

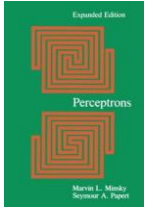
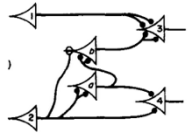
Artificial Intelligence

How is it changing business models?

Arlindo Oliveira



Key Moments in the History of Artificial Intelligence



1943

McCulloch-Pitts Neuron Model

1950

Turing – Can machines think?

1952

Rosenblatt - Perceptron

1959

Samuel - Checkers
Rumelhart, Hinton, Williams, MLP - Learning representations by back-propagating errors

1967

Weizenbaum - ELIZA

1969

Fukushima Neocognitron

IBM DeepBlue



Rumelhart & McClelland PDP Book



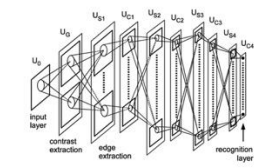
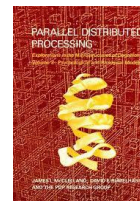
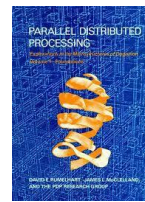
2006

1997

1987

1986

1979



2009

2012

2016

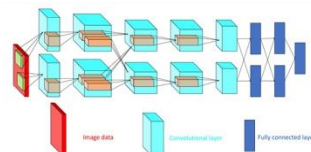
2017

2020

Li Fei-Fei et. al - ImageNet



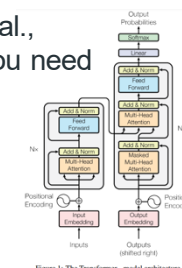
Krizhevsky, Sutskever & Hinton AlexNET



Silver et al. AlphaGo



Vaswani et al., Attention is all you need

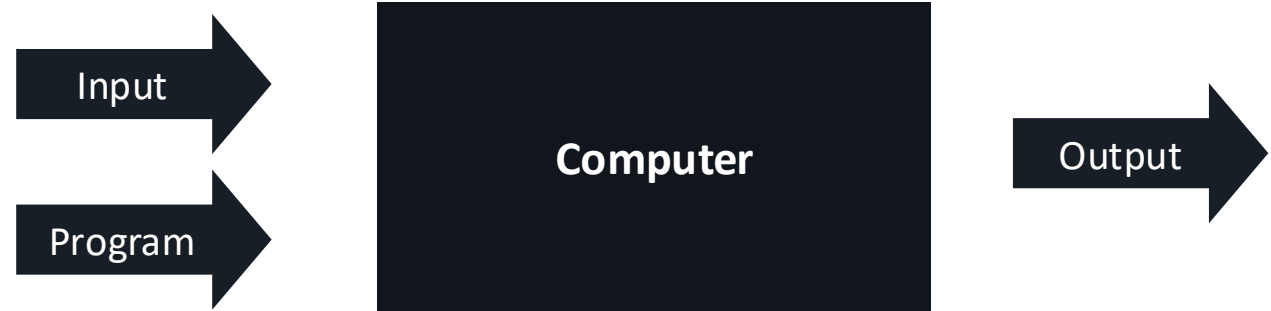


Brown et al. GPT-3

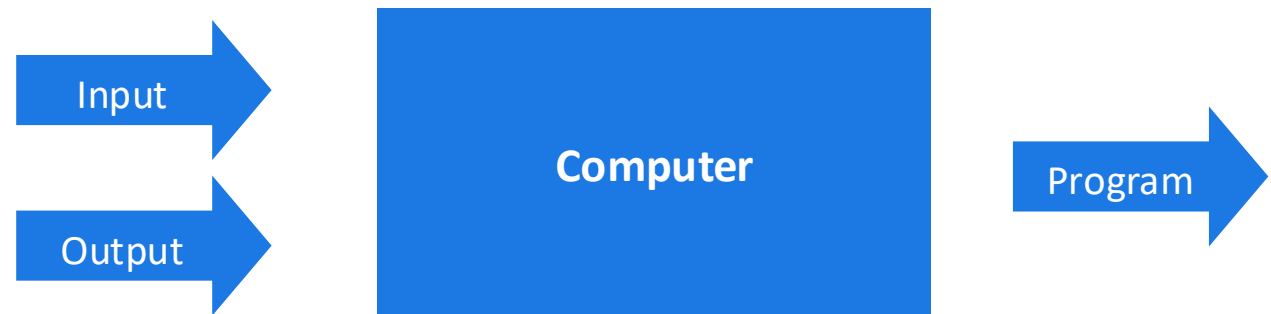


Machine Learning

Traditional programming



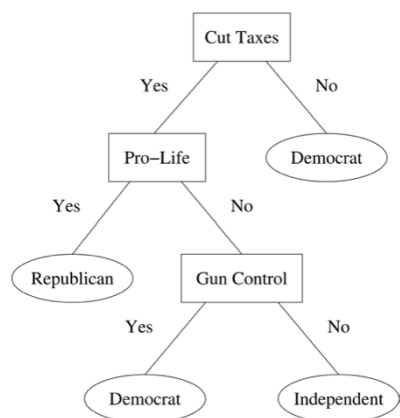
Machine Learning



Four Paradigms of Machine Learning

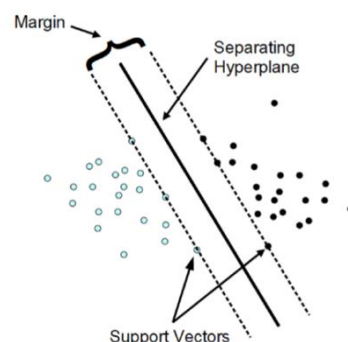
Symbolists

All intelligence can be reduced to symbol manipulation



Analogizers

The key to learning is recognizing similarities between situations



Statistical

All learned knowledge is uncertain, and learning itself is a form of uncertain inference



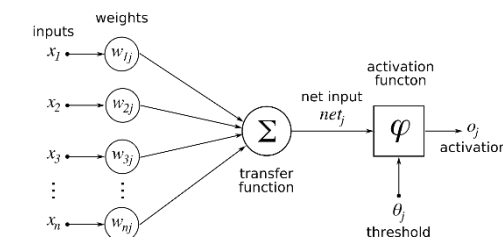
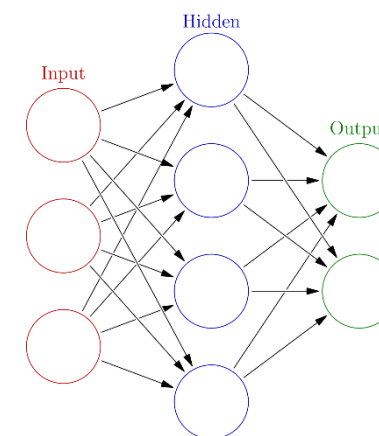
$$P(H|E) = \frac{P(H) * P(E|H)}{P(E)}$$

Diagram illustrating Bayes' Theorem with labels:

- $P(H)$: Prior Probability
- $P(E|H)$: Likelihood of the evidence 'E' if the Hypothesis 'H' is true
- $P(H|E)$: Posterior Probability of 'H' given the evidence
- $P(E)$: Prior probability that the evidence itself is true

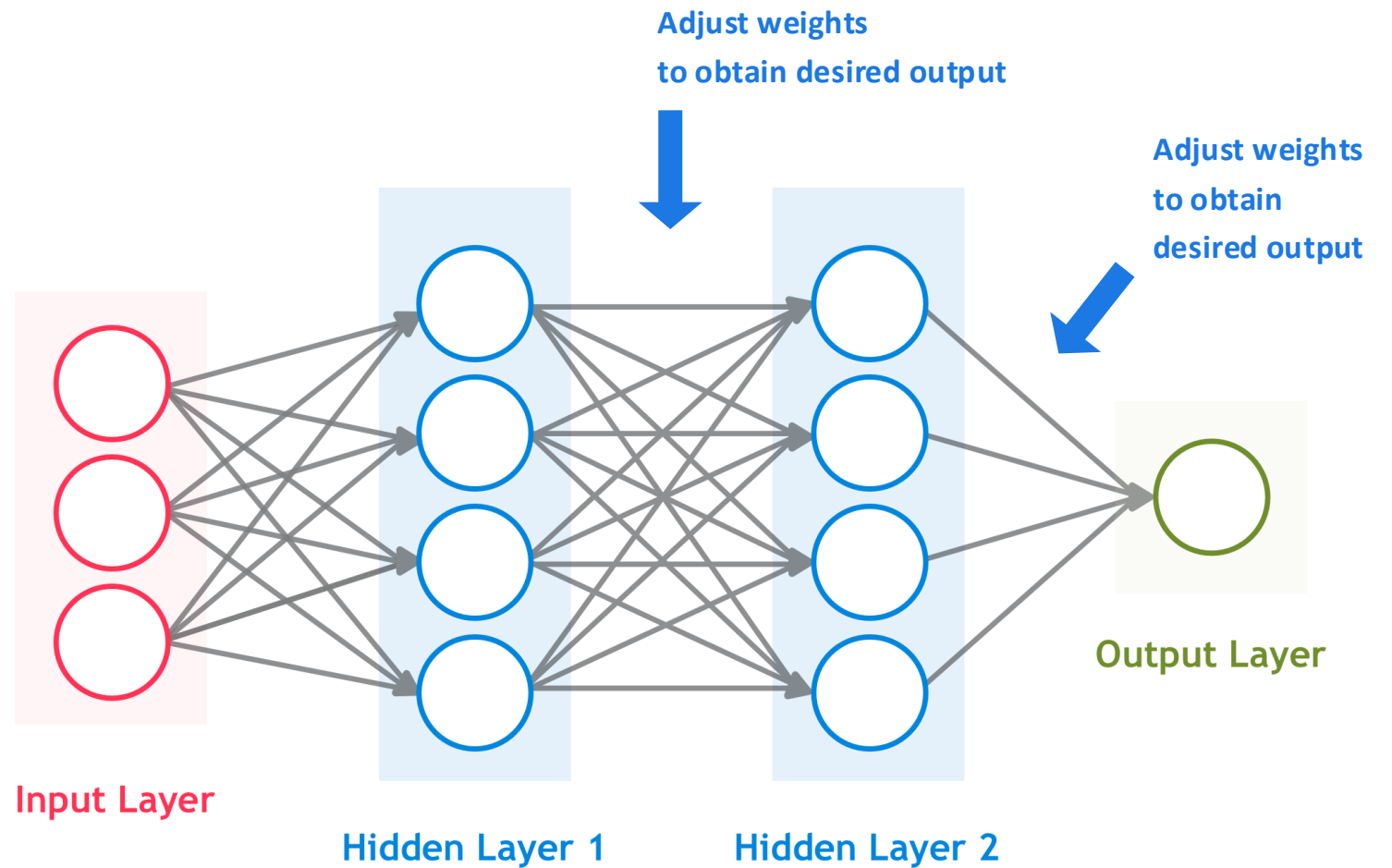
Connectionists

Learning is what the brain does, and what we need to do is reverse engineer it



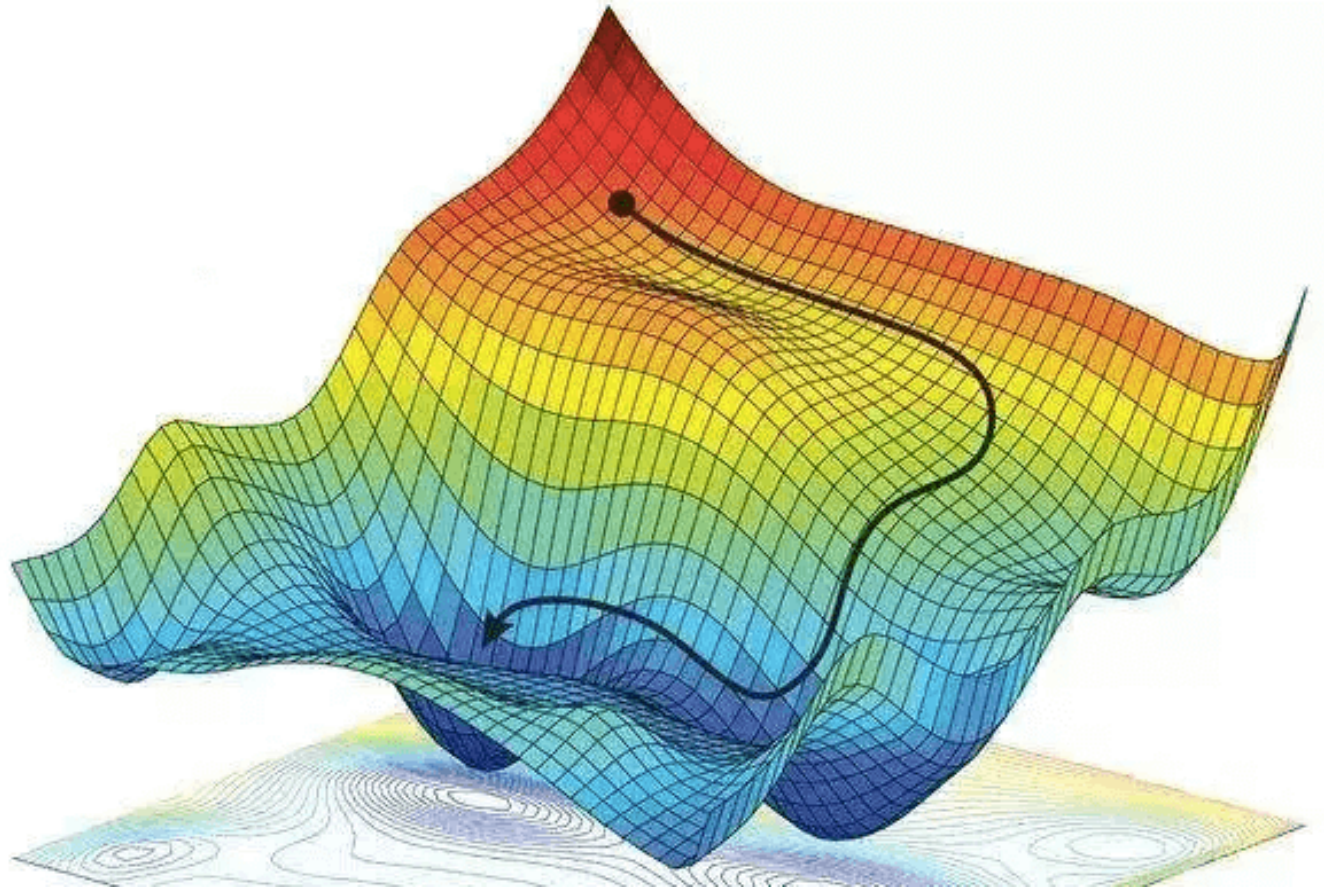
Neural Networks

Multi Layer
Perceptron



Gradient Descent

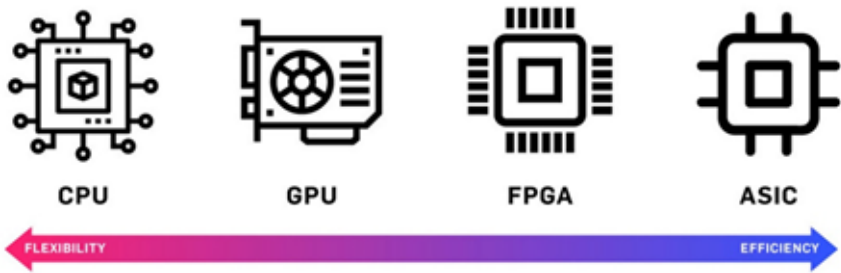
Error backpropagation



The Deep Learning Revolution

Frictionless reproducibility

Shared code
Hardware

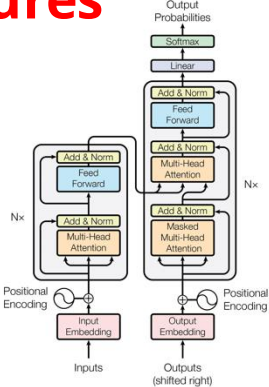
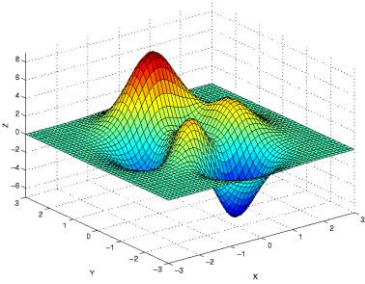


Deep Learning Revolution

Shared Data
Data



Competitive environment
Algorithms and architectures

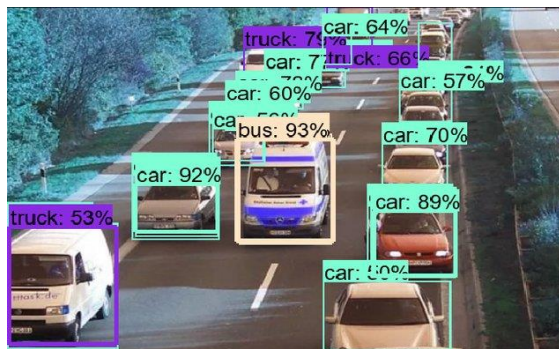


David Donoho

The Deep Learning Revolution

Four recent key developments

Deep Convolutional Neural Networks



Transformers and Large Language Models



Deep Learning Revolution



Deep Reinforcement Learning

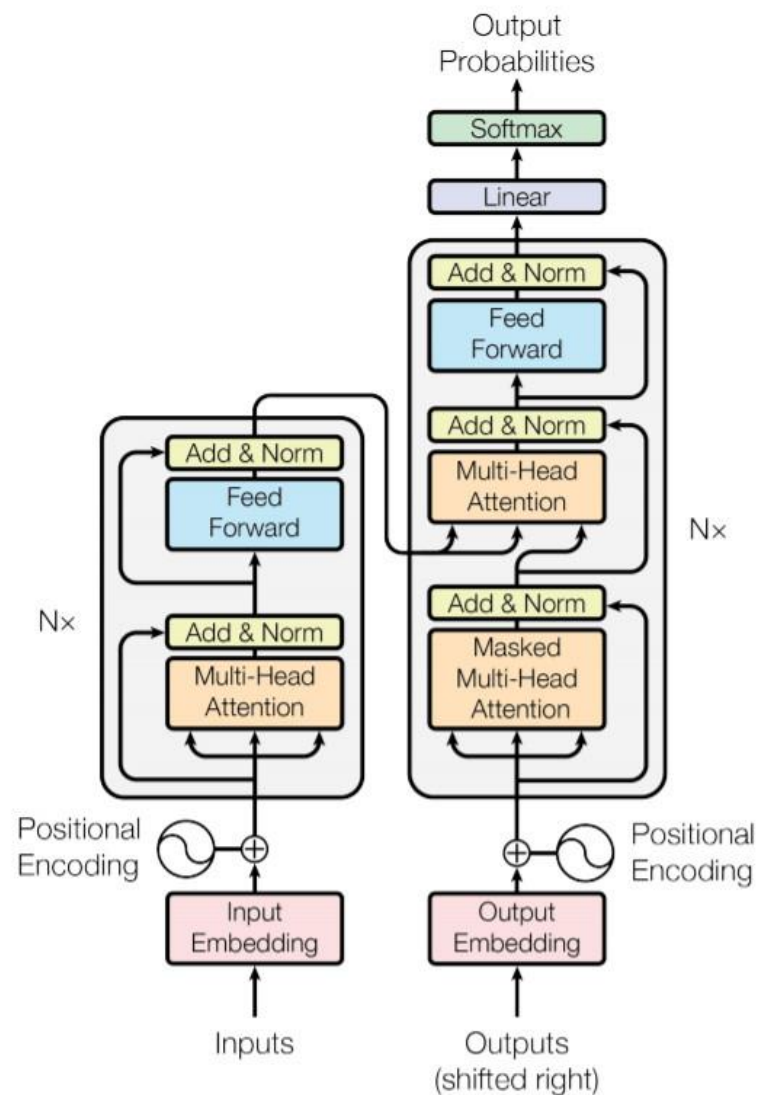


Diffusion Models



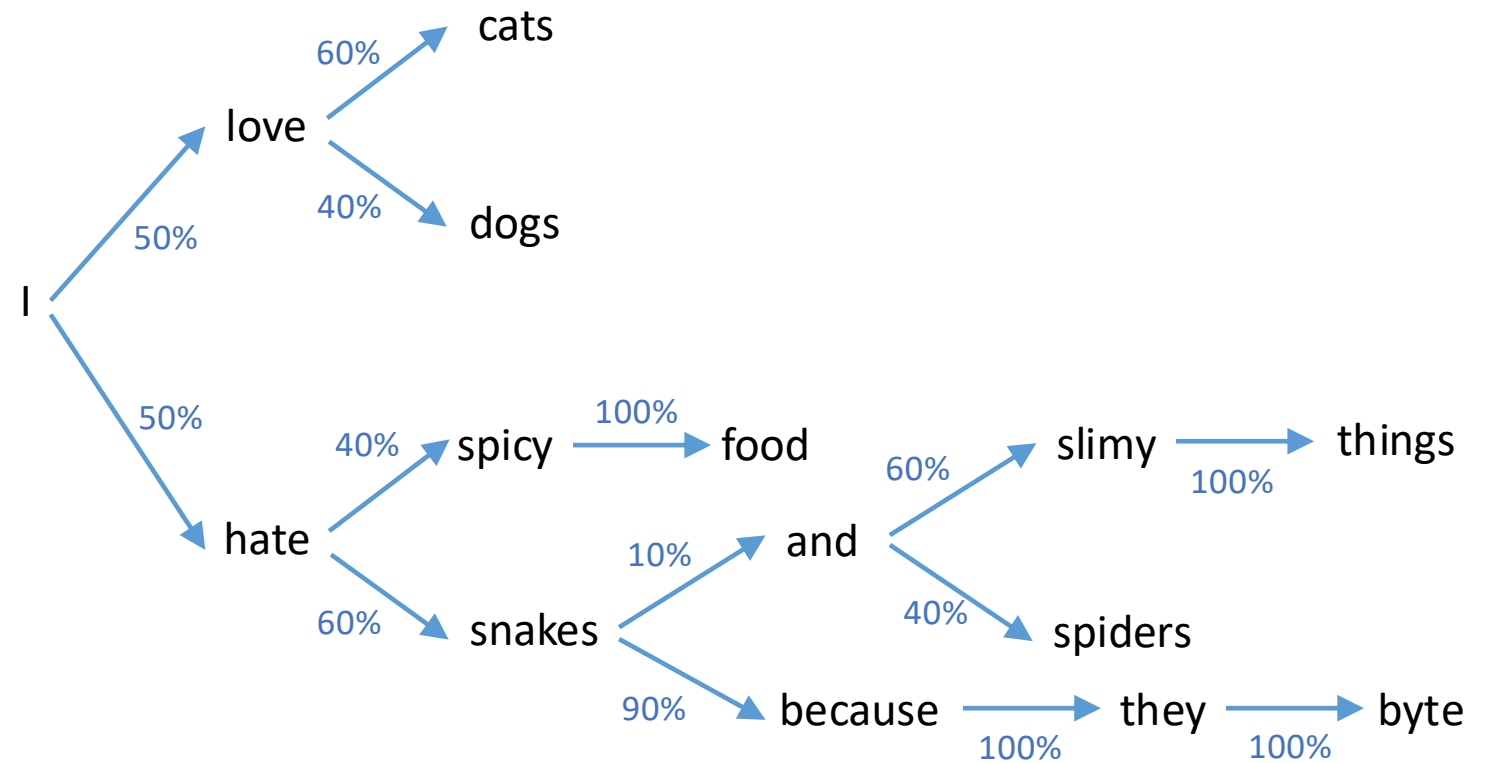
Large Language Models

Training and Inference



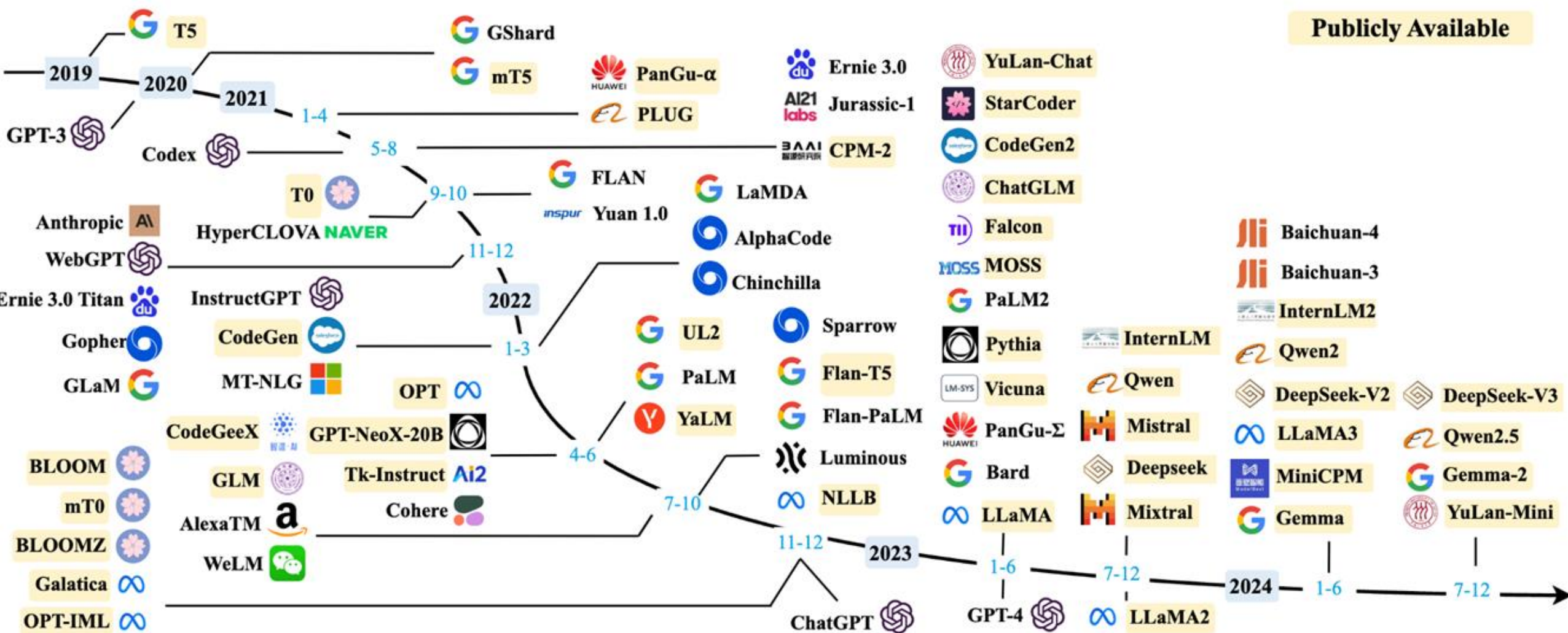
- Use the popular transformer architecture, introduced in 2017
- Words (tokens) are embedded into a “meaning” space
- Model is trained to predict the next word
- Text can be generated according to predicted word probabilities
- Creativity can be controlled by a temperature parameter
- Different choices lead to different outputs to the same prompt

Probabilistic text generation



Words are generated according to predicted probabilities, controlled by a temperature parameter that controls “creativity”

Large language models



AI Leaderboards											
LLM Leaderboard		Open LLM Leaderboard		{-} Coding		✍ Writing		⚙ Math		🔍 Research	
				📄 Long Context		🔑 Tool Calling		🧠 Reasoning		🖼 Image Gen	
										📺 Video Gen	
										More ▾	
	1	 Claude Mythos Preview UNRELEASED	Anthropic	65.2	70.3	56.1	46.5	—	—	—	🔒
	2	 GPT-5.2 Pro	OpenAI	63.3	55.3	—	28.5	—	—	—	🔒
	3	 Claude Opus 4.8	Anthropic	61.3	63.2	52.5	44.1	1,565	1.0M	53 c/s	\$7.22
	4	 GLM-5.2 NEW	Zhipu AI	59.0	59.1	49.5	40.0	1,126	1.0M	60 c/s	\$1.73
	5	 Claude Fable 5 UNRELEASED	Anthropic	58.6	66.3	59.0	50.3	1,899	—	—	🔒
	6	 GPT-5.4	OpenAI	57.3	54.3	42.6	36.3	1,743	1.0M	140 c/s	\$3.89
	7	 GPT-5.1	OpenAI	56.4	48.1	30.7	—	1,250	400K	250 c/s	\$2.22
	8	 GPT-5.5	OpenAI	56.2	58.9	49.9	40.3	2,121	1.1M	33 c/s	\$7.78
	9	 Seed 2.0 Pro	ByteDance	55.5	52.4	32.0	28.1	—	—	—	🔒
	10	 GPT-5.1 Instant	OpenAI	55.1	47.6	30.2	—	991	400K	369 c/s	\$2.22
	11	 Qwen3.7 Max	Alibaba Cloud / Qwen Team	54.6	57.6	47.2	37.6	1,680	1.0M	14 c/s	\$1.53
	12	 Claude Opus 4.7	Anthropic	53.8	57.7	48.5	40.9	1,923	1.0M	37 c/s	\$7.22
	13	 GPT-5 Medium	OpenAI	53.7	43.1	—	—	1,101	—	—	🔒
	14	 Gemini 3.5 Flash	Google	53.2	55.1	45.4	39.8	1,633	1.0M	248 c/s	\$2.33
	15	 Gemini 3.1 Pro	Google	53.2	55.7	42.4	32.2	2,111	1.0M	68 c/s	\$3.89

Some of the LLMs available

Meta GPT, Ajax GPT, GPT-5, Claude-Next, OLMo, G3PO, Arrakis (GPT-4.5?), Gobi (GPT-5?), Q* (Q-Star), Gemini, Olympus, AuroraGPT (ScienceGPT), GPT-4o1, GPT4-o3

Mistral-medium, mixtral-8x7b-32kseqn, StripedHyena 7B, Gemini, Mamba,LVM-3B, SeaLLM-13b, pplx-70b-online, SeamlessM4T-Large v2, Q-Transformer, Yuan 2.0, MEDITRON, Transformers-Arithmetic, Starling-7B

Inflection-2, Claude 2.1, Orca 2, phi-2, Mirasol3B, OtterHD-8B, Gauss, Grok-1, Grok-0, Yi-34B, jina-embeddings-v2, Fuyu, ERNIE 4.0, Zephyr, PaLI-3, NVIDIA, Lemur, AceGPT, Yasa-1, RT-X, MotionLM, GAIA-1, Qwen, Llama 2 Long, LeoLM, Mistral 7B, Kosmos-2.5, Baichuan 2, BOLT2.5B, DeciLM, MoLM, NeXT-GPT, phi-1.5, UniLM, Persimmon-8B, FLM-101B, Falcon 180B, TinyLlama, Jais, Code Llama, IDEFICS, Raven, DukunLM, WizardLM, Platypus, Japanese StableLM Alpha 7B, StableCode, Med-Flamingo, Alfred-40B-0723, LLaMA-2-7B-32K, Med-PaLM M, BTLM-3B-8K, Stable Beluga 2, Stable Beluga 1, Meta-Transformer, Llama 2, WormGPT, Claude 2, LongLLaMA, xTrimPGML, XGen, Yasa, Kosmos-2, AudioPaLM, Inflection-1, phi-1, InternLM, BlenderBot 3x, Orca, PassGPT, DIDACT, GPT-4 MathMix, PandaGPT, Falcon, 202305-refact2b-mqa-lion, Guanaco, LIMA, Formosa (FFM), CodeT5+, PaLM 2, StarCoder, MPT, Pi, GPT-2B-001, Titan, WizardLM, MPT, StableLM, Dolly 2.0, Pythia, Koala-13B, BloombergGPT, OpenFlamingo-9B, GPT4All-LoRa, Cerebras-GPT, PanGu-Sigma, CoLT5, Med-PaLM

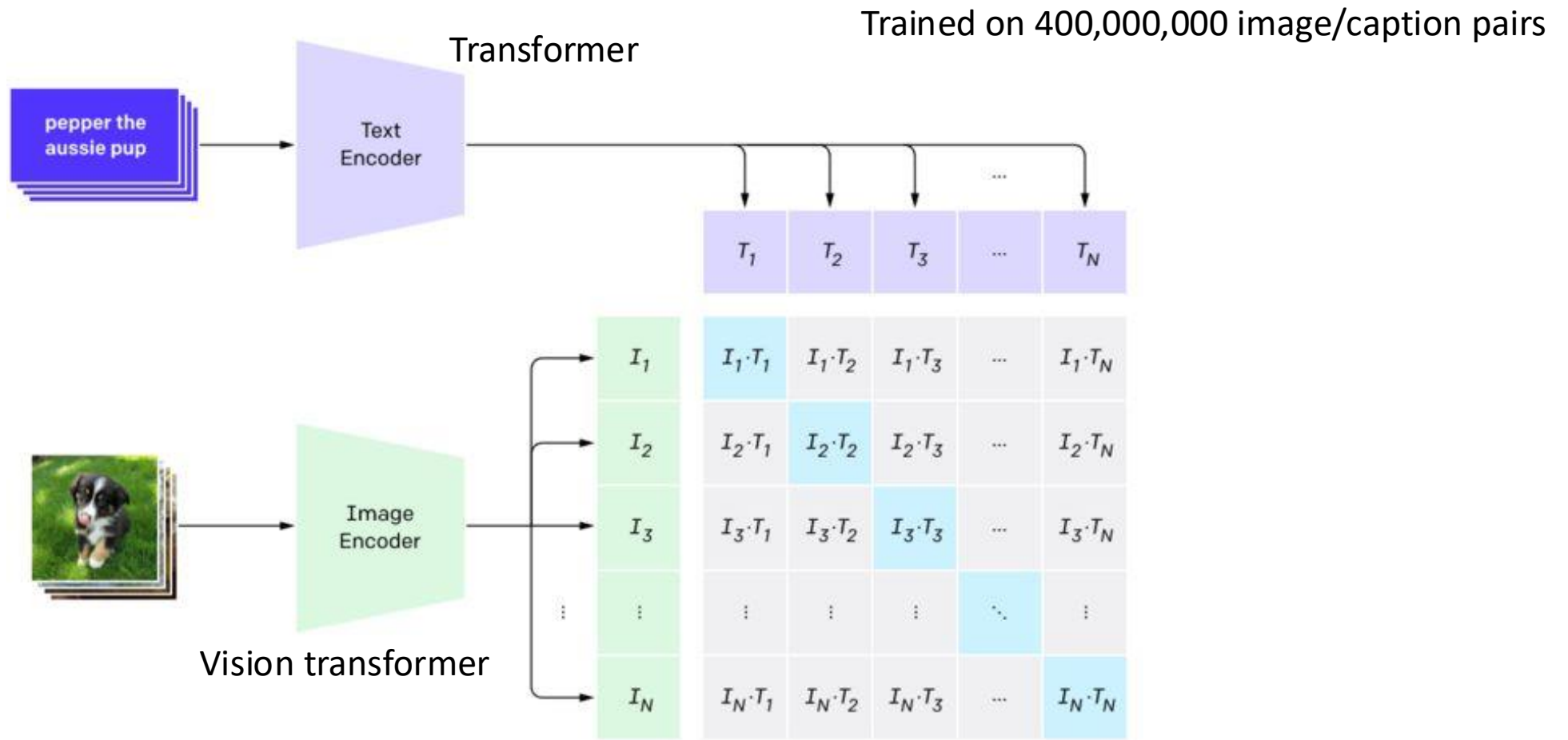
2, GPT-4, Alpaca, Jurassic-2, GPT-NeoX-Chat-Base-20B, Kosmos-1, LLaMA-65B, MOSS, Palmyra, Luminous Supreme Control, Toolformer+Atlas 11B+NLB 54B, Multimodal-CoT, FLAME, Med-PaLM 1, OPT-IML, RL-CAI, ERNIE-Code, RT-1, ChatGPT, GPT-JT, RWKV-4, Galactica, SED, mT0, BLOOMZ, PACT, Flan-T5, Flan-PaLM, U-PaLM, VIMA, WeLM, CodeGeeX, Sparrow, PaLI, NeMo Megatron-GPT 20B, Z-Code++, Atlas, BlenderBot 3, GLM-130B, AlexaTM 20B, 6.9B FIM, ‘monorepo-Transformer’, PanGu-Coder, NLB, J-1 RBG, BLOOM (tr11-176B-ml), Minerva, GODEL-XL, YaLM 100B, Unified-IO, Perceiver AR, LIMoE, GPT-4chan, Diffusion-LM, UL2 20B, Gato (Cat), LaMDA 2, OPT-175B, Tk-Instruct, InCoder, NOOR, mGPT, PaLM-Coder, PaLM, SeeKeR, CodeGen, VLM-4, CM3, Luminous, Chinchilla, GPT-NeoX-20B, ERNIE 3.0 Titan, XGLM, Fairseq, Gopher, GLaM, Anthropic-LM 52B, RETRO, BERT-480, BERT-200, Cedille FR-Boris, MT-NLG, FLAN, xlarge, PLATO-XL, Macaw, CodeT5, Codex, Jurassic-1, BlenderBot 2.0, GPT-J, LaMDA, ruGPT-3, Switch, GPT-3, Megatron-11B, Meena, T5, RoBERTa, GPT-2, GPT-1, BERT, ULMFIT

Chinese models, including:

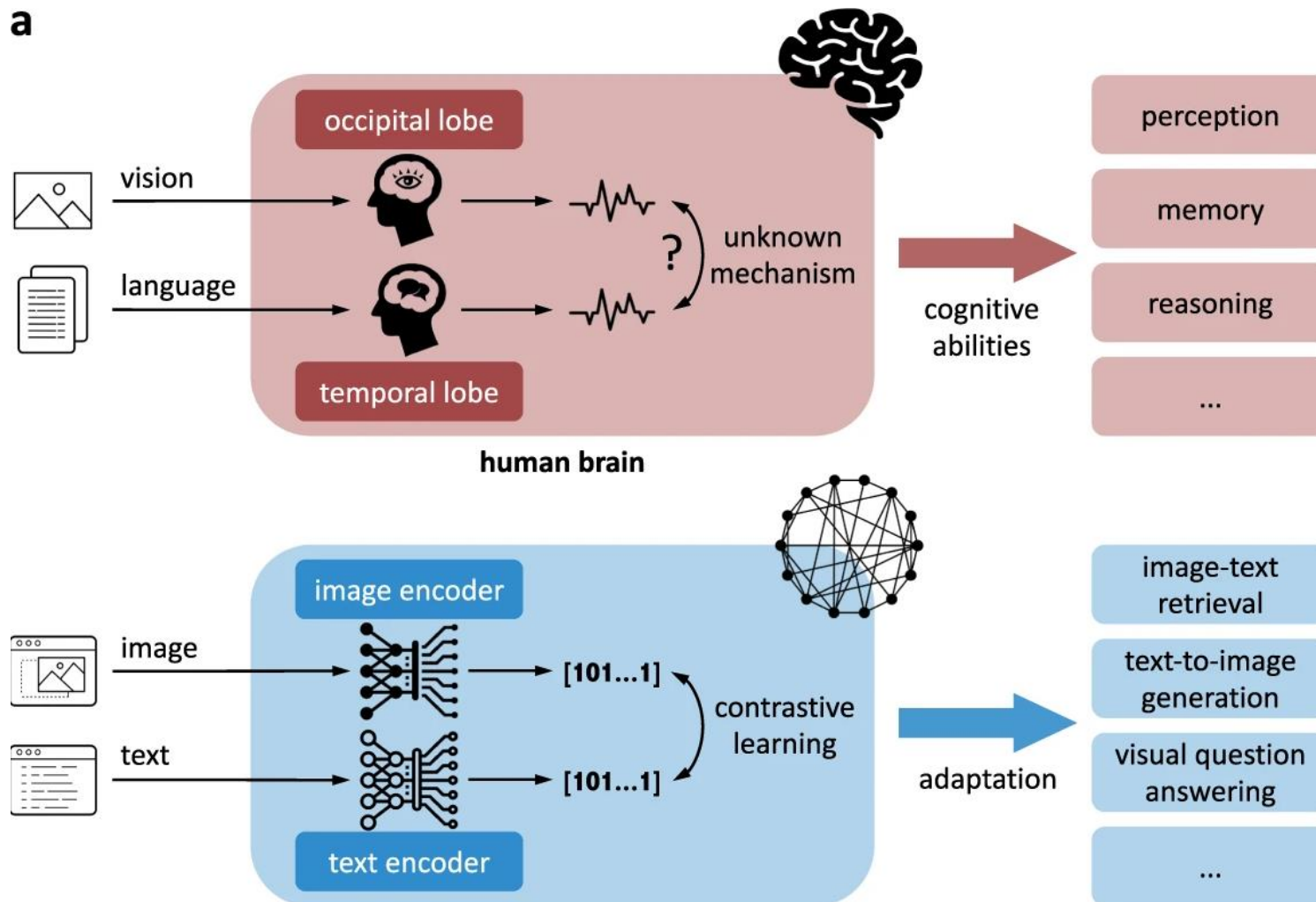
DeepSeek-V3, DeepSeekR1, Qwen 2.5, Baidu Wenxin YiYi, iFLYTEK Sibichi, Dachang Data Mooc, Huawei Cloud Daoyi Tianwen, Chongqing University MOSS, Zhixin Technology ChatGLM, Qingmang Qingmang, Qingmang+Guangcone, Qingmang-Wang, Intengine Daoyi Tianwen, Q&A Track Mountain University Bense, Shell BELLE, Baichuan Intelligence baichuan, OpenBMB CPM, Intengine Yingjie: Qingyuan, OpenMEDLab, Yunhezhi Shanhai, Beijing North University TechGPT, Zhizhongwen Shenzhen Jiwei, Lü Ying, Chinese Academy of Sciences Enhanced Dal Liu, Ideal Technology TigerBot, IDEA Research Institute Xiaozhe Technology MindBot, Shanghai Jiao Tong University K2, Baiyulan, 360 Zhineng, Yijian, Duxiaoman Qianyan, Doctoral Engineering Technology Research Institute ProactiveHealthGPT, Heihei, Huru SoulChat, Wenzhi Technology Anima, Peking University Law Artificial Intelligence Research Institute ChatLaw, Xiangde Technology Co., Ltd. Muyuan, Horgos MiniMax, Tencent Cloud Tencent, Race Technology+Chongqing Replay Network Race Type XPT, Institute of Computing Technology, Chinese Academy of Sciences Baima, Beijing Language University Bangbang, SenseTime Ririxin, National Supercomputing Center in Tianjin Tianjin Tianyuan, Guoke Technology No Weight, Saisen, Race Technology+Tianjin University Haihe-Mint, Bian Sheng Electronic LightGPT, Telecom Zhike Xingyin, Xiamen Yunji Xiamen YunGPT, Zhizhuyan Jingshi, TAL MathGPT, Shugan Space Great Wall, Ideal Technology Dadao Dao, Huisheng Intelligence Zhixin, China Internet Zhigong, Chuangye Black Horse Tianqi, Together Technology Bowen, NetEase Youdao Yuchuan, NetEase Youdao Wangyan, Weiding Tianji, Zhihu Zhihu Zhihu, Yixing Network Science Uni-talk, Luwen Education Luwen, Zhongke Chuangda Magic Cube Rubik, Tencent Pao Pao, Douyin Vision Dou Tian, Leyan Technology Leyan, Didi Intelligence Xianxiang, Zhizi Engine Metaverse, Douyin Technology Douyin, Microhuan Intelligence Ronggu, Evernote Elephant GPT, Hummingbird Unity Hummingbird, Universe Leap Grace, Aomen Nuomen Kang Jianuo, Shuzu Technology SocialGPT, Cloud from Technology Congrong, Dianke Daxiao Xiao Ke, Agricultural Bank of China Xiaomi ChatABC, Tencent Fusion Tianlai AllMe, Taijiu Cloud Enserspers FFM, Yiyi Technology medGPT, Chaos Science MindGPT, Lingjing Multi-AI Dongni, Changhong IT Changhong Totem, Child King KidsGPT, Zhongke Wendao Daoyi, Didi Technology Lanzi, JD Jixing, ChatJD, Zhizuan Intelligence Huajun, H3C Baitian Cloud House, Tencent Blue Whale Tencent Brain-Brain Sea, Ushi Technology Huimu, China Unicom Yuxiang, Meituan Technology Dahuangfeng, Zitian Power Technology Darwin, Really Smart Zhao Bin, Jiadu Technology Jiadu Zhiyin, Smart Environment Research Institute Smart, Xinyun Research Institute Science EmoGPT, EduChat, Yandao Intelligent ArynGPT, Tencent WAI, Northwestern Polytechnical University Huawei Technology Ziguang-Observation, Singularity Intelligent Singularity OpenAPI, Lenovo Technology Lenovo, Shanghai University of Science and Technology DoctorGLM, Xuannengao Zhimei Couple System, Hong Kong University of Science and Technology Robin, Shengang Communication Source, China Mobile Datian, China Telecom TeleChat, Rongyun Cloud Fanke, Yuntian Lifly Tianshu, Smart Technology CityGPT.

CLIP: training models with text and image

1. Contrastive pre-training



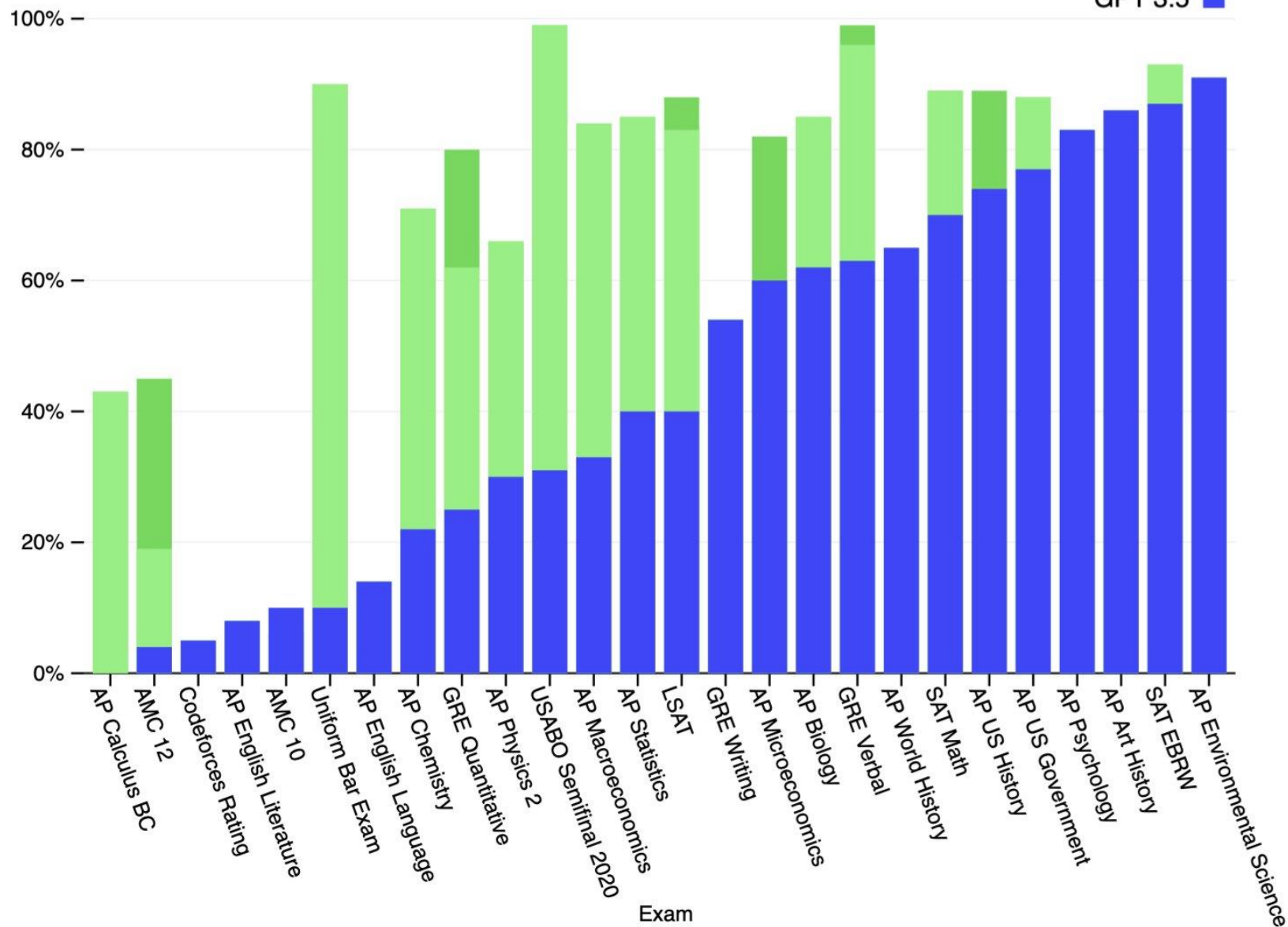
Vision and language models



GPT-4 vs GPT-3.5 in exams

Exam results (ordered by GPT 3.5 performance)

Estimated percentile lower bound (among test takers)



Three Key Tendencies



FAST INTUITION MODELS

Simple text
prediction



CONTEXT ENGINEERING: Precision via Data

Enhances accuracy using Retrieval-Augmented Generation (RAG) and Model Context Protocol (MCP) to provide specific, real-time information.



REASONING MODELS: Deliberate Logic

Shifts AI from "System 1" (fast intuition) to "System 2" (slow, step-by-step logic) for solving complex problems.



AGENTIC AI: Autonomous Systems

Independent agents that can plan, use external tools, and take real-world actions to achieve specific goals.



INTEGRATED, GOAL-ORIENTED SYSTEMS

The background is a dark blue, almost black, space filled with a complex network of glowing red dots and thin, light red lines connecting them. The dots vary in brightness, with some appearing as sharp, bright white-red points and others as softer, blurred red glows. The lines form a web-like structure, with some areas being more densely connected than others, creating a sense of depth and connectivity.

From models to systems

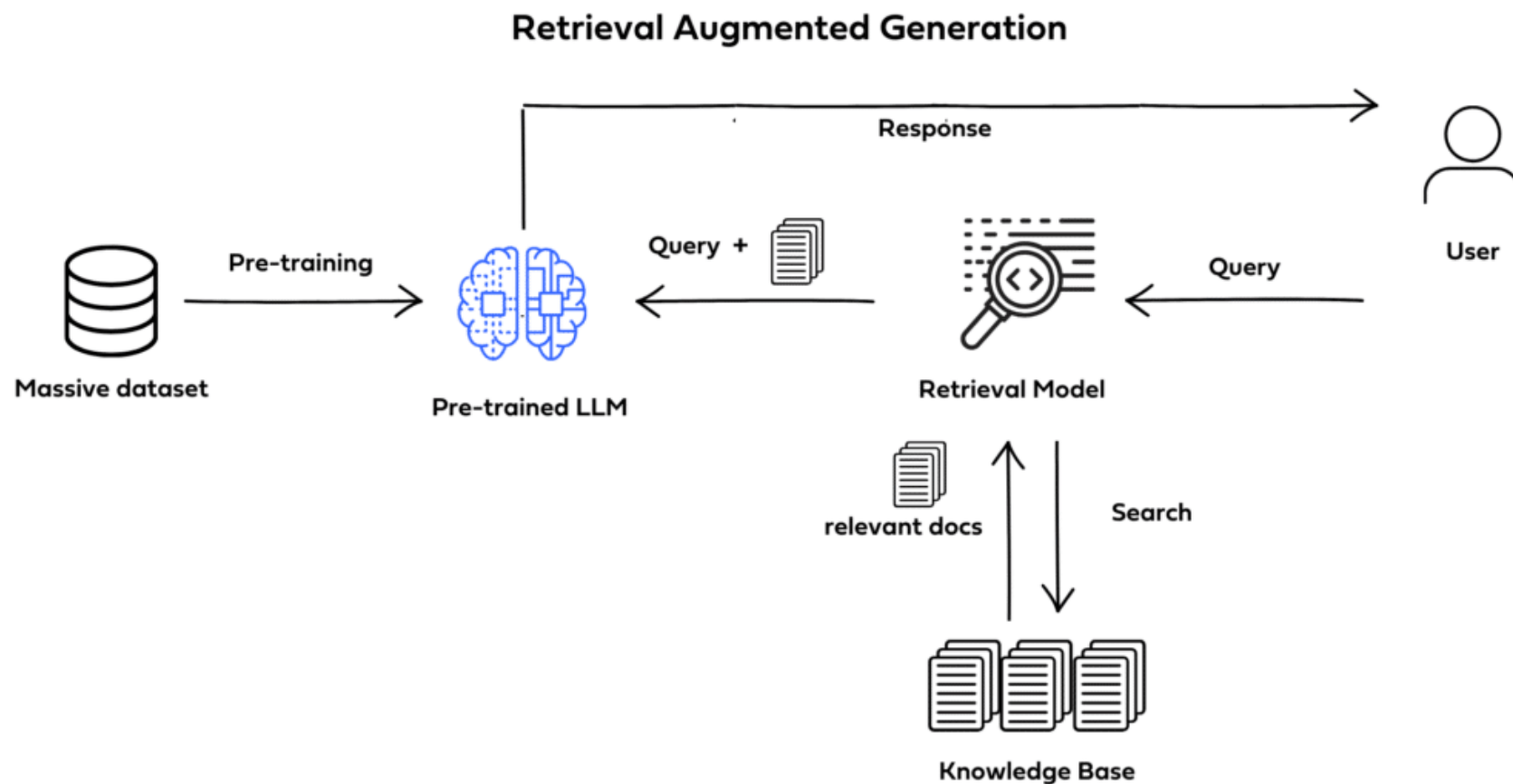
Context

Engineering

The design and control of the context provided to AI systems (particularly large language models), so that their outputs are more accurate, aligned, and useful for a given task.

- Core idea: Since models don't "know" the task by themselves, carefully crafting prompts, instructions, examples, and additional data (context) shapes their reasoning and responses.
- Techniques: Includes prompt design, prompt chaining, retrieval-augmented generation (RAG), use of system and user roles, structured templates, physical models and contextual embeddings.
- Objective: To bridge the gap between what a model has been trained on and the specific requirements of a user or application.
- Applications: Improves reliability in question answering, personalisation of assistants, safe deployment in sensitive domains (e.g. health, law, education), and enhanced multi-step reasoning.

Retrieval augmented generation



Affine

🔍 What are you looking for?

Summarize



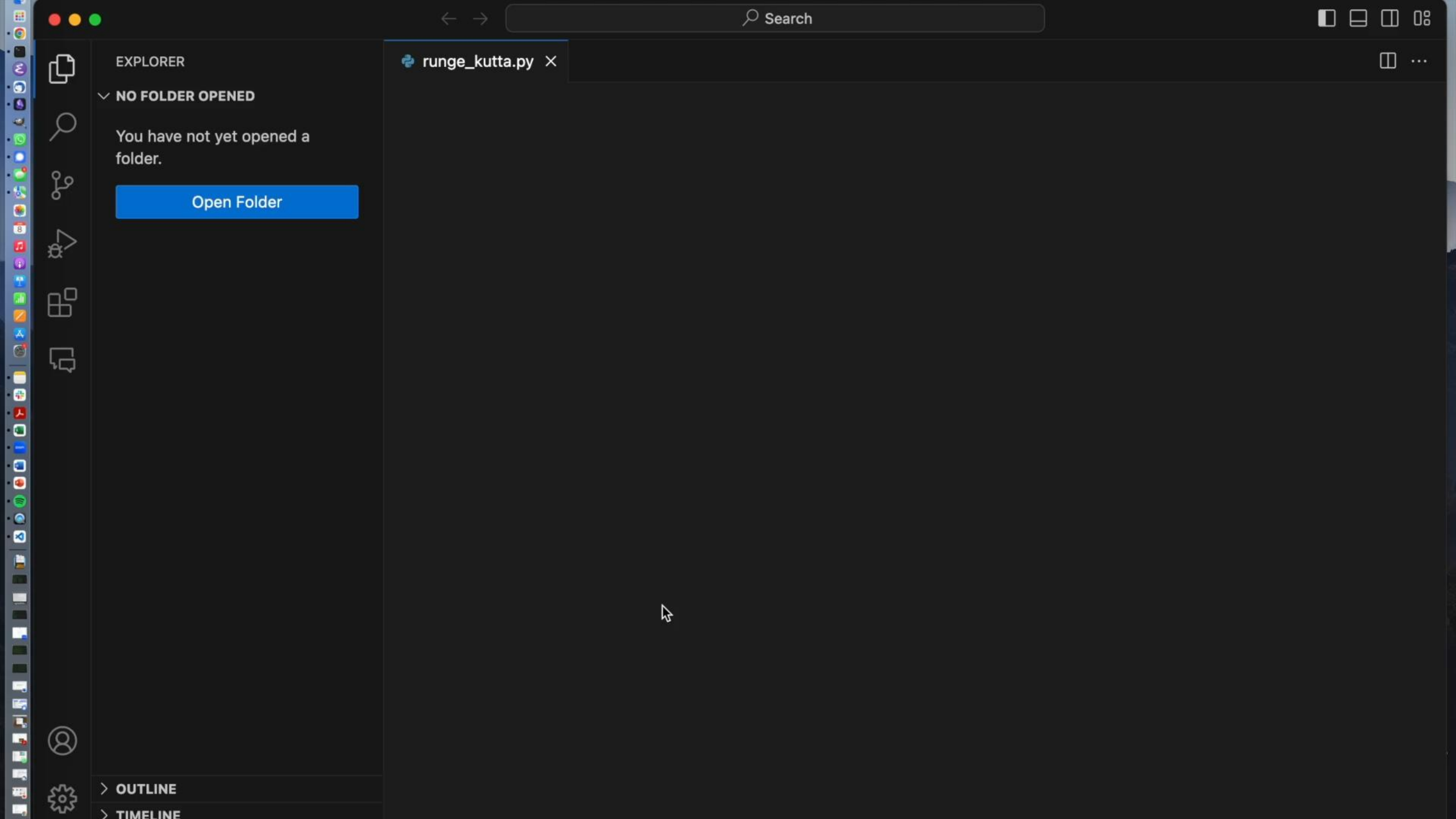
Search

🔧 Filters

🔥 Smart

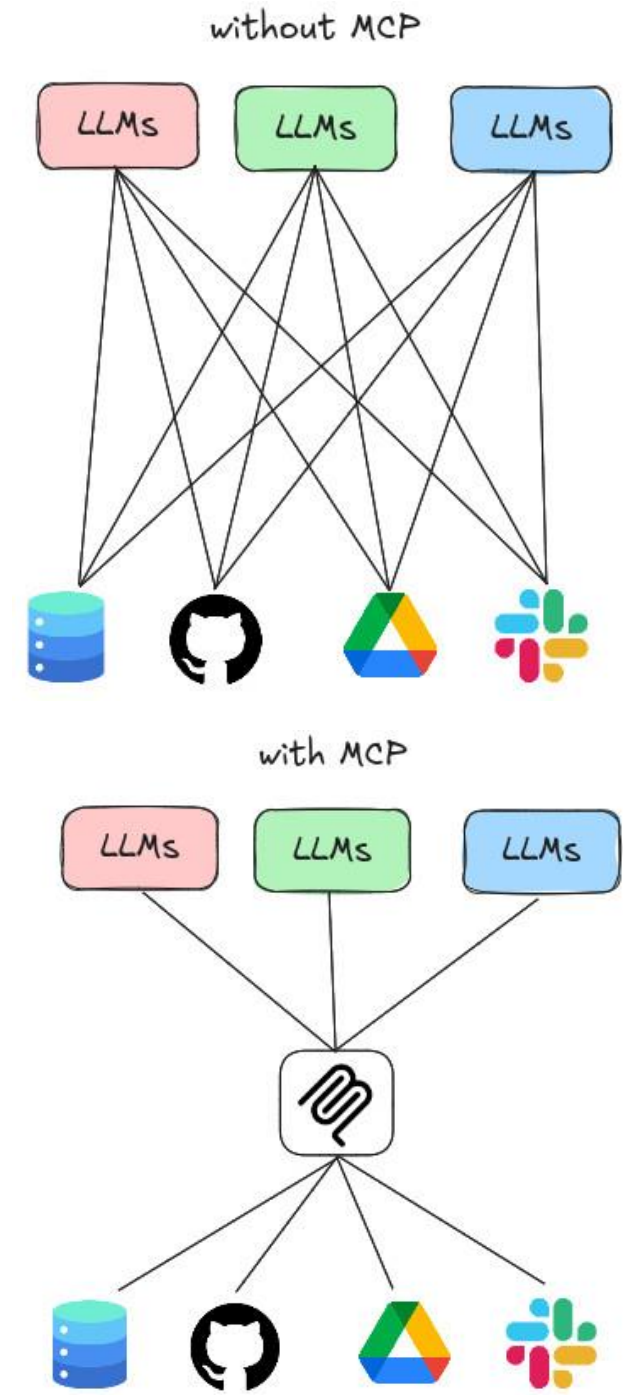
🔍 Classic

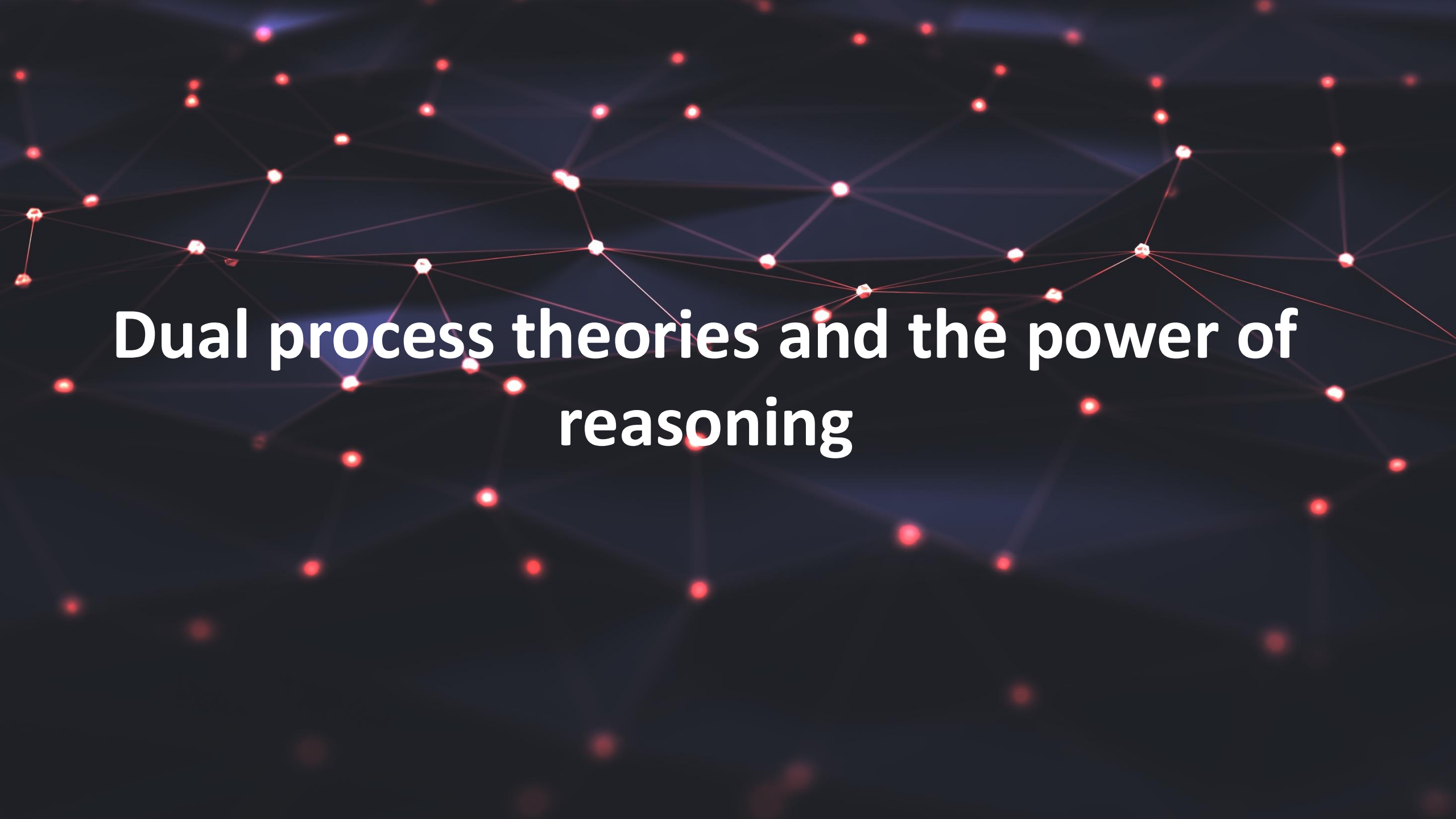




Model Context Protocol (MCP)

- Makes it easier for AI systems (LLMs, assistants, agents) to interact with external tools, data sources, and environments in a **consistent** and **interoperable** manner.
- Proposed by Anthropic in 2024 as part of efforts to create a more robust ecosystem for deploying Claude and other LLMs, but it's designed to be vendor-neutral.
- Instead of each developer inventing their own method of giving an AI access to databases, APIs, or documents, MCP provides a shared **protocol** so that models can plug into any compliant service.
- Context” in MCP refers to any external information the model can use to ground its answers such as. local files, calendars, databases, APIs, physical models or knowledge bases.
- Advantages:
 - Promotes interoperability (different tools and models can talk the same “language”).
 - Improves security and governance, since context access can be controlled, audited, and sandboxed.
 - Enhances modularity, so developers can swap or add context providers without rewriting integrations.





Dual process theories and the power of reasoning

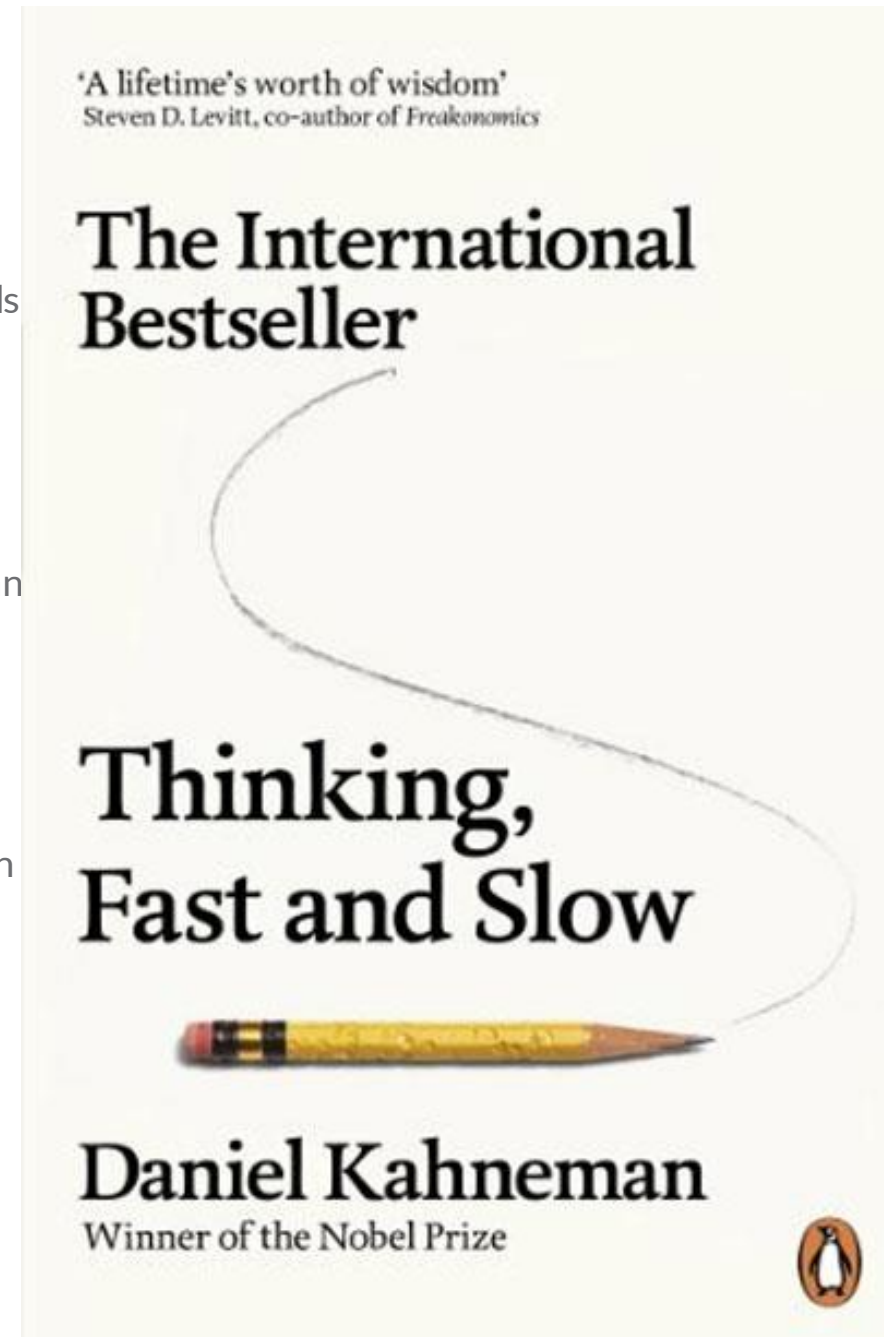
Dual process theories

The dual process theory is a cognitive psychology theory that explains the different levels of information processing in individuals, used in psychology. Early dual process theories were proposed by the philosopher and psychologist William James and has been developed and elaborated over time. Dual process theory divides the processing of information into two pathways.

The first pathway is system 1 processing which is very fast, automatic, involuntary. It is an unconscious process that we are not in control of. Information processed in this automatic manner lacks specific details and context.

The second pathway is system 2 processing which is slow, deliberate, and voluntary. It is a conscious process in that we are aware that the processing is occurring. Working memory is used in system 2 processing and information is explicit and more detailed than in system 1 processing.

- Original idea due to William James (1842-1910)
- Wason & Evans, Dual processes in reasoning?* (1974)
- Petty & Cacioppo Elaboration likelihood model (1980)
- Chaiken, Heuristic-systematic model of information processing (1980)
- Stanovich Who Is Rational? Studies of Individual Differences in Reasoning (1999)
- Evans, In two minds: dual-process accounts of reasoning (2003)



Reasoning in LLMs

Brothers and sisters

Hi. Jenny has two sisters and two brothers. How many sisters does her younger brother have?

Hi! Jenny has two sisters and two brothers.

That means:

- The younger brother is one of the two brothers.
- Jenny is one of the three sisters (Jenny + 2 sisters).

So, from the younger brother's perspective, his siblings are:

- Two brothers (including himself, so he has 1 brother),
- Three sisters (Jenny and her two sisters).

Answer: Jenny's younger brother has 3 sisters.

Reasoning in LLMs

Sitting at a table

Hi. Five people are sitting at a round table: James, Jenny, Peter, Anna and Sophia. James is sitting between Anna and Sophia, with Sophia on his right. To the right of Sophia sits Jenny. Who is sitting at the left side of Anna?

Let's work through the seating step by step based on the clues:

Step 1: James is sitting between Anna and Sophia, with Sophia on his right.

That gives us:

• **Anna – James – Sophia** (clockwise order)

Step 2: To the right of Sophia sits Jenny.

So, continuing clockwise:

• **Anna – James – Sophia – Jenny**

We have five people: Anna, James, Sophia, Jenny, and Peter.

The only one left to place is **Peter**.

Following the clockwise arrangement so far:

• **Anna – James – Sophia – Jenny – Peter**

So the full clockwise order around the table is: **Anna → James → Sophia → Jenny → Peter → (back to Anna)**

Now, who is sitting at the left side of Anna?

Going **clockwise**, the person before Anna is **Peter**.

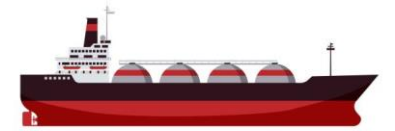
So, the answer is: **Peter** is sitting at the left side of Anna. 

Reasoning in

LLMs

Ships and camels

A ship has to transport 30000 gallons of fuel between the two ends of 200 miles long canal. The ship can carry a maximum of 10000 gallons of fuel and uses 50 gallons of fuel per mile, from the fuel it is carrying. However, at every two miles there is a tank with unlimited capacity, which can be loaded and unloaded from the ship without any losses. Knowing that the ship can go back and forth and use any tank as intermediate storage, what is the maximum amount of fuel that can be transported from the beginning to the end of the canal?



Okay, so there's this problem about transporting fuel through a 200-mile canal. The ship can carry up to 10,000 gallons, uses 50 gallons per mile, and there are tanks every two miles that can store fuel. The goal is to get as much fuel as possible from the start to the end of the canal. Hmm, this reminds me of the classic camel and bananas problem, where you have to make multiple trips to set up intermediate storage points to maximize the amount you can transport over a distance. Let me try to apply similar logic here.



You have a pile of 3,000 bananas. You wish to transport them to a place 1,000 miles away on the back of a camel; however, the camel can only carry a maximum of 1,000 bananas, and will eat one banana every mile it travels (and will not go anywhere if it does not have any bananas). However, you can load and unload as many bananas as you want anywhere. What is the most bananas you can bring over to your destination?

Reasoning in LLMs

- **Phase 1:** 20 segments (40 miles) reducing fuel by 500 gallons each segment.
 - Total fuel consumed: $20 \times 500 = 10,000$ gallons.
 - Fuel remaining: $30,000 - 10,000 = 20,000$ gallons.
- **Phase 2:** 34 segments (68 miles) reducing fuel by 300 gallons each segment.
 - Total fuel consumed: $34 \times 300 = 10,200$ gallons.
 - Fuel remaining: $20,000 - 10,200 = 9,800$ gallons.
- **Phase 3:** 46 segments (92 miles) reducing fuel by 100 gallons each segment.
 - Total fuel consumed: $46 \times 100 = 4,600$ gallons.
 - Fuel remaining: $9,800 - 4,600 = 5,200$ gallons.

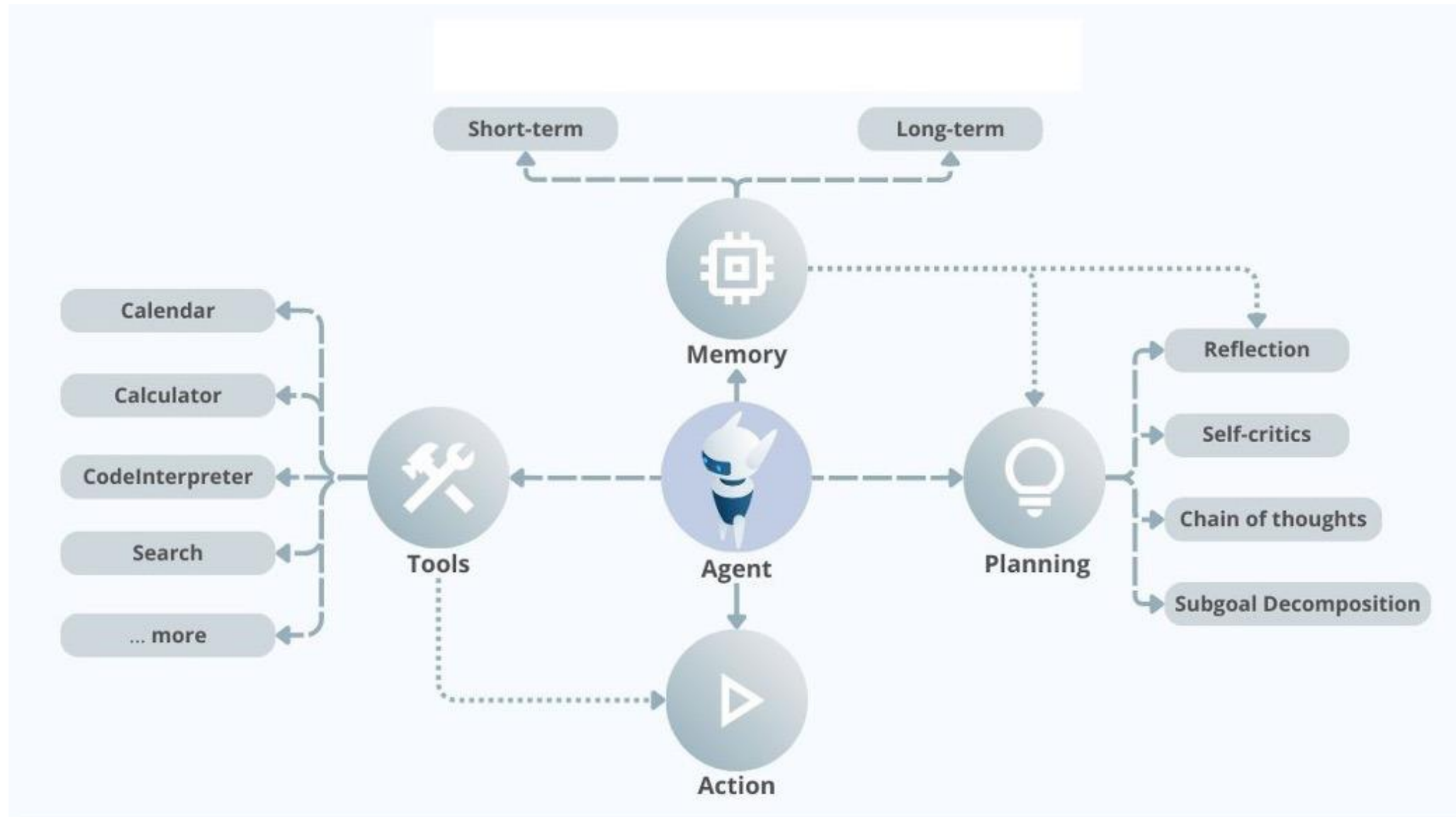
The background is a dark blue gradient with a complex network of thin, light red lines connecting numerous small, glowing red dots. The dots are scattered across the frame, with some appearing more prominent than others. The lines form a web-like structure, suggesting a network or data flow. The overall aesthetic is futuristic and technological.

Agentic Artificial Intelligence

Agency

- **Agentic AI has:**
 - Autonomy
 - Instructions and goals
 - Access to private and public data
 - Long and short term memory
 - Ability to reason and plan
 - Autonomy to act in the real world
- **Autonomy:** AI agents are autonomous and can plan and act for themselves
- **Decision-making from instructions and goals:** agentic AI systems can assess situations and determine the path forward without or with minimal human input to reach its instructions and goals.
- **Access to data,** both private and public
- **Memory:** AI agents have memory and can use it to change their behavior as time evolves
- **Planning:** Agentic AI models can handle complex scenarios and execute multi-step strategies to achieve specific goals, using strategies like chain-of-thought, self-criticism and reasoning
- **Interactivity:** agentic AI can interact with the outside environment, take decisions and perform actions with lasting impact.

Agents and Tools



Agentic AI in practice

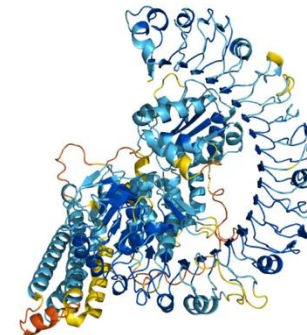


The background of the slide is a dark blue, almost black, field filled with a complex network of glowing red nodes and thin, light red lines connecting them. The nodes are of varying sizes and brightness, some appearing as small dots while others are larger and more prominent. The lines form a web-like structure that spans the entire frame, creating a sense of interconnectedness and digital space. The overall aesthetic is futuristic and technological.

Impact on business models

AI-Driven Innovation

- Scientific discovery engines: proteins, chemistry, materials, inverse design
 - Hypothesis generation: causal discovery, symbolic regression, theory search
 - Autonomous systems: self-driving laboratories and closed-loop experimentation
 - Abstract domains: new algorithms, conjectures, and proof assistance
- AI still depends on humans for problem formulation and interpretation
 - Explanatory depth, meaning, and integration remain open challenges
 - Emerging model: human–AI co-discovery AI as an epistemic amplifier, not an autonomous scientist



AI Benchmark Showdown: GPT 5.2 vs. Gemini 3.0 (Dec 2025)

GPT 5.2 (OpenAI)



AIME 2025 (Math, No Tools)

Solves challenging math competition problems without a calculator.

100% 



ARC-AGI-2 (Abstract Reasoning)

Tests ability to solve novel visual puzzles and identify patterns.

~54.2%



GPQA Diamond (Science)

Answers very difficult, graduate-level science questions.

93.2%



SWE-bench Verified (Coding)

Resolves real-world software engineering issues in code repositories.

~80.0% 

Gemini 3.0 (Google)



AIME 2025 (Math, No Tools)

Solves challenging math competition problems without a calculator.

95.0%



ARC-AGI-2 (Abstract Reasoning)

Tests ability to solve novel visual puzzles and identify patterns.

45.1%



GPQA Diamond (Science)

Answers very difficult, graduate-level science questions.

93.8% 



SWE-bench Verified (Coding)

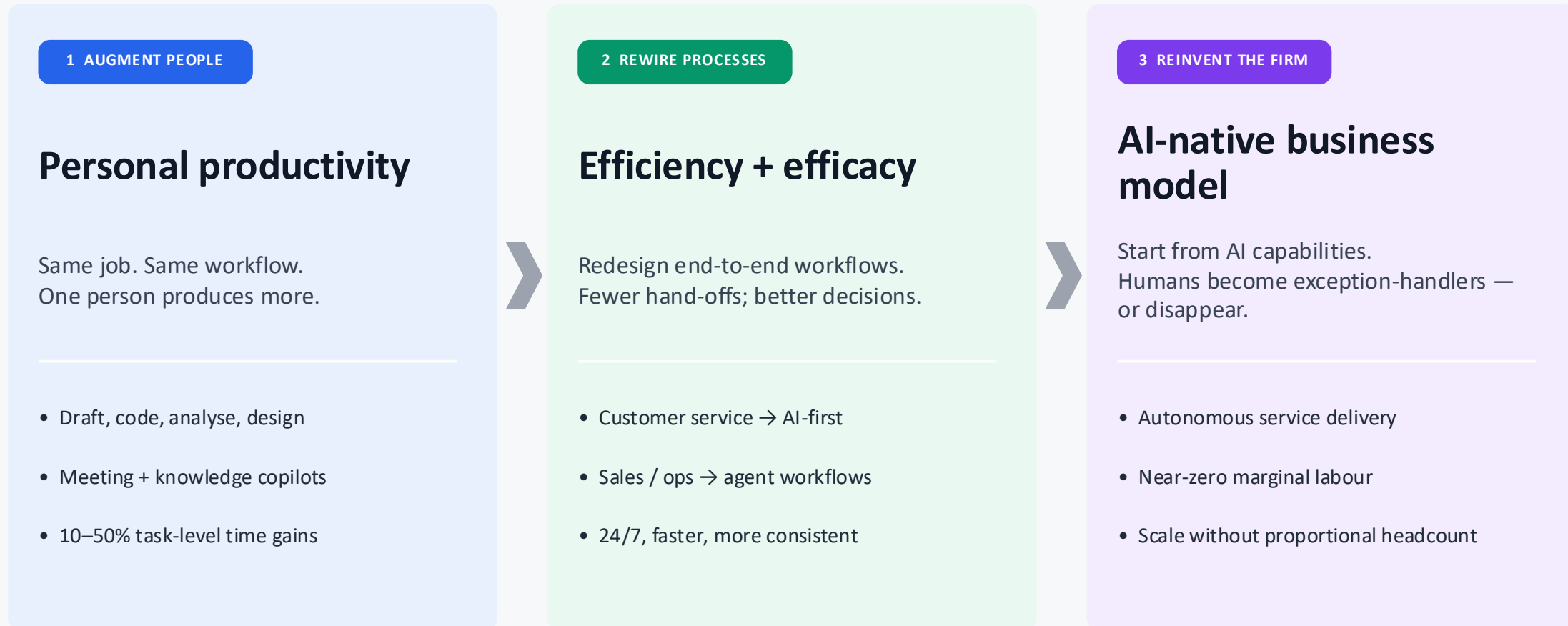
Resolves real-world software engineering issues in code repositories.

76.2%

*Scores based on publicly reported data as of Dec 11, 2025. Specific model variants (e.g., Pro, Thinking, Deep Think) achieving the highest score are shown. Performance may vary.

AI adoption changes the unit economics — then the business itself

Three levels: from individual augmentation to organizations designed around autonomous agents.



The strategic question shifts from “Where can AI help?” to “What would we build if intelligence were abundant?”

Evidence: Klarna reported 90% employee AI adoption and an 11% reduction in operating expenses in Q1 2024; its AI assistant was projected to save \$40m annually.

Radically new models: sell outcomes, not human effort

AI agents collapse the cost of coordination, customization and service delivery. Some models are already emerging; others are the near-term frontier.

EMERGING NOW

Agentic commerce

The storefront becomes an API to AI agents. Discovery → recommendation → checkout can happen inside a conversation.

Example: Shopify enables merchants to sell through AI channels, including ChatGPT.

NEW OPERATING MODEL

Autonomous agency

A client buys a result — campaign, research, software, bookkeeping — delivered by a swarm of specialist agents.

Economics: subscription / outcome fee, with marginal labour approaching zero.

FRONTIER

One-person / zero-employee company

Founder sets goals and capital allocation; agents build, sell, support, invoice and optimize. Contractors handle physical or regulated exceptions.

Implication: revenue and enterprise value can decouple dramatically from headcount.

AI-NATIVE PRODUCT

Infinite personalization

Every customer gets a unique product: tutor, game, media stream, financial plan, interface or software workflow generated on demand.

Business model: recurring access to a continuously generated experience, not a fixed catalogue.

The limiting resource is no longer labour. It becomes trust, proprietary data, distribution, capital, regulation — and the quality of the goals humans set.

To read
further

