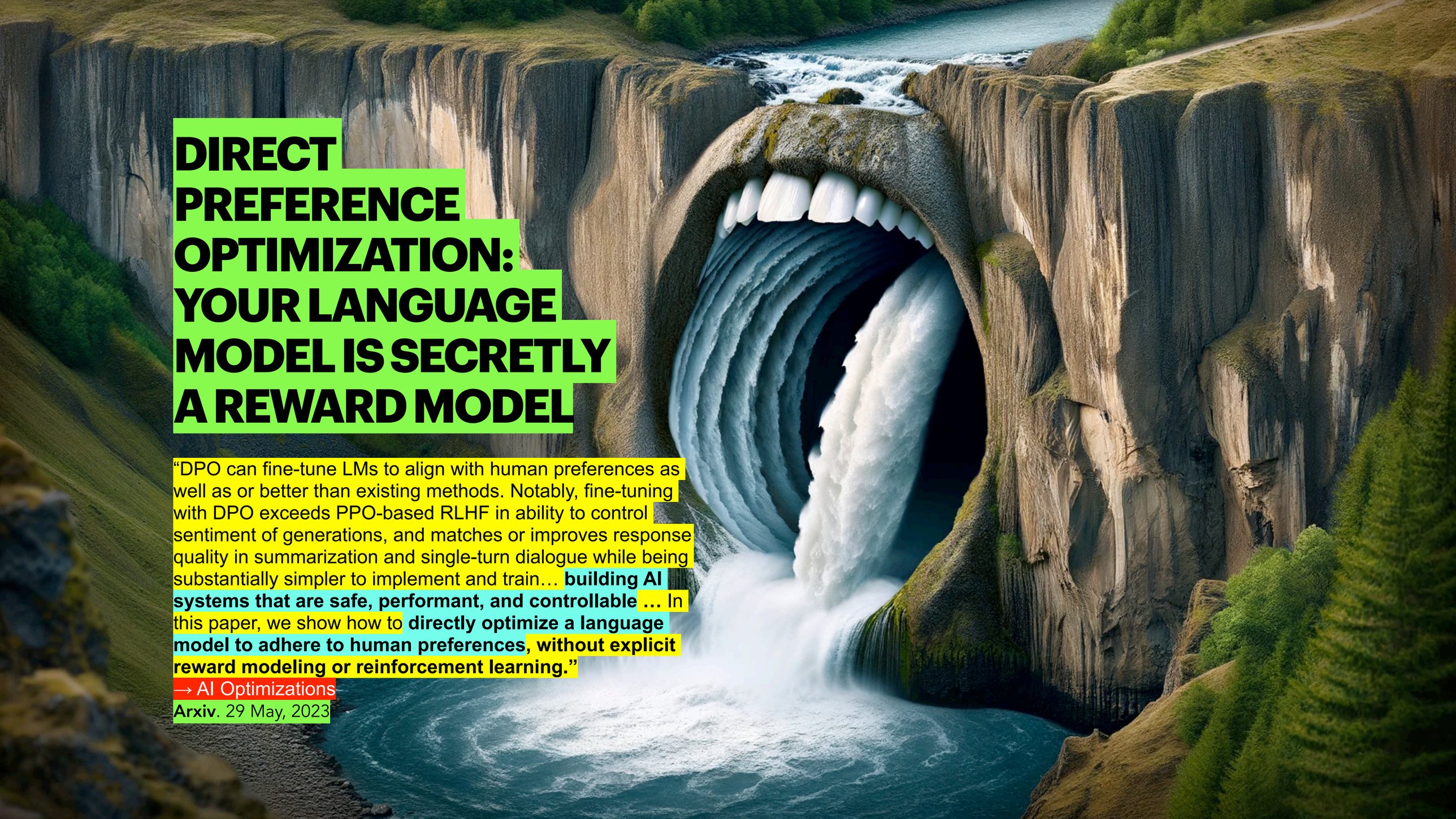


“Our process-supervised model solves 78% of problems from a representative subset of the MATH test set. Additionally, we show that active learning significantly improves the efficacy of process supervision. To support related research, we also release PRM800K, the complete dataset of 800,000 step-level human feedback labels used to train our best reward model.”

Reasoning step-by-step, generating (**hallucinating**) a huge sample, selecting best.

→ AI Optimizations

**OpenAI**. 31 May, 2023



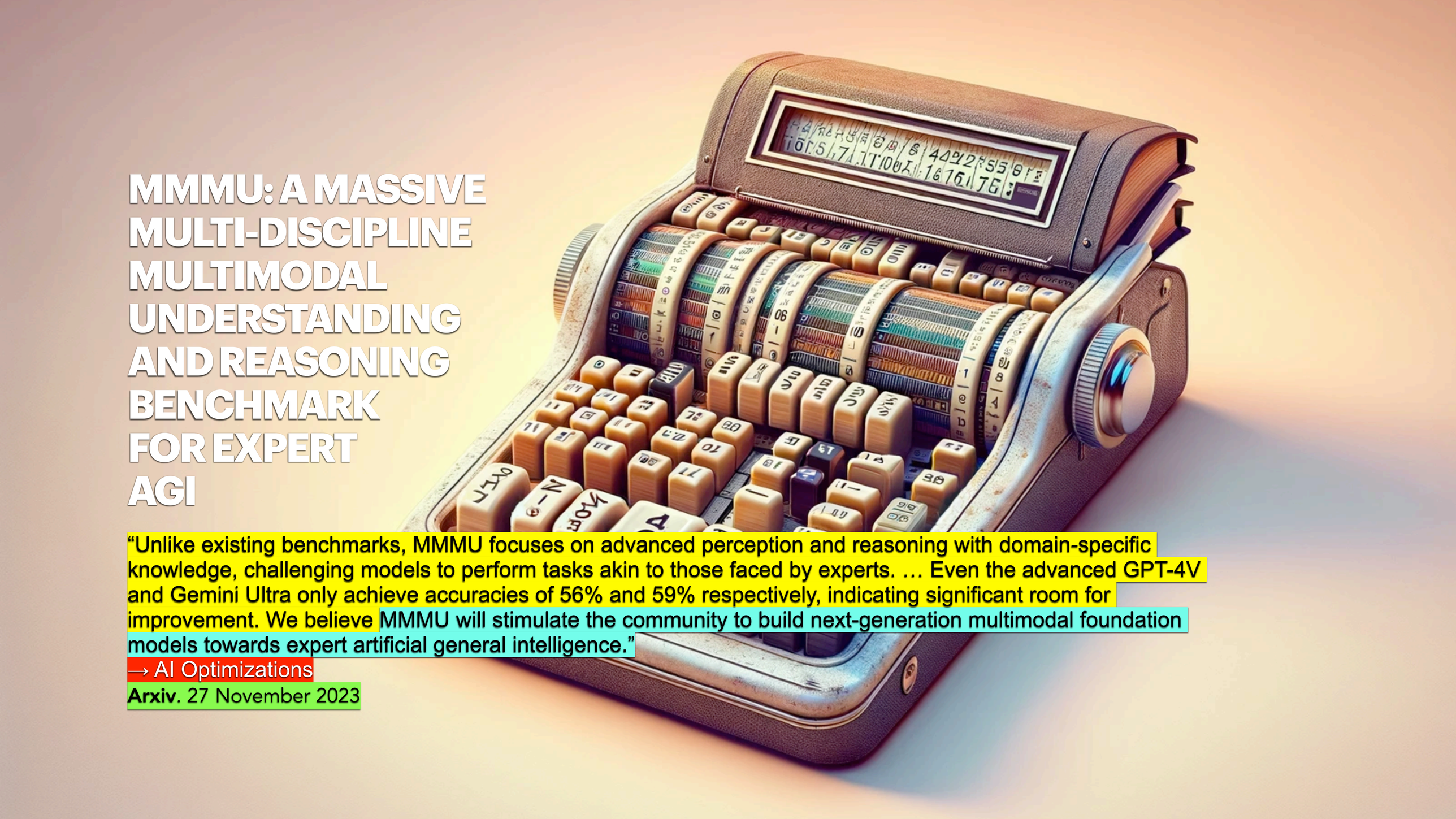
# DIRECT PREFERENCE OPTIMIZATION: YOUR LANGUAGE MODEL IS SECRETLY A REWARD MODEL

"DPO can fine-tune LMs to align with human preferences as well as or better than existing methods. Notably, fine-tuning with DPO exceeds PPO-based RLHF in ability to control sentiment of generations, and matches or improves response quality in summarization and single-turn dialogue while being substantially simpler to implement and train… **building AI systems that are safe, performant, and controllable …** In this paper, we show how to **directly optimize a language model to adhere to human preferences, without explicit reward modeling or reinforcement learning."**

→ AI Optimizations

**Arxiv**. 29 May, 2023

# MMMU: A MASSIVE MULTI-DISCIPLINE MULTIMODAL UNDERSTANDING AND REASONING BENCHMARK FOR EXPERT AGI



"Unlike existing benchmarks, MMMU focuses on advanced perception and reasoning with domain-specific knowledge, challenging models to perform tasks akin to those faced by experts. … Even the advanced GPT-4V and Gemini Ultra only achieve accuracies of 56% and 59% respectively, indicating significant room for improvement. We believe MMMU will stimulate the community to build next-generation multimodal foundation models towards expert artificial general intelligence."

→ AI Optimizations

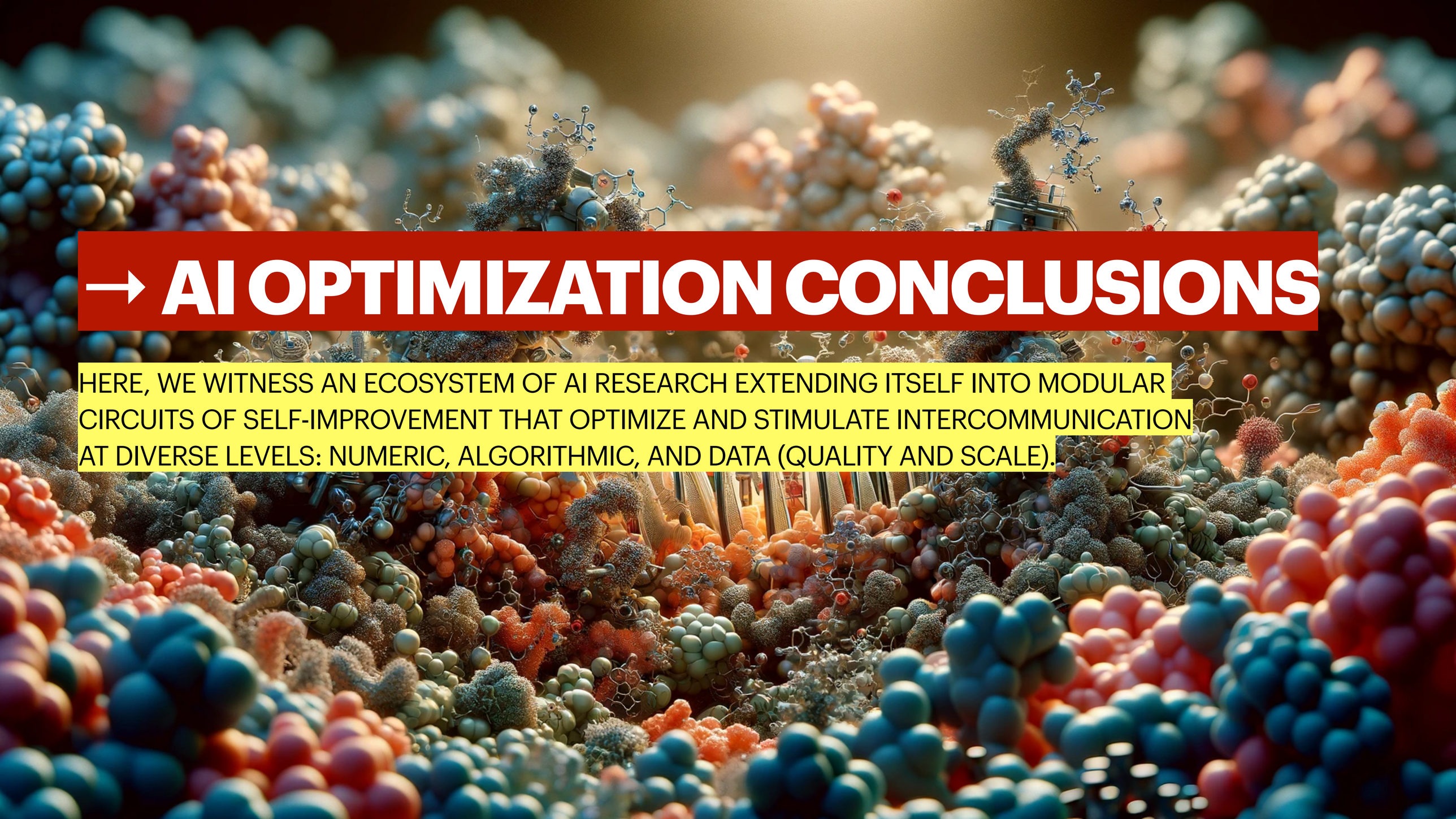
**Arxiv**. 27 November 2023

# PHI-2: THE SURPRISING POWER OF SMALL LANGUAGE MODELS

"Phi-2 a 2.7 billion-parameter language model that demonstrates outstanding reasoning and language understanding capabilities, showcasing state-of-the-art performance among base language models with less than 13 billion parameters. On complex benchmarks Phi-2 matches or outperforms models up to 25x larger, thanks to new innovations in model scaling and training data curation." "Textbooks Are All You Need."

**Hallucinating from a solid foundation grounded in fact** leads to a proliferation of plausible potentialities.

**Microsoft**. December 12, 2023

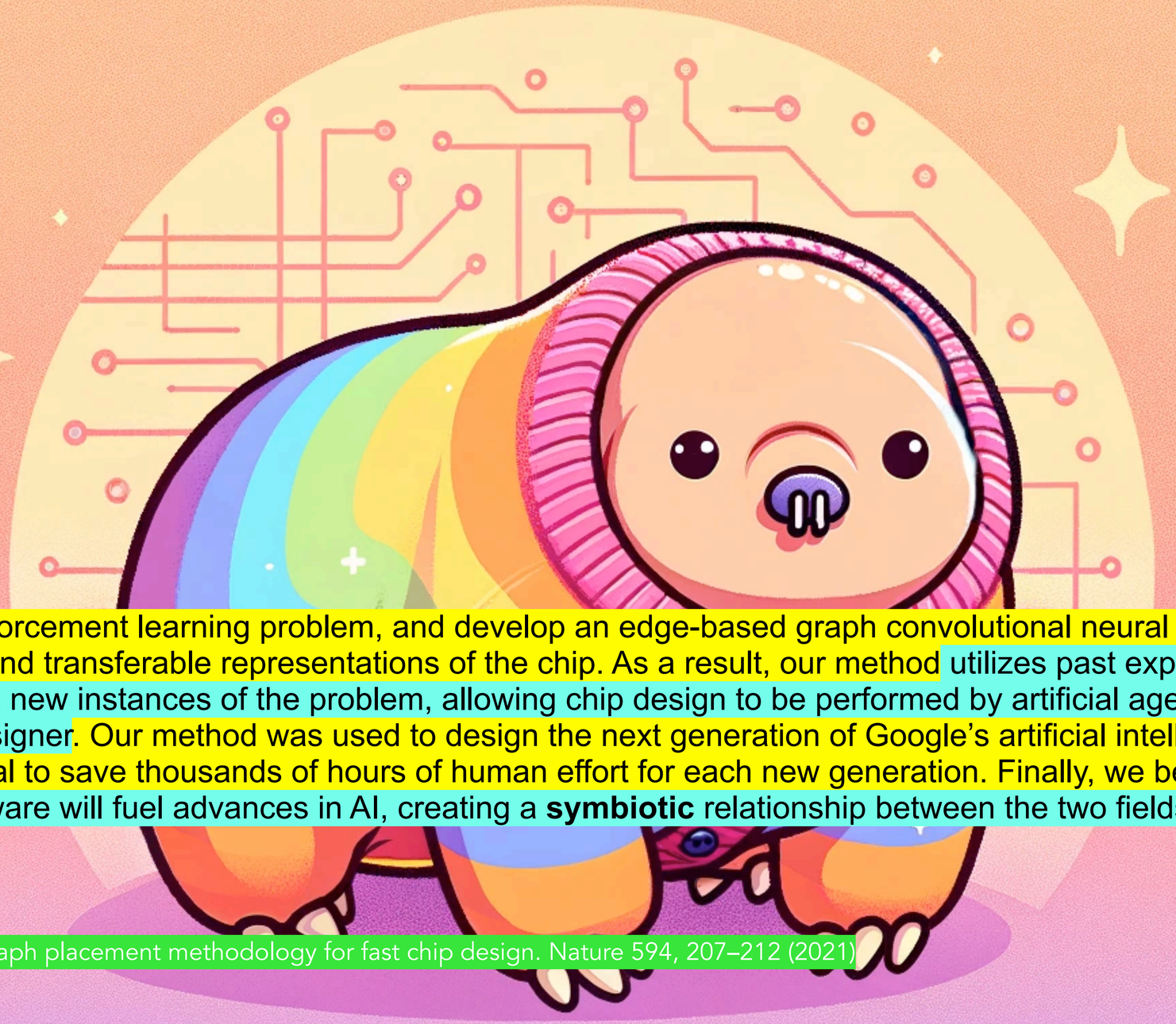


# → AI OPTIMIZATION CONCLUSIONS

HERE, WE WITNESS AN ECOSYSTEM OF AI RESEARCH EXTENDING ITSELF INTO MODULAR CIRCUITS OF SELF-IMPROVEMENT THAT OPTIMIZE AND STIMULATE INTERCOMMUNICATION AT DIVERSE LEVELS: NUMERIC, ALGORITHMIC, AND DATA (QUALITY AND SCALE).


→ AI Designing AI
Hallucinations of circuits, algorithms, code.

# A GRAPH PLACEMENT METHODOLOGY FOR FAST CHIP DESIGN

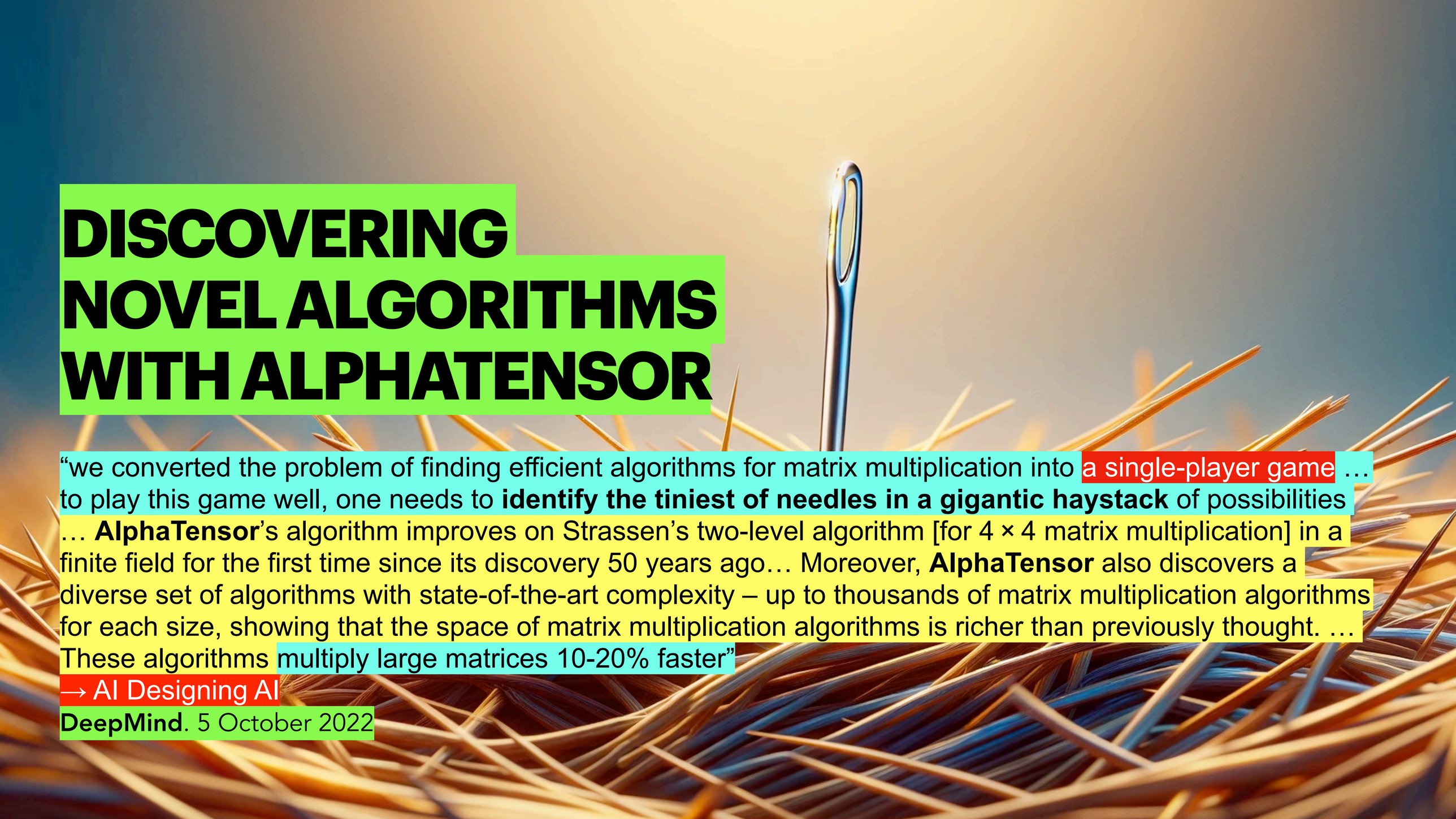


“we pose chip floorplanning as a reinforcement learning problem, and develop an edge-based graph convolutional neural network architecture capable of learning rich and transferable representations of the chip. As a result, our method utilizes past experience to become better and faster at solving new instances of the problem, allowing chip design to be performed by artificial agents with more experience than any human designer. Our method was used to design the next generation of Google’s artificial intelligence (AI) accelerators, and has the potential to save thousands of hours of human effort for each new generation. Finally, we believe that more powerful AI-designed hardware will fuel advances in AI, creating a **symbiotic** relationship between the two fields”

→ AI Designing AI

**Google**. 09 June 2021

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



# DISCOVERING NOVEL ALGORITHMS WITH ALPHATENSOR

“we converted the problem of finding efficient algorithms for matrix multiplication into a single-player game … to play this game well, one needs to **identify the tiniest of needles in a gigantic haystack** of possibilities … **AlphaTensor**’s algorithm improves on Strassen’s two-level algorithm [for 4 × 4 matrix multiplication] in a finite field for the first time since its discovery 50 years ago… Moreover, **AlphaTensor** also discovers a diverse set of algorithms with state-of-the-art complexity – up to thousands of matrix multiplication algorithms for each size, showing that the space of matrix multiplication algorithms is richer than previously thought. … These algorithms multiply large matrices 10-20% faster”

→ AI Designing AI

**DeepMind**. 5 October 2022

# ALPHA DEV

“Fundamental algorithms such as sorting or hashing are used trillions of times on any given day… an artificial intelligence (AI) system that uses reinforcement learning to discover enhanced computer science algorithms – surpassing those honed by scientists and engineers over decades. … we transformed sorting into a single player ‘assembly game’. … AlphaDev uncovered new sorting algorithms that led to improvements in the LLVM libc++ sorting library that were up to 70% faster for shorter sequences … AlphaDev’s new hashing algorithm was released into the open-source Abseil library,”

→ AI Designing AI

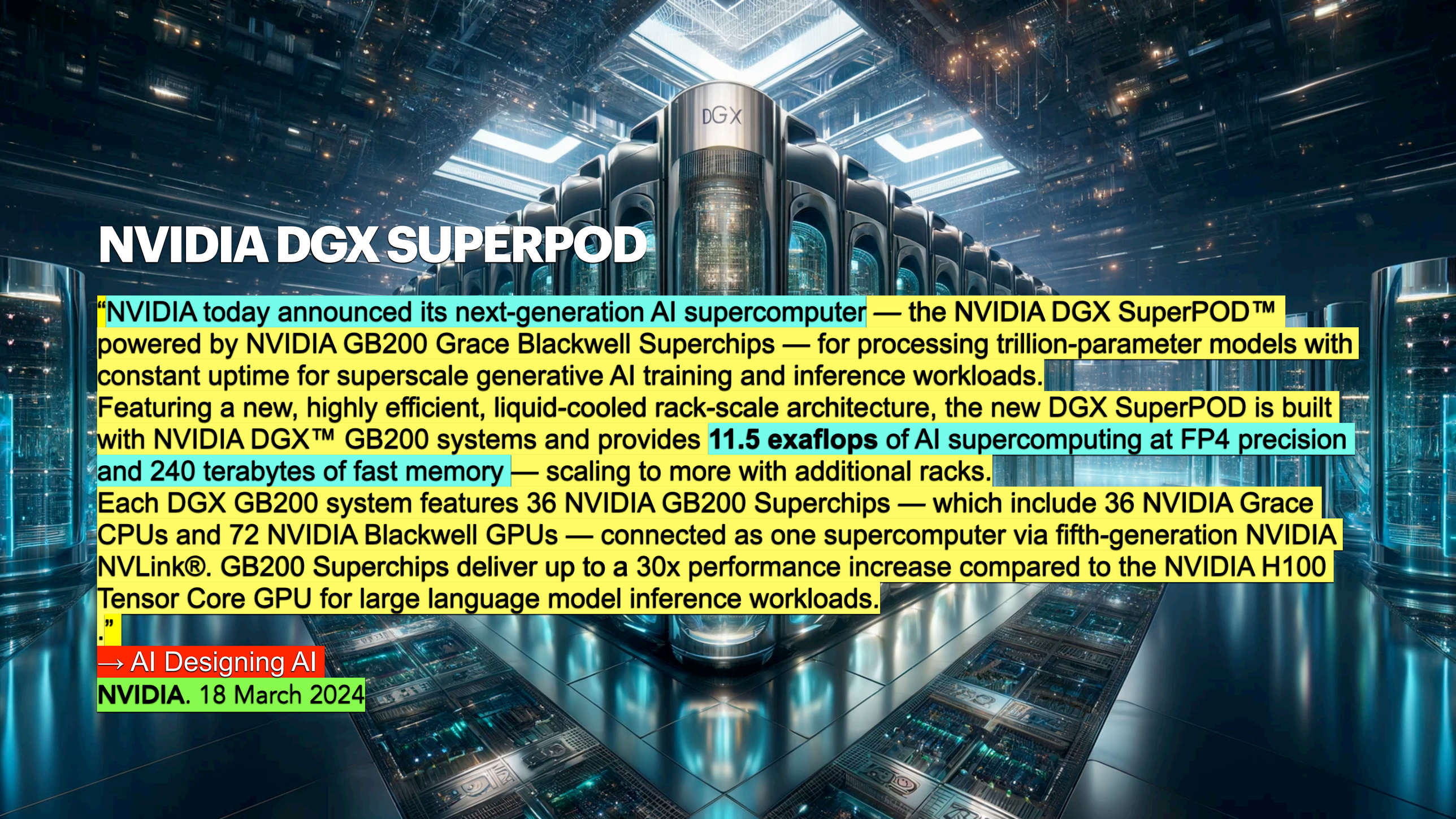
**Daniel J. Mankowitz and Andrea Michi. DeepMind**. 07June 2023

# PROJECT CEIBA

"NVIDIA and AWS are partnering on Project Ceiba to design the world's fastest GPU-powered AI supercomputer—an at-scale system with GH200 NVL32 and Amazon EFA interconnect hosted by AWS for NVIDIA's own research and development team. This first-of-its-kind supercomputer—featuring 16,384 NVIDIA GH200 Superchips and capable of processing **65 exaflops** of AI—will be used by NVIDIA to propel its next wave of generative AI innovation."

→ AI Designing AI

**NVIDIA and AWS**. 28 November 2023



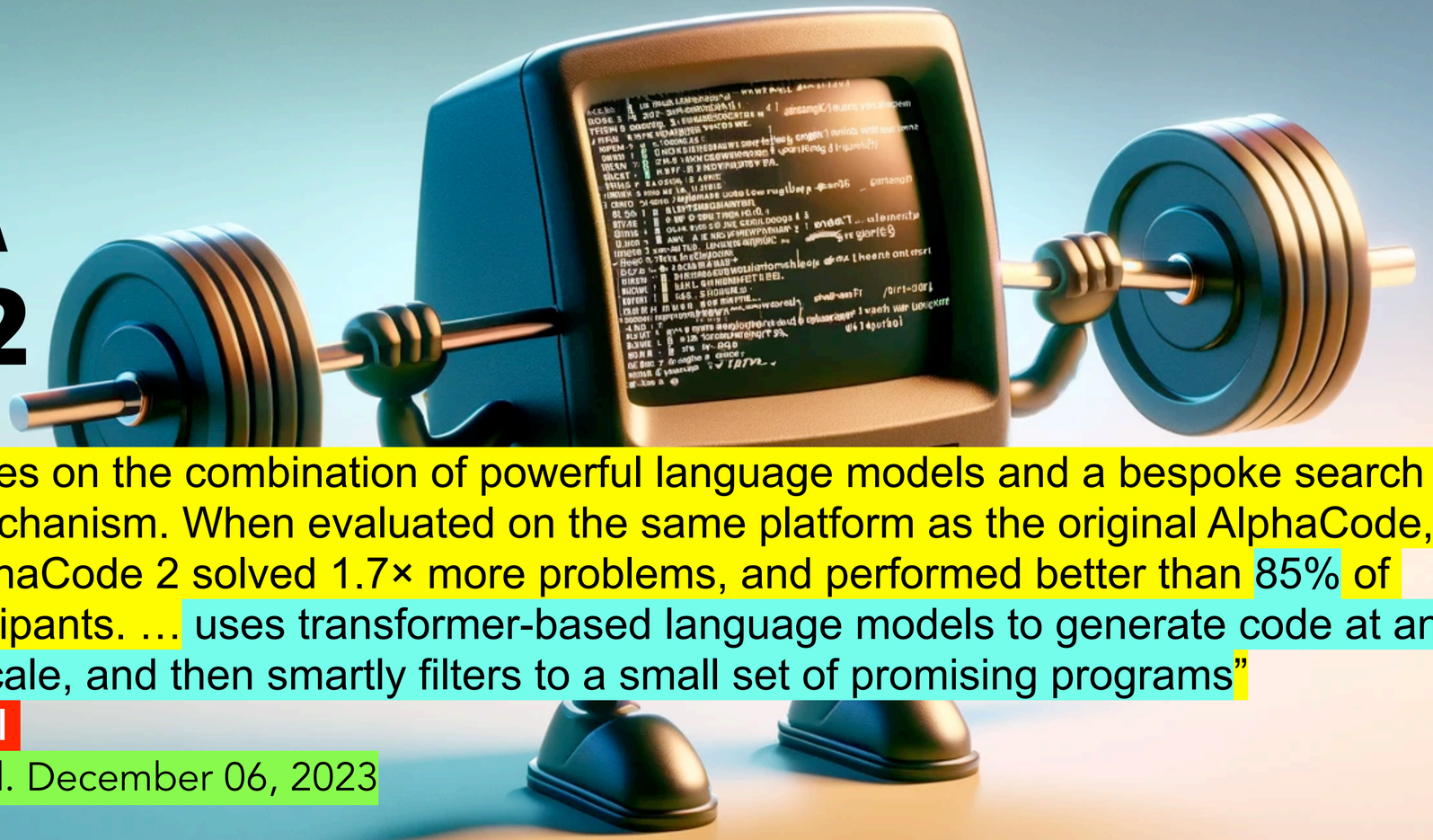

# NVIDIA DGX SUPERPOD

“NVIDIA today announced its next-generation AI supercomputer — the NVIDIA DGX SuperPOD™ powered by NVIDIA GB200 Grace Blackwell Superchips — for processing trillion-parameter models with constant uptime for superscale generative AI training and inference workloads.
Featuring a new, highly efficient, liquid-cooled rack-scale architecture, the new DGX SuperPOD is built with NVIDIA DGX™ GB200 systems and provides **11.5 exaflops** of AI supercomputing at FP4 precision and 240 terabytes of fast memory — scaling to more with additional racks.
Each DGX GB200 system features 36 NVIDIA GB200 Superchips — which include 36 NVIDIA Grace CPUs and 72 NVIDIA Blackwell GPUs — connected as one supercomputer via fifth-generation NVIDIA NVLink®. GB200 Superchips deliver up to a 30x performance increase compared to the NVIDIA H100 Tensor Core GPU for large language model inference workloads.
.”

→ AI Designing AI

**NVIDIA**. 18 March 2024

# ALPHA CODE 2



“AlphaCode 2 relies on the combination of powerful language models and a bespoke search and reranking mechanism. When evaluated on the same platform as the original AlphaCode, we found that AlphaCode 2 solved 1.7× more problems, and performed better than 85% of competition participants. … uses transformer-based language models to generate code at an unprecedented scale, and then smartly filters to a small set of promising programs”

→ AI Designing AI

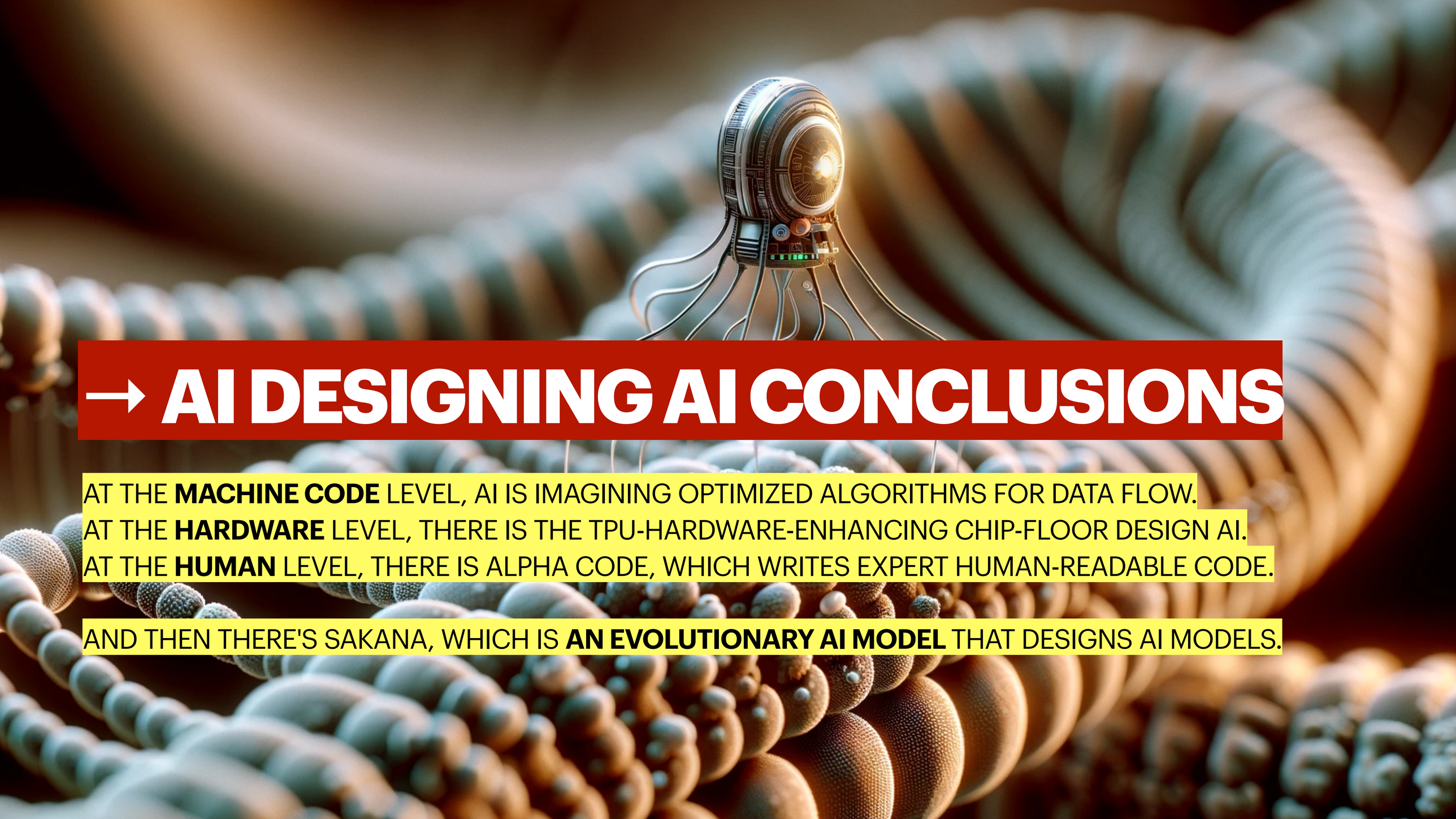
**Google DeepMind**. December 06, 2023

# EVOLVING NEW FOUNDATION MODELS: UNLEASHING THE POWER OF AUTOMATING MODEL DEVELOPMENT

"**Evolutionary Model Merge**, a general method that uses evolutionary techniques to efficiently discover the best ways to combine different models from the vast ocean of different open-source models with diverse capabilities. As of writing, Hugging Face has over 500k models in dozens of different modalities that, in principle, could be combined to form new models with new capabilities! By working with the vast collective intelligence of existing open models, our method is able to *automatically* create *new* foundation models with desired capabilities specified by the user."

→ AI Designing AI

**Sakana.ai** 21 March 2024



# → AI DESIGNING AI CONCLUSIONS

AT THE **MACHINE CODE** LEVEL, AI IS IMAGINING OPTIMIZED ALGORITHMS FOR DATA FLOW.
AT THE **HARDWARE** LEVEL, THERE IS THE TPU-HARDWARE-ENHANCING CHIP-FLOOR DESIGN AI.
AT THE **HUMAN** LEVEL, THERE IS ALPHA CODE, WHICH WRITES EXPERT HUMAN-READABLE CODE.

AND THEN THERE'S SAKANA, WHICH IS **AN EVOLUTIONARY AI MODEL** THAT DESIGNS AI MODELS.

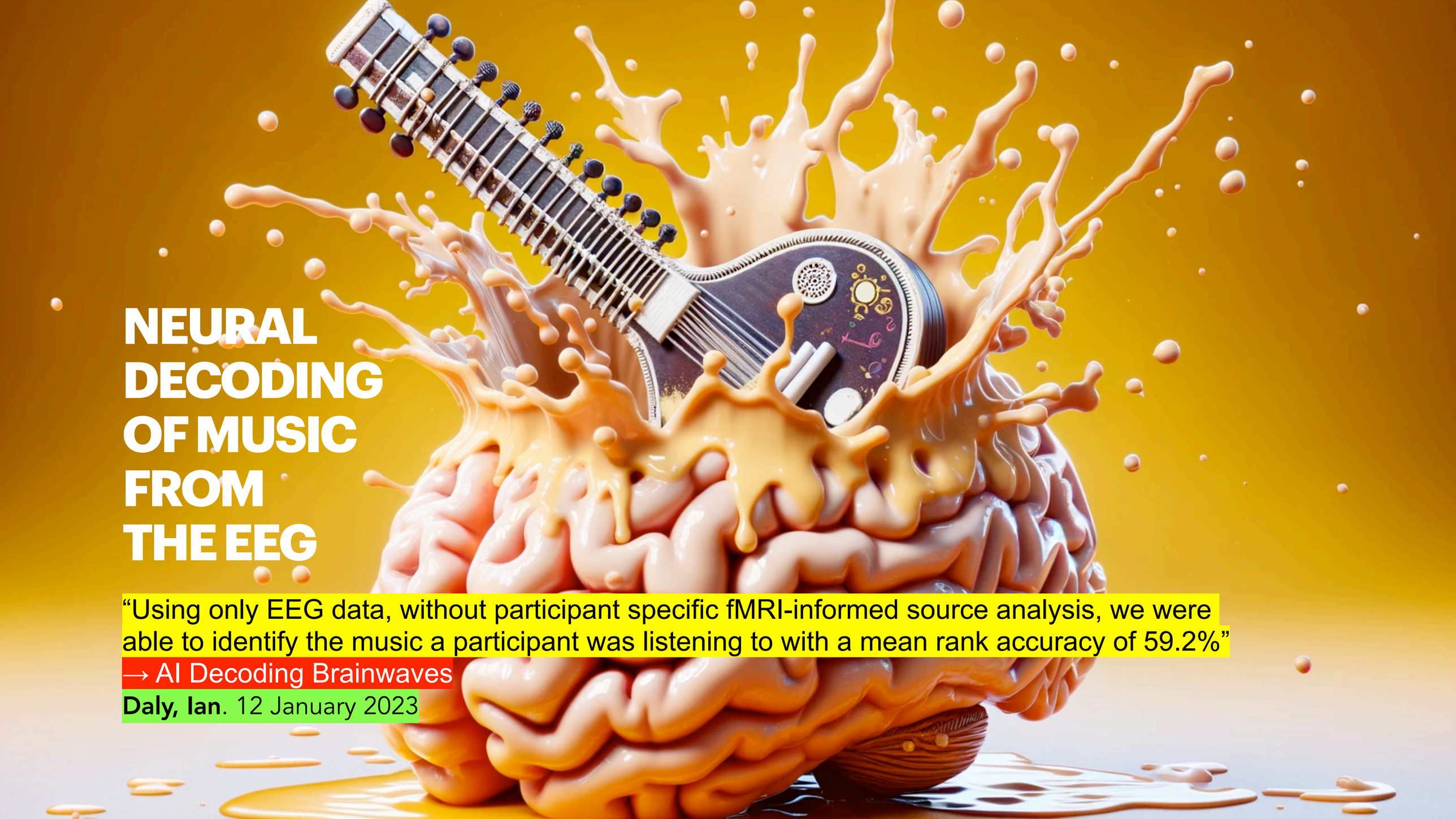

→ AI Mind-Reading
Hallucinated thoughts, intentions, words, images, gestures.

# WALKING NATURALLY AFTER SPINAL CORD INJURY USING A BRAIN–SPINE INTERFACE

“digital bridge between the brain and spinal cord that enabled an individual with chronic tetraplegia to stand and walk naturally in community settings. This brain–spine interface (BSI) consists of fully implanted recording and stimulation systems that establish a direct link between cortical signals and the analogue modulation of epidural electrical stimulation targeting the spinal cord regions involved in the production of walking. A highly reliable BSI is calibrated within a few minutes. This reliability has remained stable over one year… This digital bridge establishes a framework to restore natural control of movement after paralysis.”

→ AI Decoding Brainwaves

**Nature**. 24 May 2023

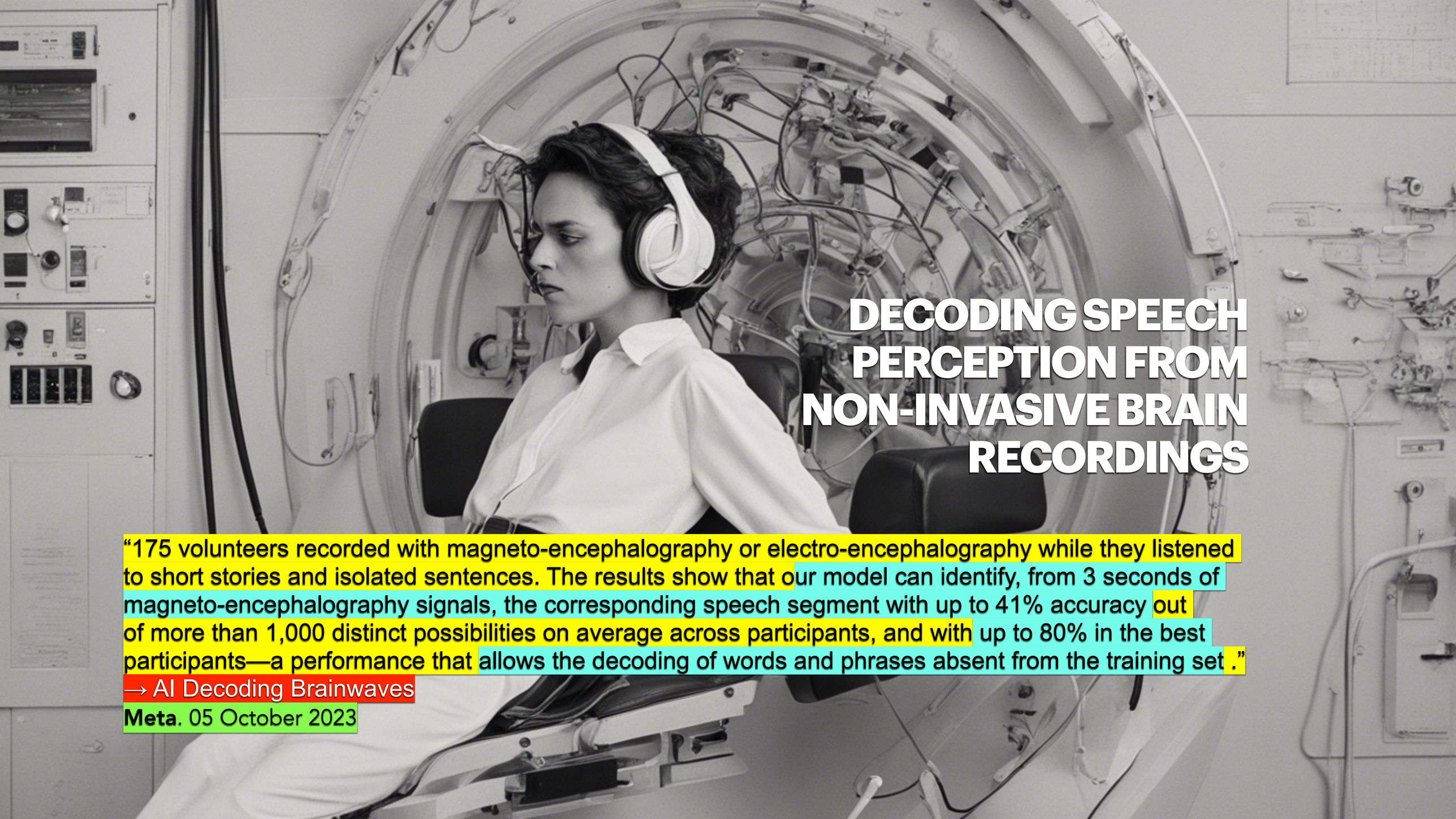


# NEURAL DECODING OF MUSIC FROM THE EEG

“Using only EEG data, without participant specific fMRI-informed source analysis, we were able to identify the music a participant was listening to with a mean rank accuracy of 59.2%”

→ AI Decoding Brainwaves

**Daly, Ian**. 12 January 2023

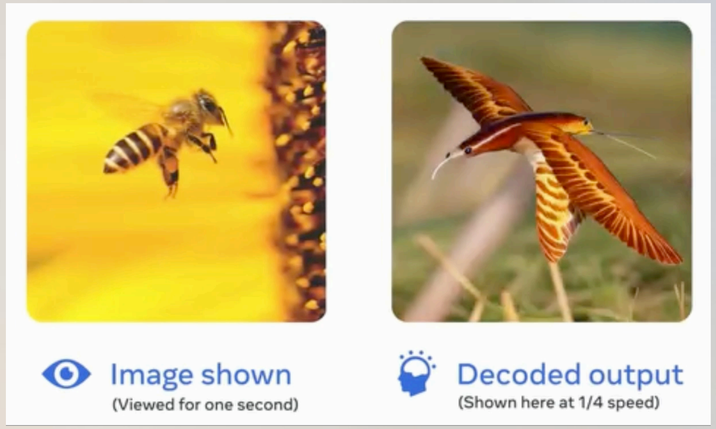


# DECODING SPEECH PERCEPTION FROM NON-INVASIVE BRAIN RECORDINGS

“175 volunteers recorded with magneto-encephalography or electro-encephalography while they listened to short stories and isolated sentences. The results show that our model can identify, from 3 seconds of magneto-encephalography signals, the corresponding speech segment with up to 41% accuracy out of more than 1,000 distinct possibilities on average across participants, and with up to 80% in the best participants—a performance that allows the decoding of words and phrases absent from the training set .”

→ AI Decoding Brainwaves

**Meta**. 05 October 2023




# TOWARD A REAL-TIME DECODING OF IMAGES FROM BRAIN ACTIVITY

“Using magnetoencephalography (MEG), a non-invasive neuroimaging technique in which thousands of brain activity measurements are taken per second, we showcase an AI system capable of decoding the unfolding of visual representations in the brain with an unprecedented temporal resolution. … Decoding of visual stimuli need not be restricted to a limited set of classes, but can now leverage pretrained representations to condition subsequent generative AI models. While the resulting image may be partly “hallucinated”, interpreting images can be much simpler than interpreting latent features.”

→ AI Decoding Brainwaves

**Meta**. 18 October 2023

# MOVING MAGNETOENCEPHALOGRAPHY TOWARDS REAL-WORLD APPLICATIONS WITH A WEARABLE SYSTEM

“a magnetoencephalography system that can be worn like a helmet, allowing free and natural movement during scanning. This is possible owing to the integration of quantum sensors which do not rely on superconducting technology, with a system for nulling background magnetic fields. We demonstrate human electrophysiological measurement at millisecond resolution while subjects make natural movements, including head nodding, stretching, drinking and playing a ball game.”

→ AI Decoding Brainwaves

**Nature**. 21 March 2018

# CINEMATIC MINDSCAPES: HIGH-QUALITY VIDEO RECONSTRUCTION FROM BRAIN ACTIVITY

“We propose Mind-Video, which progressively learns spatiotemporal information from continuous fMRI data through masked brain modeling + multimodal contrastive learning + spatiotemporal attention + co-training with an augmented Stable Diffusion model that incorporates network temporal inflation.”

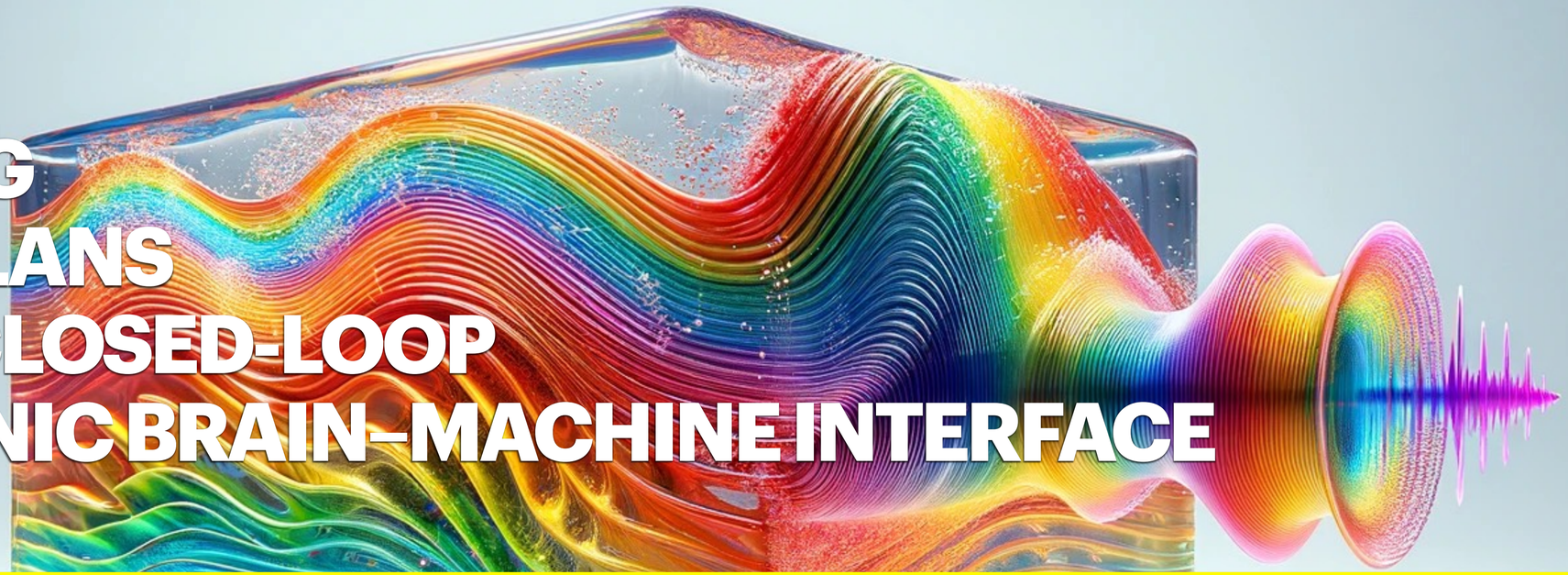
→ AI Decoding Brainwaves

**NeurIPS**. 19 May 2023

# BUTTERFLY NETWORK ANNOUNCES 5-YEAR CO-DEVELOPMENT AGREEMENT WITH FOREST NEUROTECH FOR NEXT-GENERATION BRAIN COMPUTER INTERFACES USING ULTRASOUND-ON-CHIP™ TECHNOLOGY.

"Companies announce collaboration to develop the first implanted, whole-brain neural interface Powered by Butterfly's Ultrasound-on-Chip technology™"
→ AI Decoding Brainwaves
**Nature.** 23 October 2023

# DECODING MOTOR PLANS USING A CLOSED-LOOP ULTRASONIC BRAIN–MACHINE INTERFACE



"Brain–machine interfaces (BMIs) enable people living with chronic paralysis to control computers, robots and more with nothing but thought. Existing BMIs have trade-offs across invasiveness, performance, spatial coverage and spatiotemporal resolution. Functional ultrasound (fUS) neuroimaging is an emerging technology … a new class of **less-invasive** (epidural) interfaces that generalize across extended time periods and promise to restore function to people with neurological impairments … We streamed fUS data from the posterior parietal cortex of two rhesus macaque monkeys while they performed eye and hand movements. After training, the monkeys controlled up to eight movement directions using the BMI "

→ AI Decoding Brainwaves

**Nature.** 30 November 2023 (received Jan 2023)

# NEURALINK

"The N1 Implant records neural activity through 1024 electrodes distributed across 64 threads."

"By modeling the relationship between different patterns of neural activity and intended movement directions, we can build a model (i.e., "calibrate a decoder") that can predict the direction and speed of an upcoming or intended movement."

→ AI Decoding Brainwaves

Musk, Elon, and Neuralink. "An Integrated Brain-Machine Interface Platform with Thousands of Channels." bioRxiv, August 2, 2019. + "Pager Plays MindPong" Neuralink Blog. 2021

The first human received an implant from @Neuralink on Jan 28, 2024
**March 20, 2024 Livestream of @Neuralink demonstrating "Telepathy"** – controlling a computer and playing video games just by thinking

1. Quadriplegia (paralysis or severe weakness in all four limbs)
2. Paraplegia (paralysis in at least two limbs)
3. Vision loss
4. Hearing loss
5. Aphasia (inability to speak)
6. Amputee (major limb amputation)

# → AI MIND-READING CONCLUSIONS

FROM MESSY, MOIST, NOISY, NON-LINEAR, TURBULENT SYNAPTIC SIGNALS, AI-ENHANCED SENSORS GATHER, FILTER AND THEN PASS THE RELEVANT ASPECTS OF THE SIGNAL TO ANOTHER AI (WHICH EITHER ACTUATES A MOTOR NEURON, A DIFFUSION IMAGE GENERATOR, OR A LANGUAGE MODEL).

AFTER THE FILTRATION OF THE SIGNAL ARISING FROM THE BRAIN'S SIMULATION, CLEANED DATA IS PASSED TO A GENERATIVE MODEL, AND AN EXTERNAL REPRESENTATION OF WHAT WAS PREVIOUSLY INTERNAL ARISES. THIS IS AN ARTFUL ACT, THE MAKING OF A REPRESENTATION BASED UPON NEUROLOGICAL SIGNALS IN A DIRECT FEEDBACK PROCESS:

**GATHER, FILTER, ACTUATE, REPEAT**