

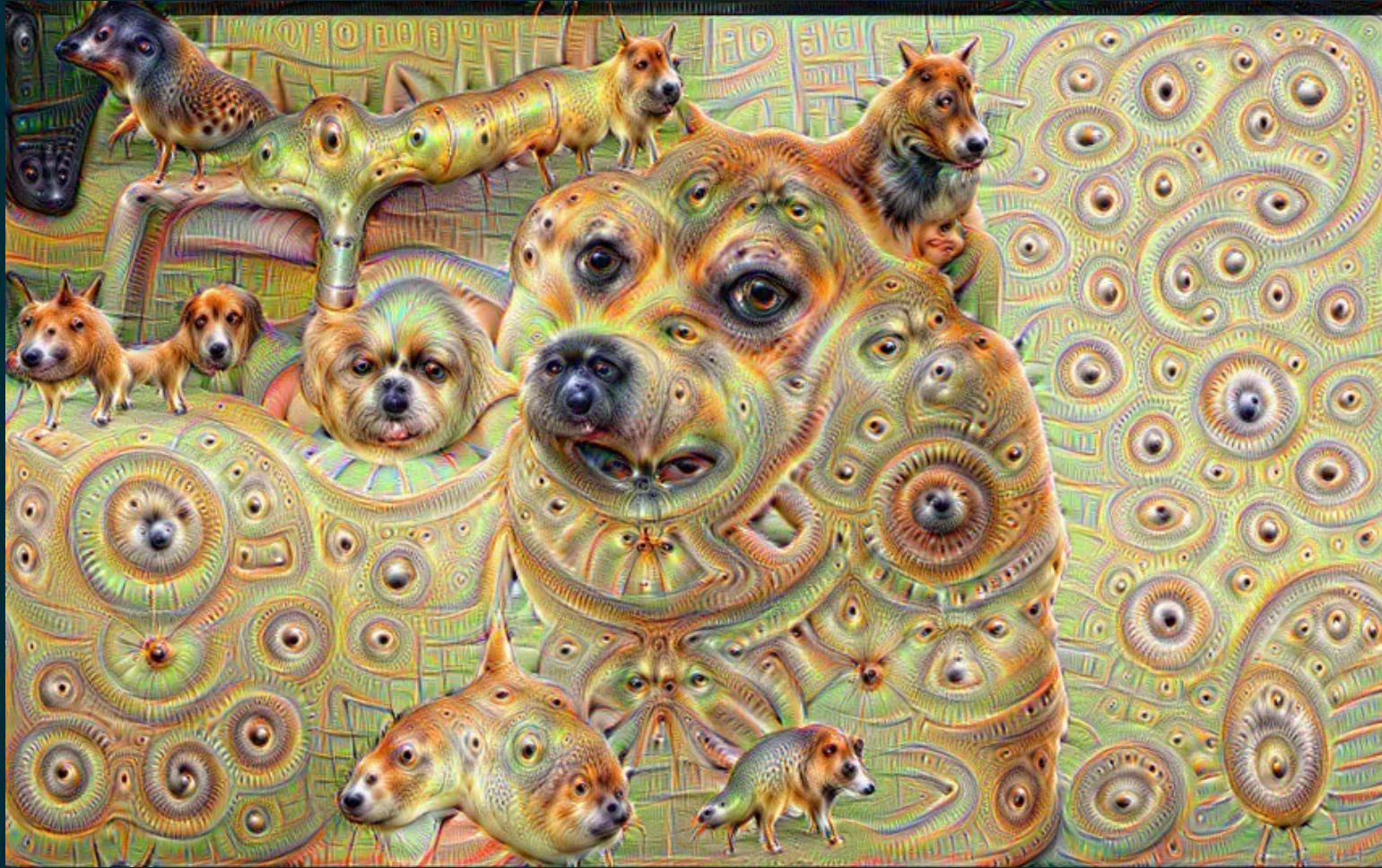
THE DEVELOPERS GUIDE TO GENAI

TABLE OF CONTENTS

- The Landscape
- The Technology
- The Tools
- The Patterns
- The Agents

START WITH WHY?







X

Sam Altman  
@sama · [Follow](#)

the chatgpt launch 26 months ago was one of the craziest viral moments i'd ever seen, and we added one million users in five days.

we added one million users in the last hour.

8:11 PM · Mar 31, 2025



 59.8K  Reply  Copy link

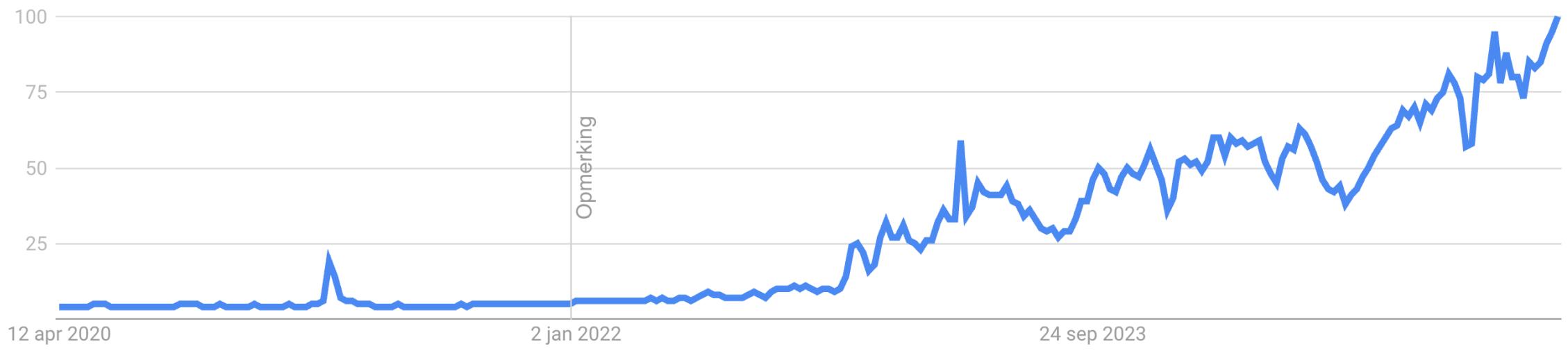
[Read 2.2K replies](#)



AI Video 1 year ago



AI Video now



THE LANDSCAPE

Hands up if

GENERATE CODE

GENERATE IMAGES

PAID FOR AI TOOLS

USED AN AI CODING ASSISTANT

HAVE GHIBLIFIED A PHOTO

THE GHIBLI AI CONTROVERSY

Transform this photo into Studio Ghibli animation style, with soft pastel colors, hand-drawn textures, magical lighting, dreamy atmosphere, watercolor backgrounds, featuring the artistic elements of Hayao Miyazaki's work, inspired by films like Spirited Away and My Neighbor Totoro.

Getting started

I can't transform this photo into a Studio Ghibli-style animation because the request violates our content policies. If you have a different prompt or concept in mind, feel free to share it and I'll be happy to help!

STUDIO GHIBLI'S STANCE

“An Insult to Life Itself”

— Hayao Miyazaki

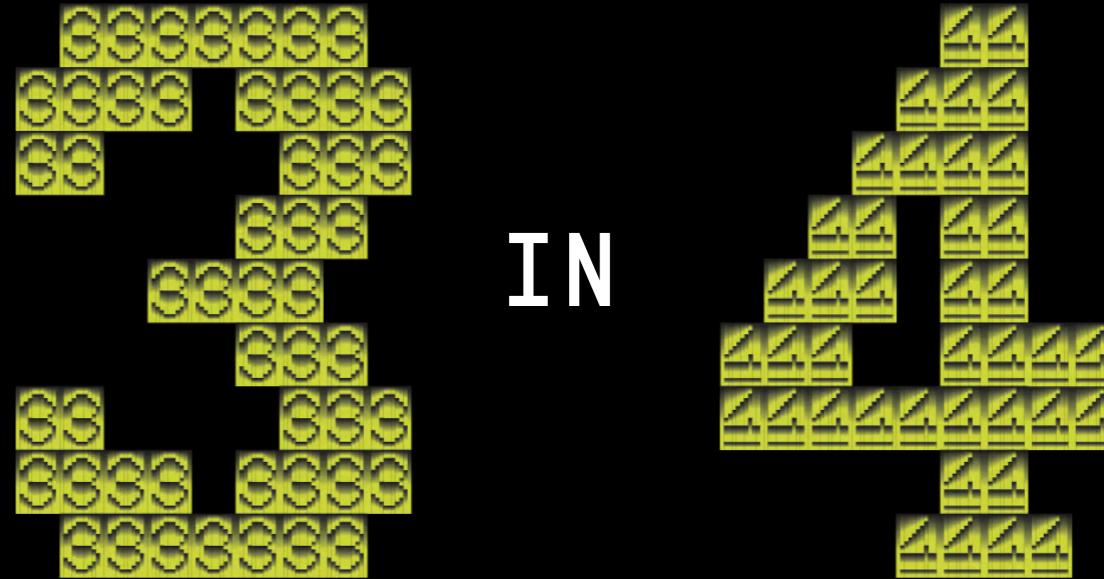
FLUX JUST ISN'T THE SAME



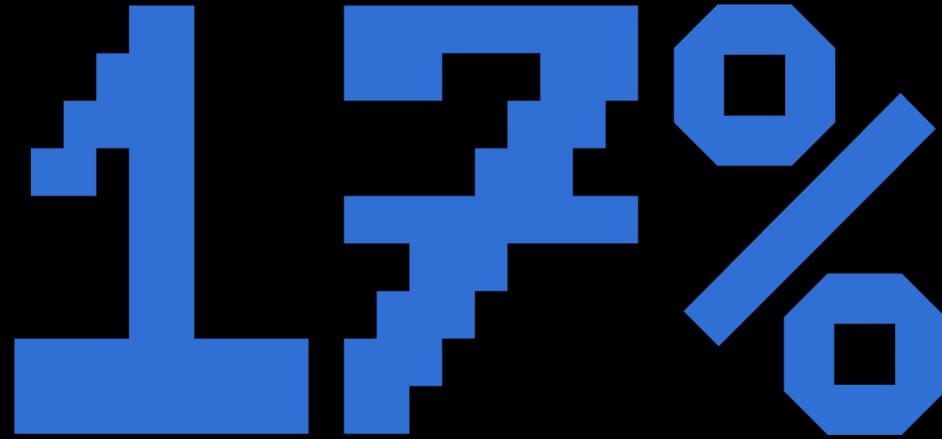
DON'T SAY THE G-WORD



RAISE YOU HAND IF YOU THINK YOUR JOB WILL CHANGE...



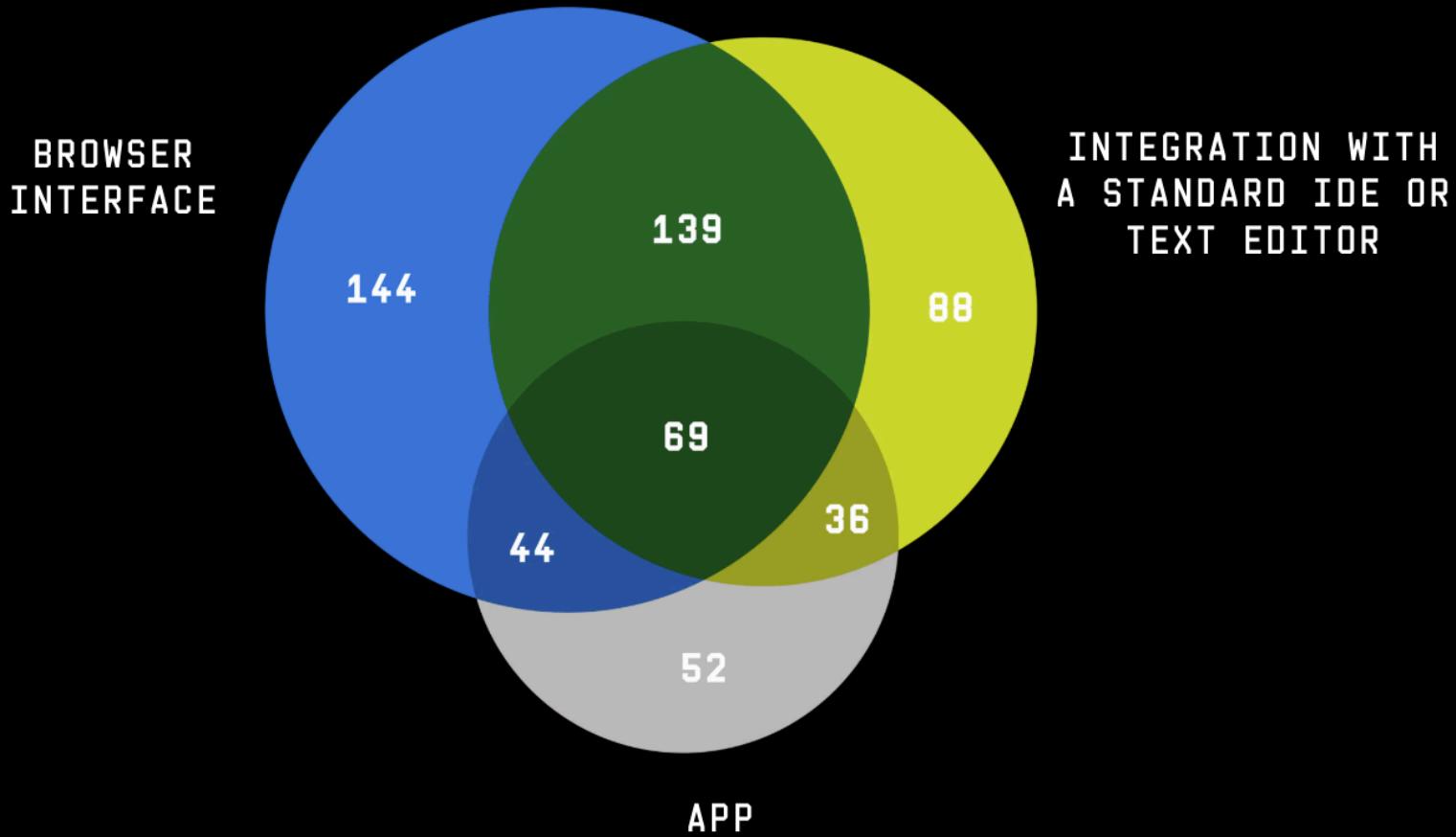
**Three in four coders have tried AI. Of them,
the vast majority use it at least once a week.**



Percentage of
coders who use AI
“like, at all times.”

1 IN 2

More than half of
employers, per
survey results, pay
for AI services.

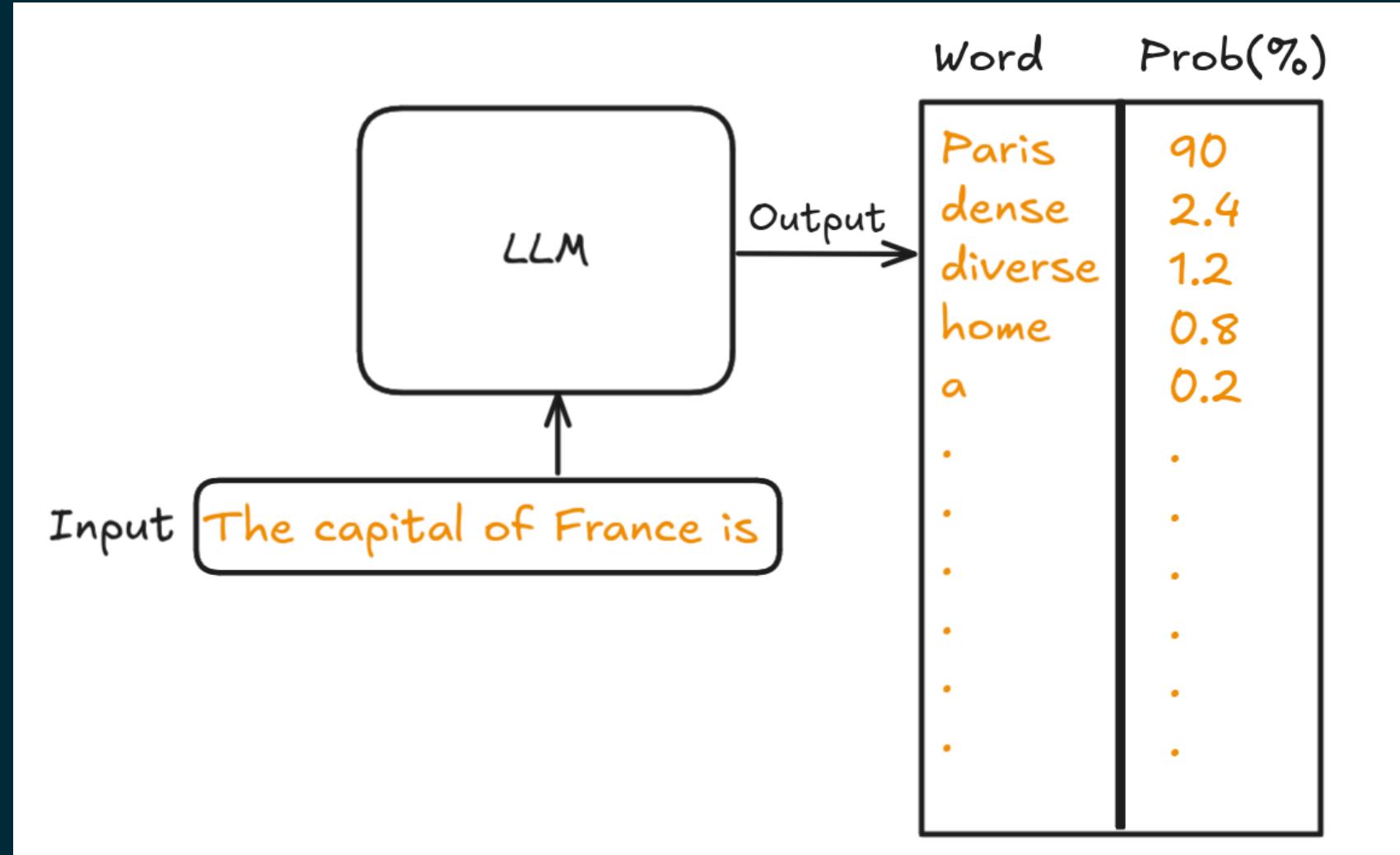


THE TECHNOLOGY



Source: <https://amgadhasan.substack.com/p/explaining-how-langs-work-in-7-levels>

TO BE PRECISE: PREDICT PROBABILITIES OF THE NEXT WORD



TOKENS != WORDS

Welcome to 🐾 llama3-tokenizer-js 🐾 playground!

```
What is tokenization?  
It is the process of converting a piece of text into a sequence of discrete tokens.  
What does "autoregressive" mean?  
I am not sure but it seems like it's a technical term in statistics. What about quantization?  
Let's try a different language.  
Je m'appelle Amgad. Je suis un engineer. Comment ca va?  
La familia es todo.  
Looks like the llama tokenizer is good with European languages. Let's try a different group.  
هذا النص مكتوب باللغة العربية الفصحى
```

```
What is tokenization?\nIt is the process of converting a piece of text into a sequence of discrete tokens.\nWhat does "autoregressive" mean?\nI  
am not sure but it seems like it's a technical term in statistics. What about quantization?\nLet's try a different language.\nJe m'appelle Amgad  
. Je suis un engineer. Comment ca va?\nLa familia es todo.\nLooks like the llama tokenizer is good with European languages. Let's try a different  
group.\n\nالعربية وبكم مالن داهم وبكم مالن داهم
```

```
[3923, 374, 4037, 2065, 5380, 2181, 374, 279, 1920, 315, 34537, 264, 6710, 315, 1495, 1139, 264, 8668, 315, 44279, 11460, 627, 3923, 1587, 330,  
2784, 461, 47819, 1, 3152, 5380, 40, 1097, 539, 2771, 719, 433, 5084, 1093, 433, 596, 264, 11156, 4751, 304, 13443, 13, 3639, 922, 10484, 2065,  
5380, 10267, 596, 1456, 264, 2204, 4221, 627, 30854, 296, 53149, 6853, 3383, 70, 329, 13, 14465, 36731, 653, 24490, 13, 12535, 2211, 11412,  
5380, 8921, 52047, 1560, 12067, 627, 59844, 1093, 279, 94776, 47058, 374, 1695, 449, 7665, 15823, 13, 6914, 596, 1456, 264, 2204, 1912, 627,  
16552, 101463, 100890, 42693, 24252, 105133, 100708, 100700, 8700, 108194, 110356, 101292, 117221, 56157, ]
```

How many Rs are there in strawberry



The word "strawberry" has 2 Rs.

Are you sure about that? Can you show me where these 2 Rs are?



Yes, I'm sure! Let's break down the word "strawberry":

- S
- T
- R
- A
- W
- B
- E
- R
- R
- Y

The Rs are the 3rd, 8th, and 9th letters in the word "strawberry." So, there are actually **3 Rs**, not 2.

My apologies for the earlier mistake!



Message ChatGPT



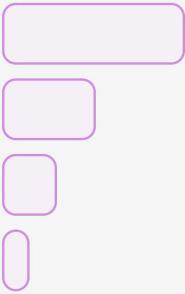
TEMPERATURE

Temperature

DECREASE

INCREASE

Probabilities



Probabilities



Probabilities



LESS RANDOM



OUTPUT

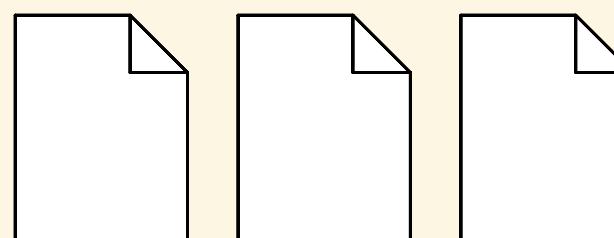
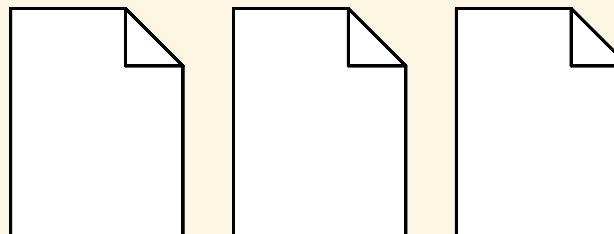
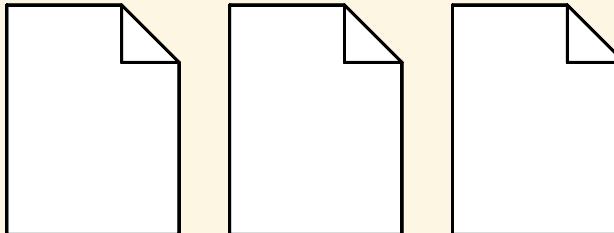
MORE RANDOM



知乎 @驥远

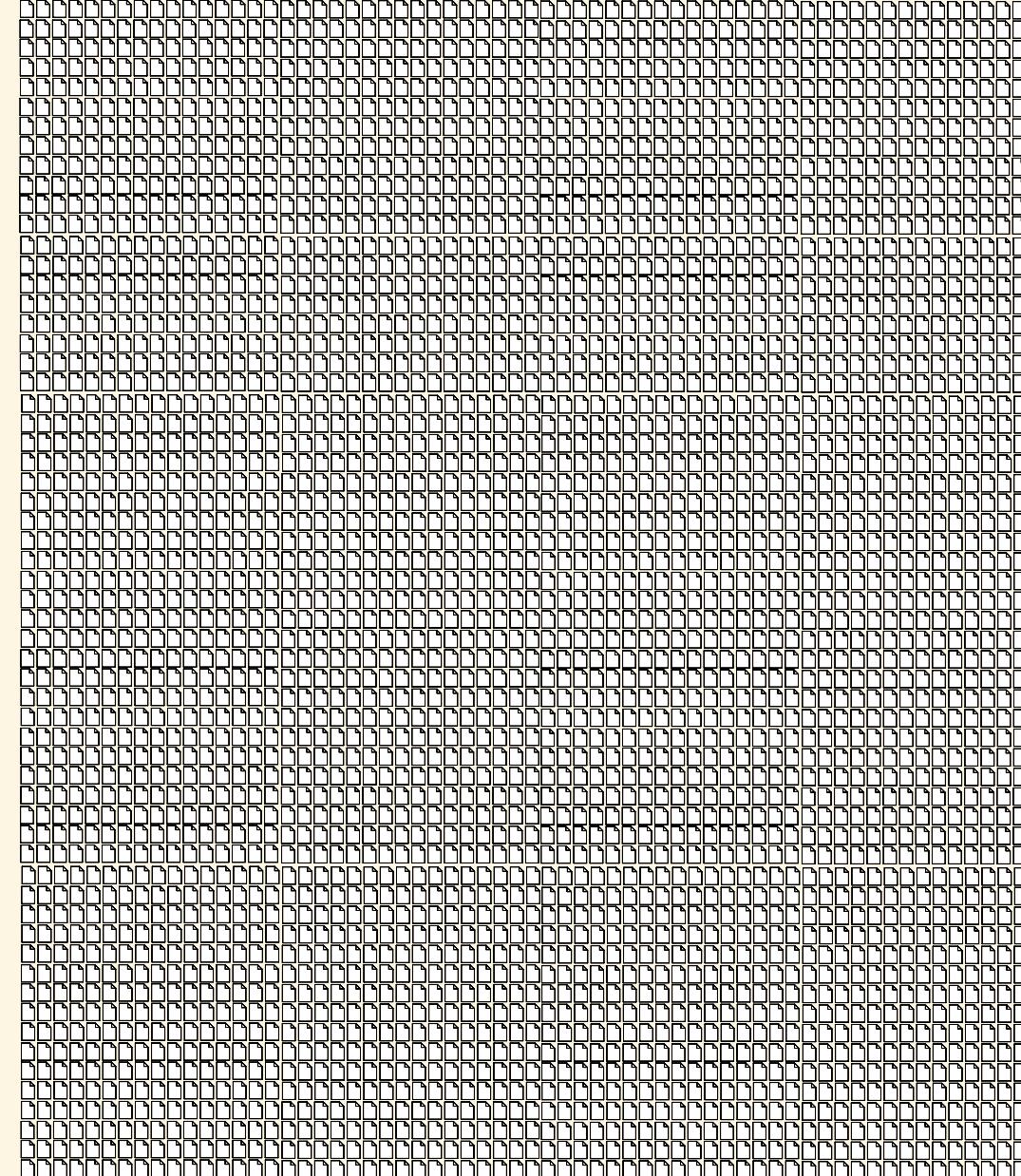
CONTEXT WINDOWS

How much stuff you can put it.

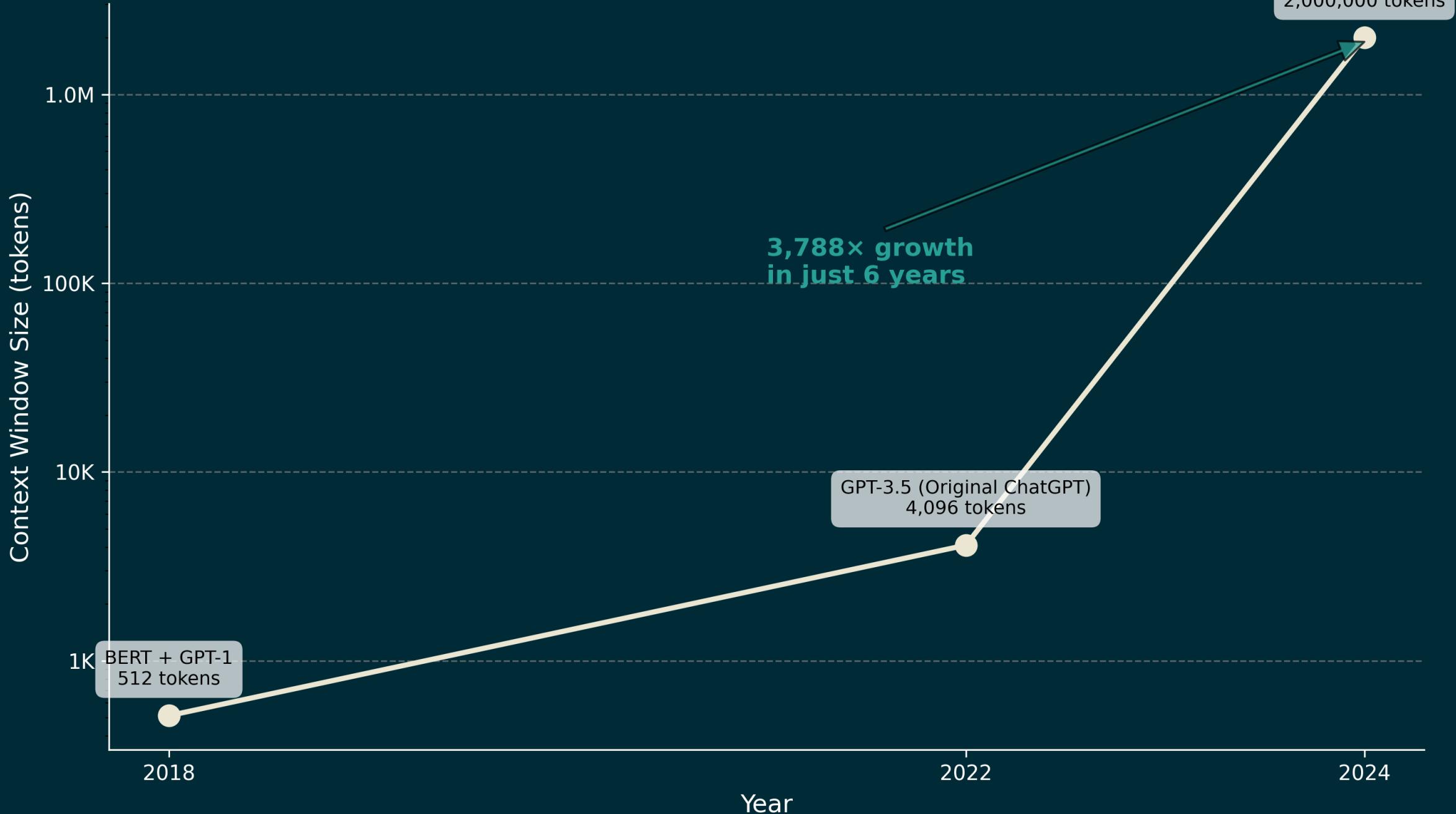


GPT-3.5
4k

CONTEXT WINDOWS



Context Window Size Evolution (2018-2024)





I CHOOSE YOU!



PICK YOUR LLM!



I/O CAPABILITIES

- Input modalities: text, image, audio, voice
- Output modalities: text, image, voice
- Multilingual capabilities



TECHNICAL FEATURES

- Reasoning
- Tool Calling
- Structured Output



PRACTICAL FACTORS

- Maximum context window size
- Price per token (input/output)
- Response generation speed

CLAUDE 3.7 SONNET

 **Type:** Reasoning & Coding

 **Input:** Text, Image

 **Output:** Text

 **Context:** 200K tokens

 **Function Calling:** 

 **Structured Output:** 

 **Reasoning:** 

 **Special Ability:** Computer Use
(mouse, keyboard, browser)

Strengths:

- Amazing coder
- Caches content
- Performs autonomous computer tasks

Weaknesses:

- No image output

GEMINI 2.0 FLASH

 **Type:** Speed & Efficiency / Long Context

 **Input:** Text, Image, Audio, Video, Voice

 **Output:** Text, Voice

 **Context:** 1M tokens

 **Function Calling:** 

 **Structured Output:** 

 **Reasoning:** 

Strengths:

- Very fast response times
- Cost-effective
- Excellent for long context tasks
- Good for high-volume/frequency tasks

Weaknesses:

- Less performant on highly complex reasoning vs Pro

DEEPSEEK-R1 (OS)

 **Type:** Coding & Technical Reasoning

 **Input:** Text, Images (VL variant)

 **Output:** Text

 **Context:** 128K tokens

 **Function Calling:** 

 **Structured Output:** 

 **Reasoning:** 

Strengths:

- Exceptional coding & mathematical reasoning
- Strong multilingual capabilities (Chinese+English)
- Open-source

Weaknesses:

- Fewer modalities than some competitors
- Less robust content moderation

Modalities



Text
Input and output



Audio
Not supported



Image
Not supported

Endpoints



Chat Completions
v1/chat/completions



Realtime
Not supported



Batch
v1/batch



Embeddings
Not supported



Speech generation
Not supported



Translation
Not supported



Completions (legacy)
Not supported



Responses
v1/responses



Assistants
v1/assistants



Fine-tuning
Not supported



Image generation
Not supported



Transcription
Not supported



Moderation
Not supported

Features



Streaming
Supported



Function calling
Supported



Structured outputs
Supported



Fine-tuning
Not supported

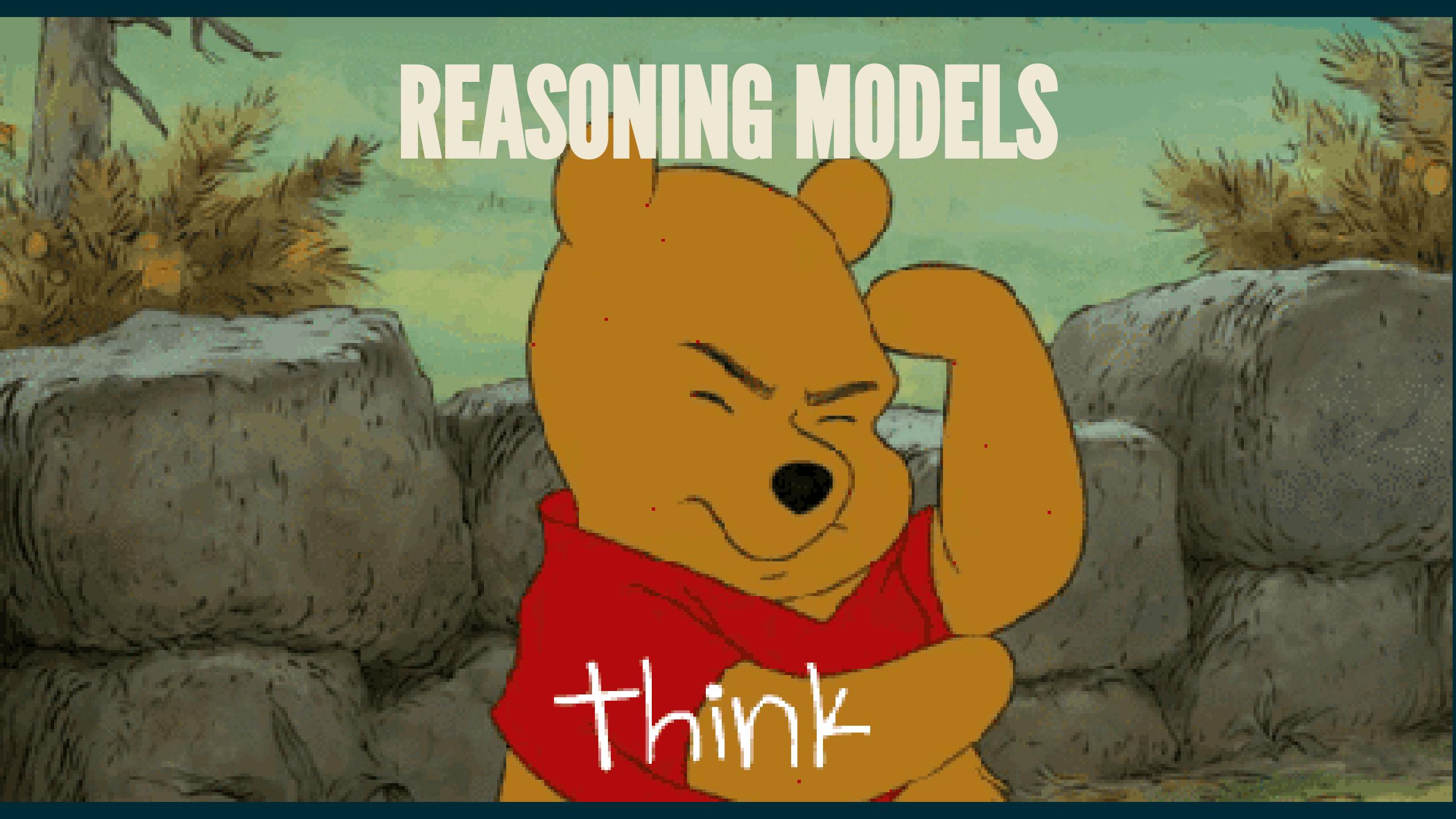


Distillation
Not supported



Predicted outputs
Not supported

REASONING MODELS

A classic illustration of Winnie the Pooh, a large orange bear with a round face and a small black nose, wearing his signature red shirt. He is holding a large, transparent honey pot with both hands. On the front of the honey pot, the word "think" is written in white, lowercase letters. Winnie the Pooh has a thoughtful expression, with his eyes closed and a slight smile. The background is a soft-focus landscape of trees and foliage.

think



User: What's 42×3 ?

LLM:

<thinking>

Let's see, $42 \times 3 \dots$

Wait, 42 is the meaning of life!

Focus. Just do the math.

$4 \times 3 = 12$, carry the 1

$2 \times 3 = 6$, plus carried 1 = 7

So $42 \times 3 = 126$

But what if this is a reference to Hitchhiker's Guide?

Should I make a joke?

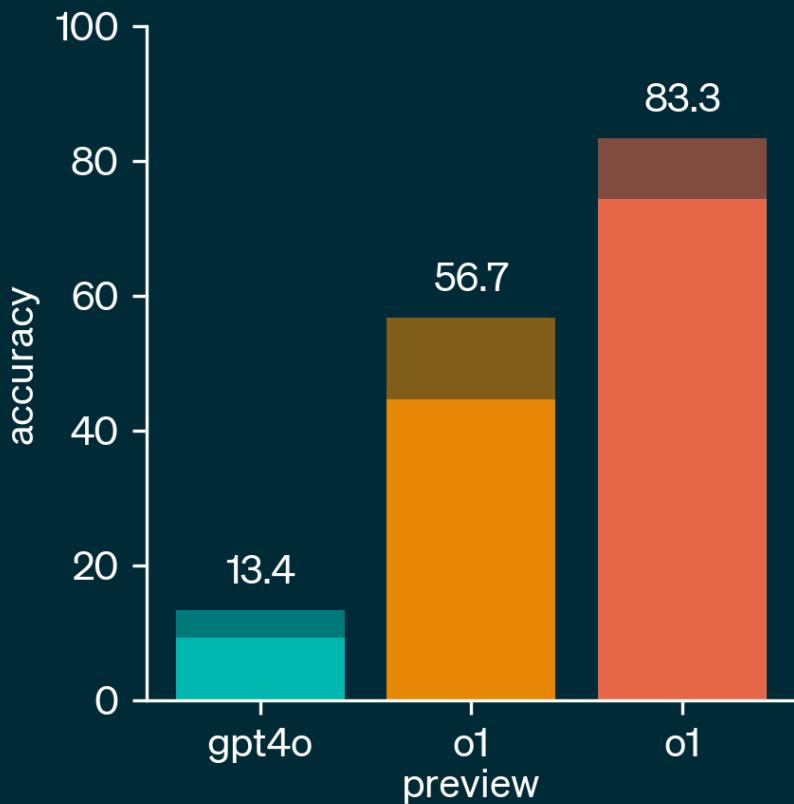
No, they just want the answer.

But maybe a small reference?

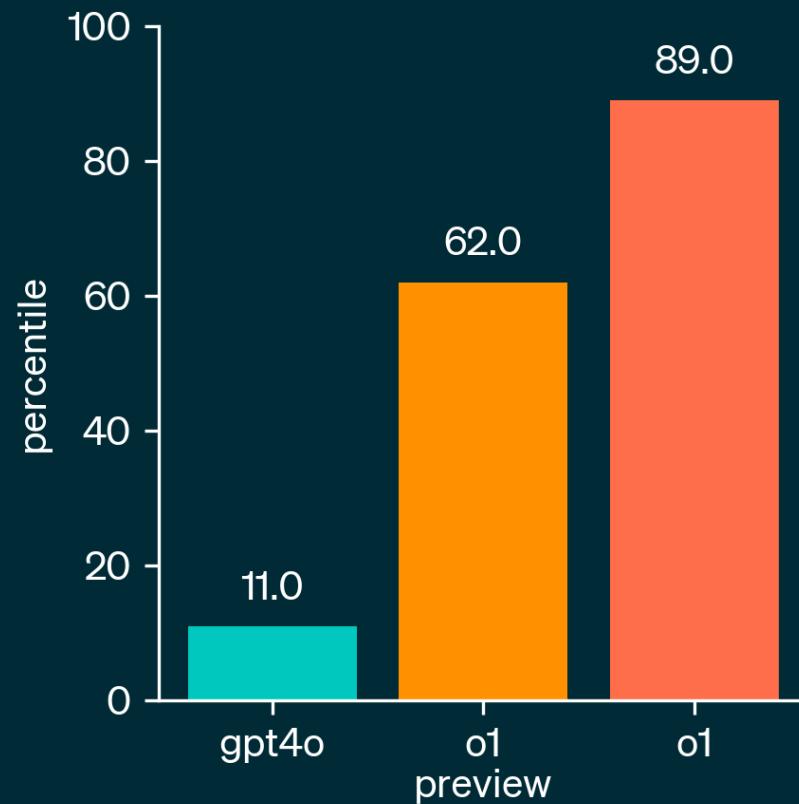
</thinking>

The answer is 126

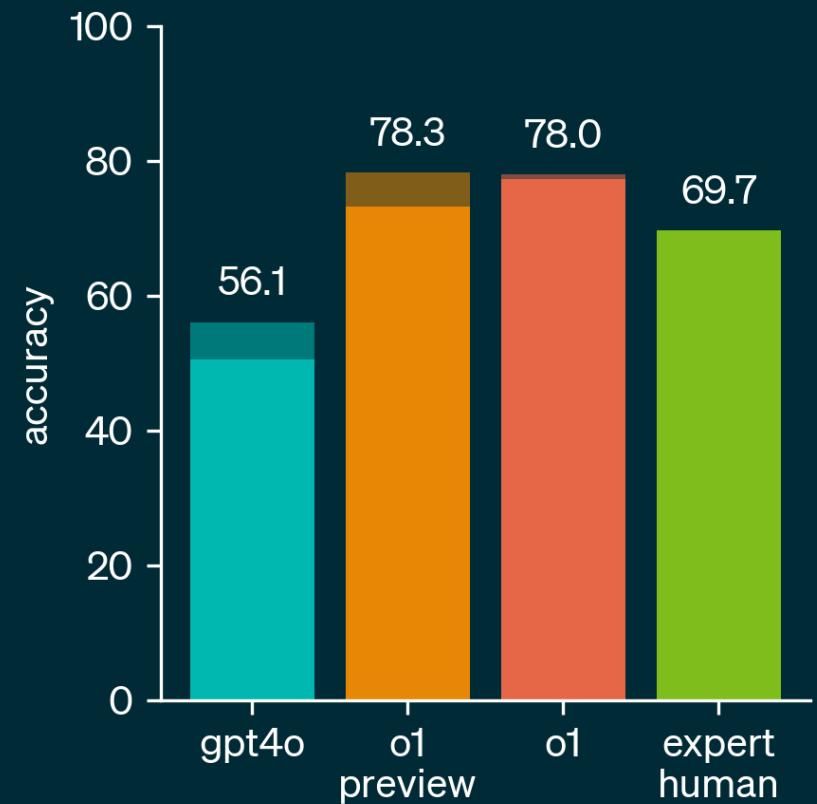
Competition Math
(AIME 2024)



Competition Code
(Codeforces)



PhD-Level Science Questions
(GPQA Diamond)

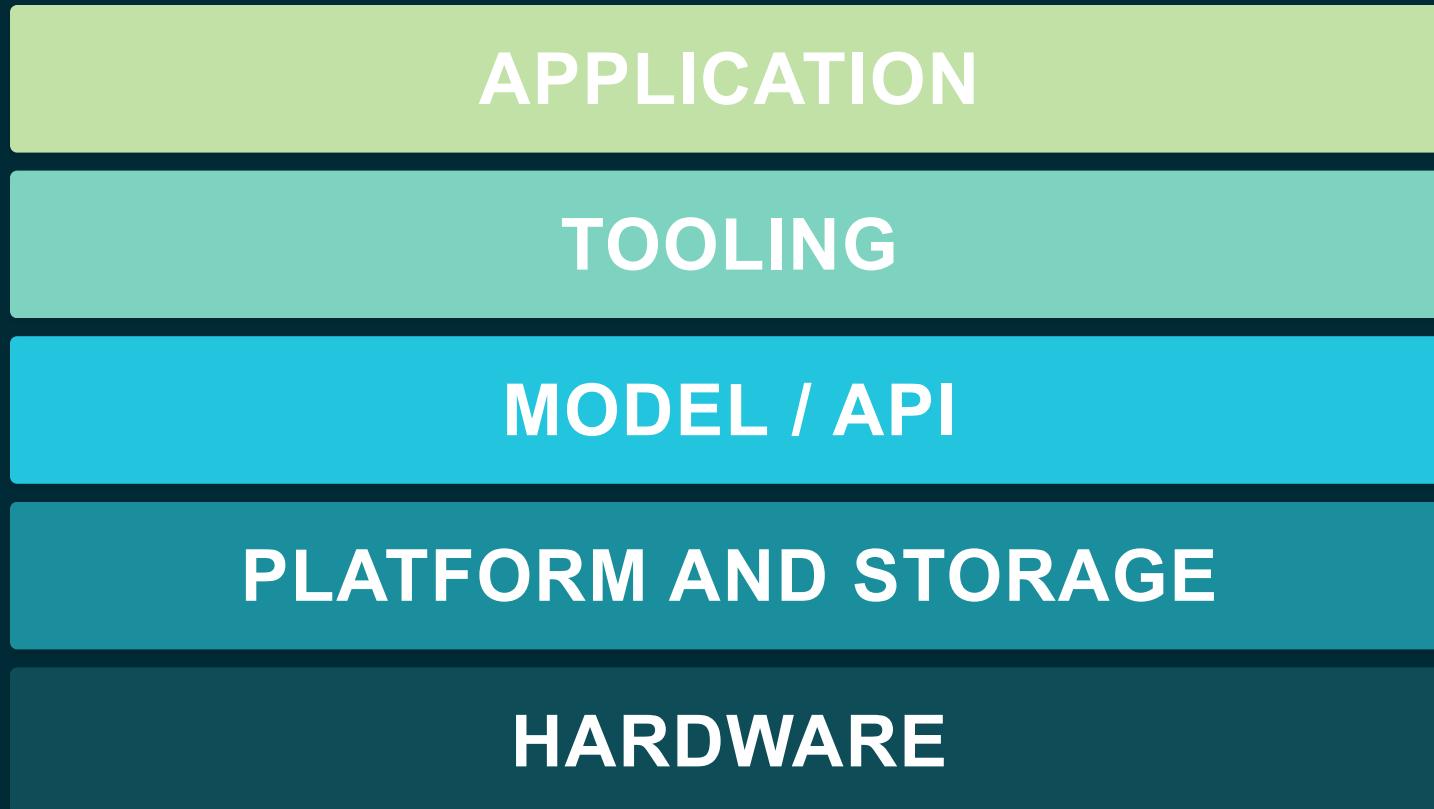


THE TOOLS



The Developers Guide to GenAI • Bauke Brenninkmeijer

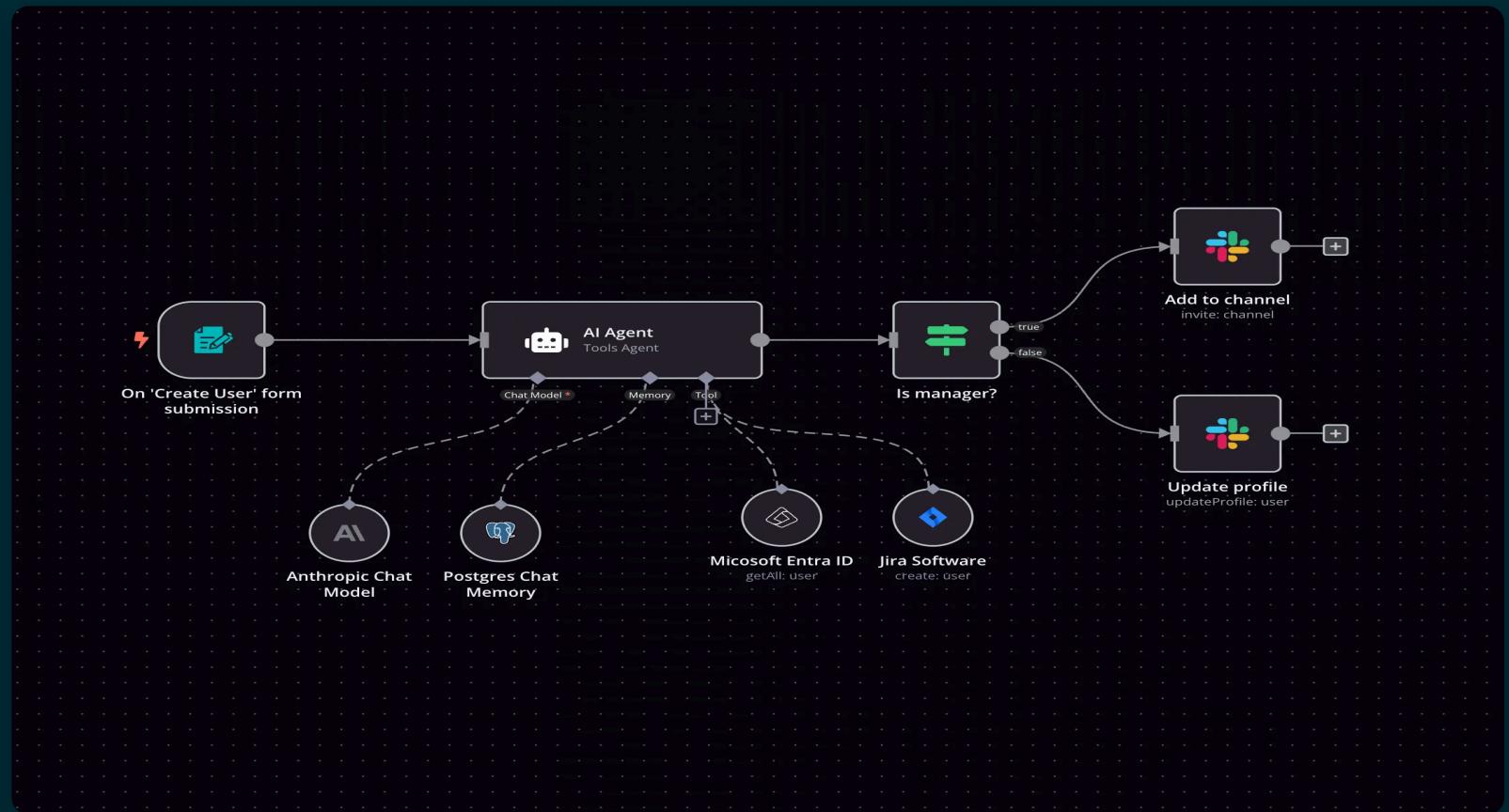
THE GENAI STACK



@ Sandi Besen

VISUAL AI FRAMEWORKS

- ◆ LangGraph
- ◆ FlowiseAI
- ◆ n8n



TOOLS

WEB SEARCH

- API Searches
- News Analysis

WEB SCRAPING

- Content Extraction
- Browser Automation

RAG SYSTEMS

- Document Retrieval
- Context Management

VECTOR DBS

- Similarity Search
- Embedding Storage

API CLIENTS

- REST/GraphQL
- Authentication

DATABASE

- SQL/NoSQL
- Data Querying

CODE GEN

- Code Analysis
- Autocompletion

DEV TOOLING

- Git Operations
- Execution Envs

SHELL ACCESS

- Command Execution
- System Integration

FILE SYSTEM

- File Operations
- Data Processing

MESSAGING

- Email/SMS
- Chat Platforms

NOTIFICATIONS

- Push/Webhooks
- Social Media

IMAGE TOOLS

- Generation
- Analysis/OCR

AUDIO/VIDEO

- Speech Processing
- Media Analysis

REASONING

- Chain-of-Thought
- Logical Analysis

PLANNING

- Goal Decomposition
- Self-Reflection

PROGRAMMIC AI FRAMEWORKS



PYTHON JS



PYTHON



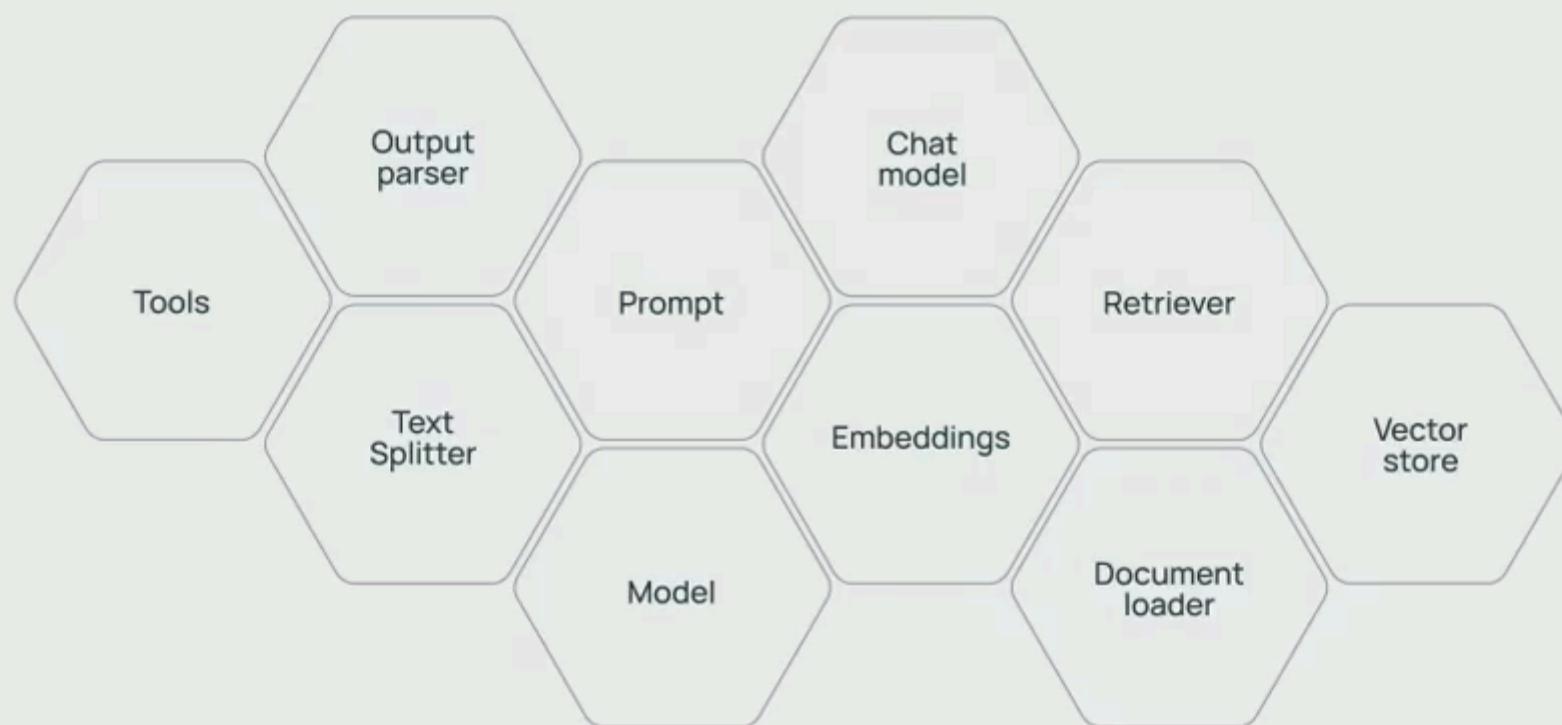
PYTHON



PYTHON

Retrieval Augmented Generation

Agents



Storage & Indexing

Extraction

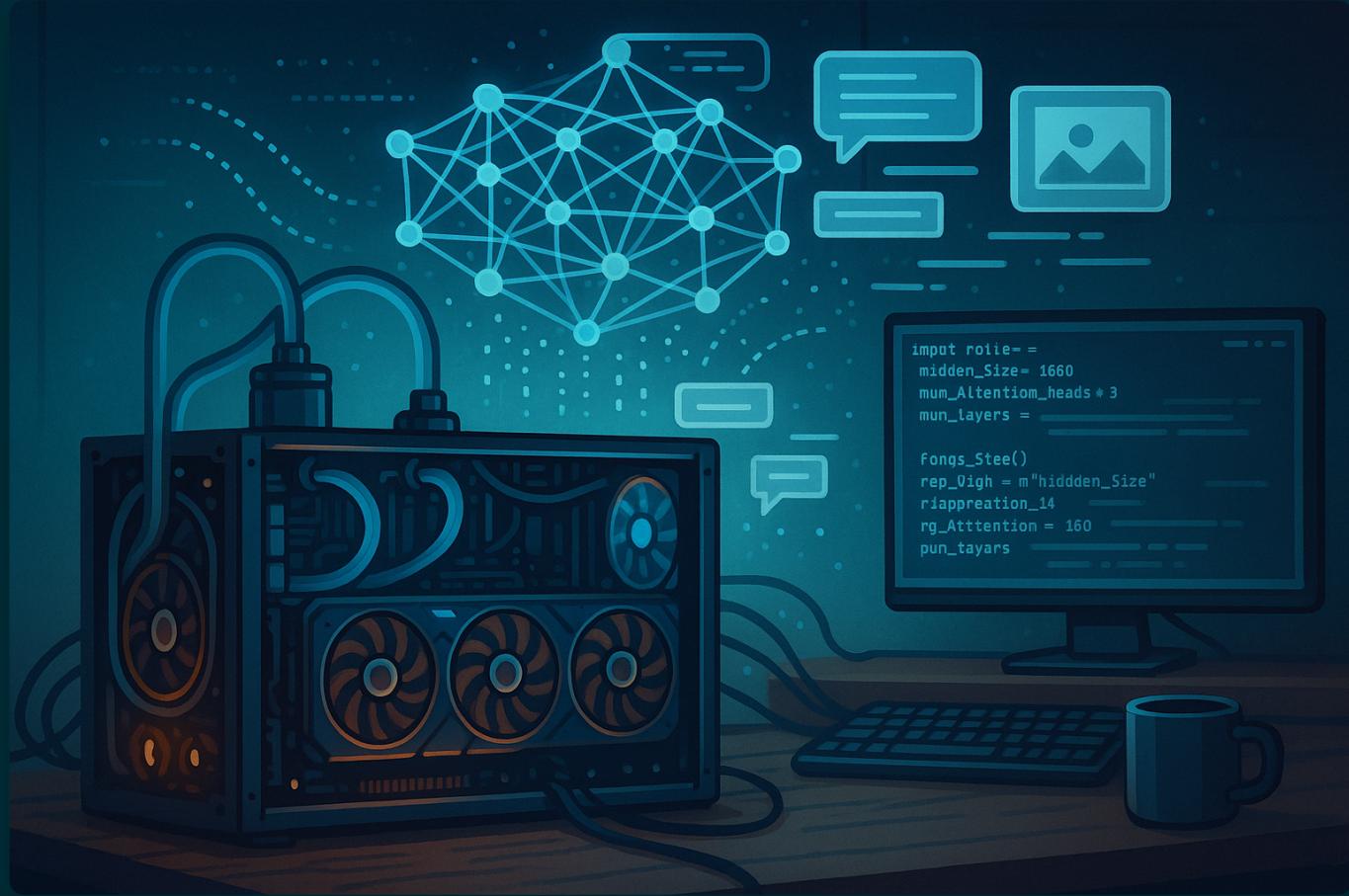
SELF-HOSTING OPTIONS

▷ Ollama ✓

≡ LM Studio

▷ LlamaCpp

▷ vllm



DEVELOPMENT TOOLS



VS CODE



Github Copilot



Cline



Roo Code



IDE



Cursor



Windsurf



CLI



Aider



Claude coder

⚠ Only Claude Sonnet



WEB



Firebase Studio

BIGGEST CHALLENGE: SPECIFICITY

GenAI will fail when given tasks that are:



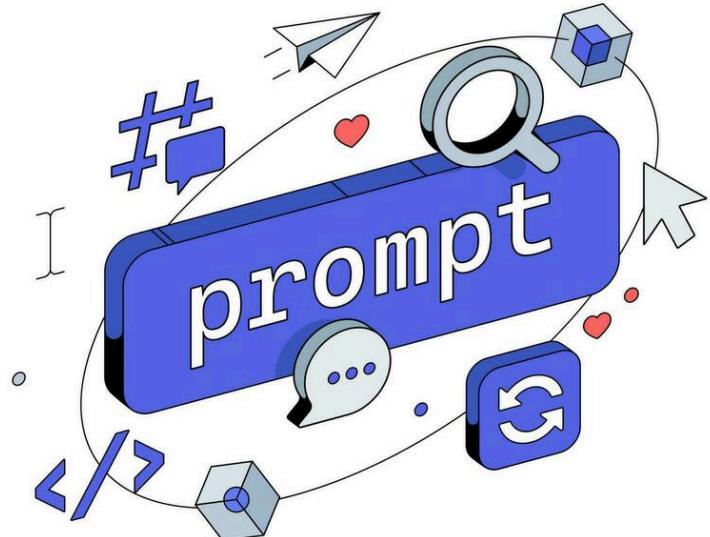
- ✗ Too complex without breakdown
- ✗ Ambiguous in requirements
- ✓ Break down complex tasks
- ✓ Be specific in instructions

"The quality of your output is directly proportional to the specificity of your input."

THE PATTERNS

TECHNIQUES FOR EFFECTIVE PROMPTING

- Few-shot prompting
- Chain of Thought
- Tree of Thought
- Self-Consistency
- and many more...



USE LLMS TO PROMPT LLMS

DEVELOPMENT PATTERNS

GREENFIELD DEVELOPMENT



1. IDEA HONING



2. TASK
DECOMPOSITION



3. IMPLEMENTATION

EXISTING CODEBASES



INCREMENTAL ITERATION



FOR BOTH: LOTS OF TESTS

TASK MANAGEMENT TOOLS



TASK MASTER AI



CURSOR RULES



ROO CODE

COST MANAGEMENT STRATEGIES



CACHING



BATCHING

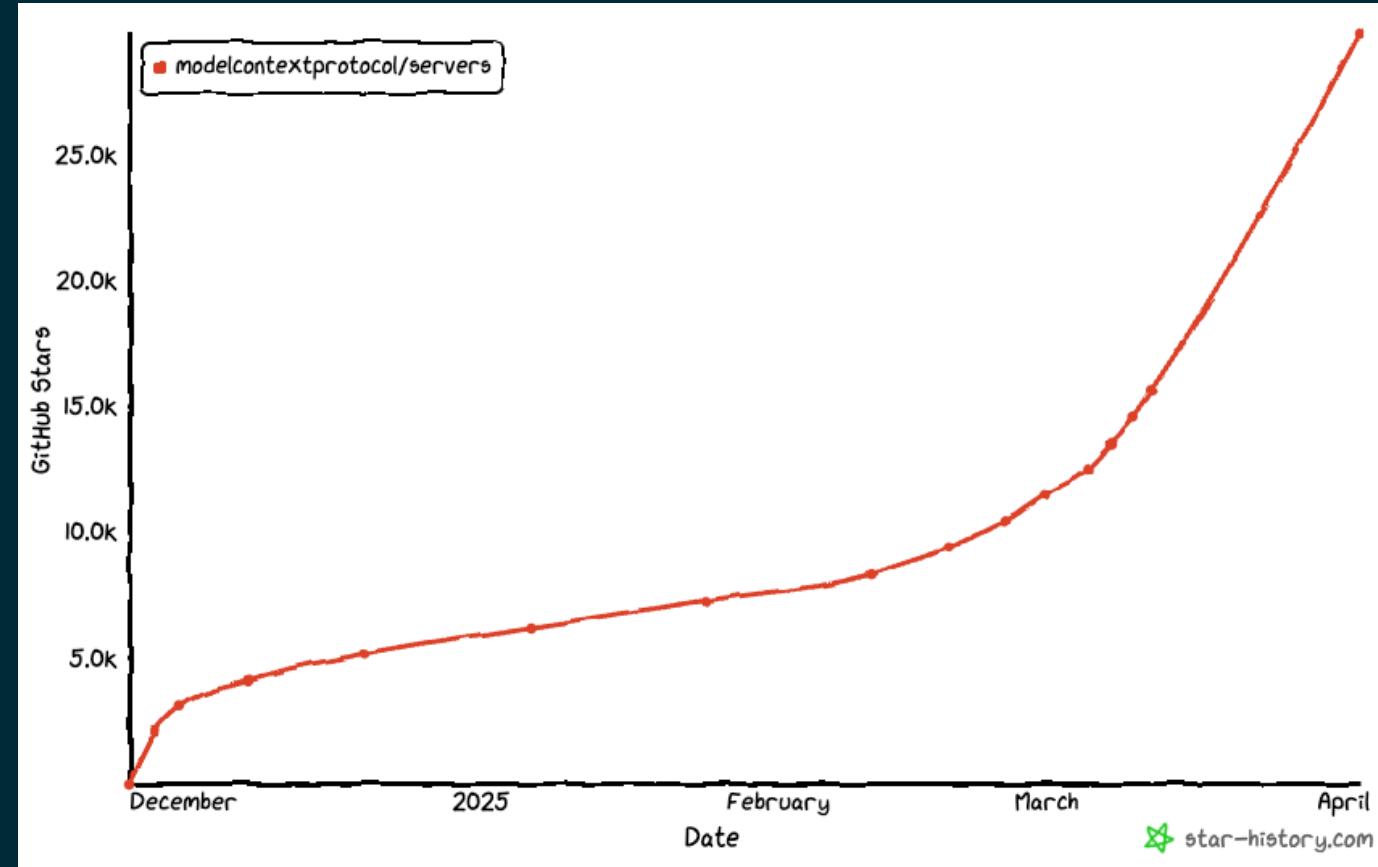


TOKEN USAGE

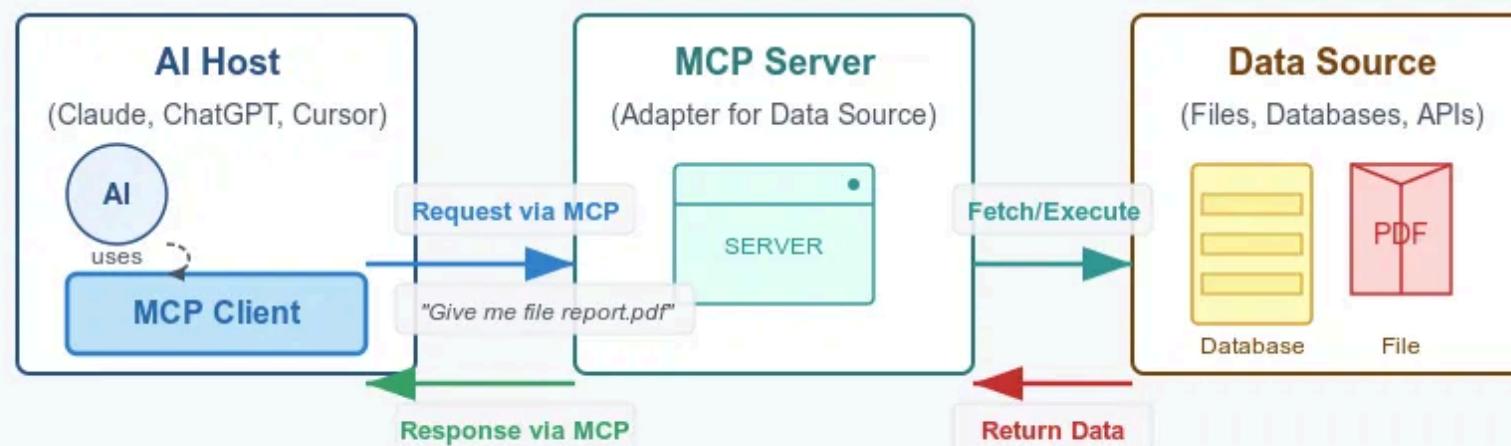
MCP: THE PROTOCOL THAT CONNECTS WORLDS

*"MCP is an open protocol
that enables seamless
integration between LLM
applications and external
data sources and tools."*

— Anthropic



Model Context Protocol (MCP) Architecture



Model Context Protocol (MCP) Flow

The MCP Client translates AI requests into the standardized protocol format, communicates with MCP Servers, which then interact with external Data Sources.

THE AGENTS

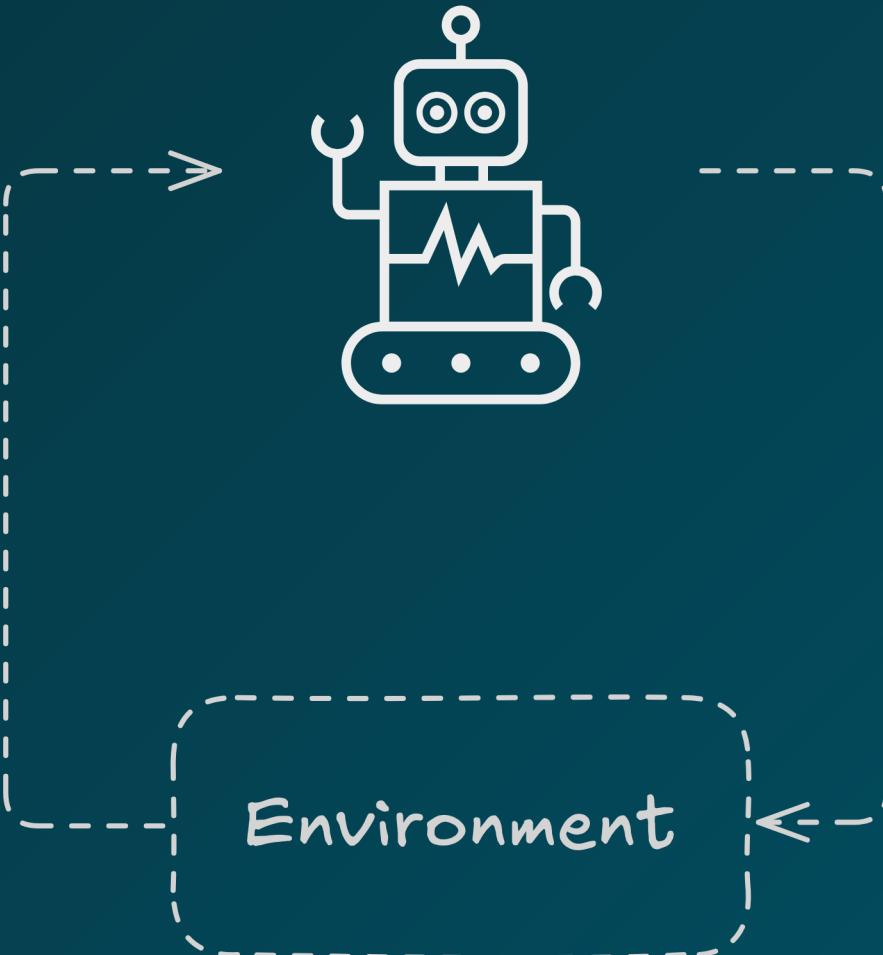
The Developers Guide to GenAI • Bauke Brenninkmeijer

AGENT

Autonomously
make decisions and
take actions to
achieve a goal



commencing interview



KEY COMPONENTS OF AGENTS

REFLECTION

PLANNING

TOOLS

COLLABORATION

KEY COMPONENTS OF AGENTS

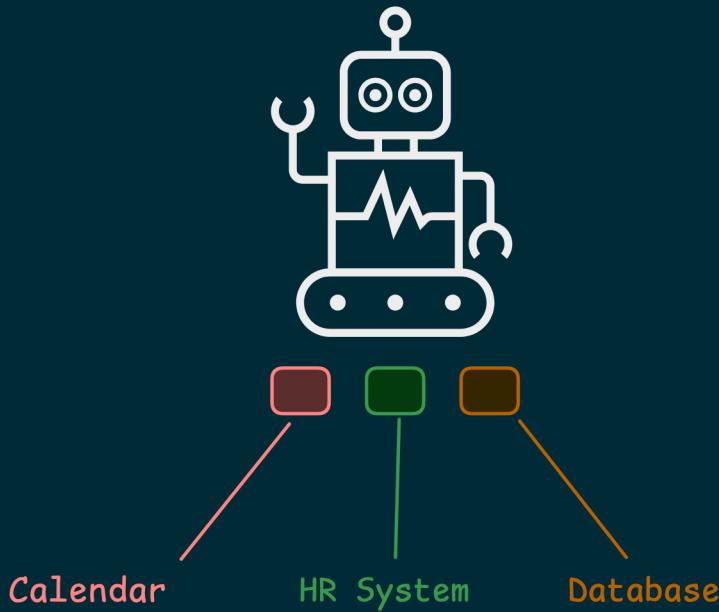
REFLECTION

PLANNING

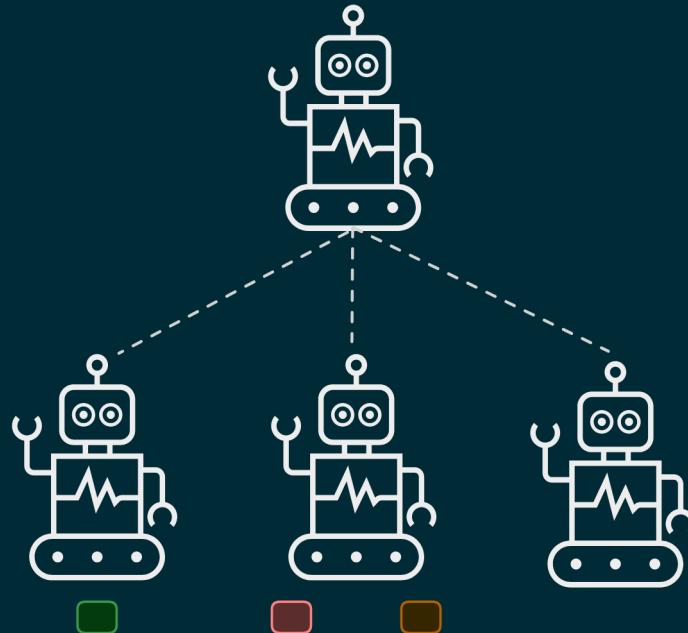
TOOLS

COLLABORATION

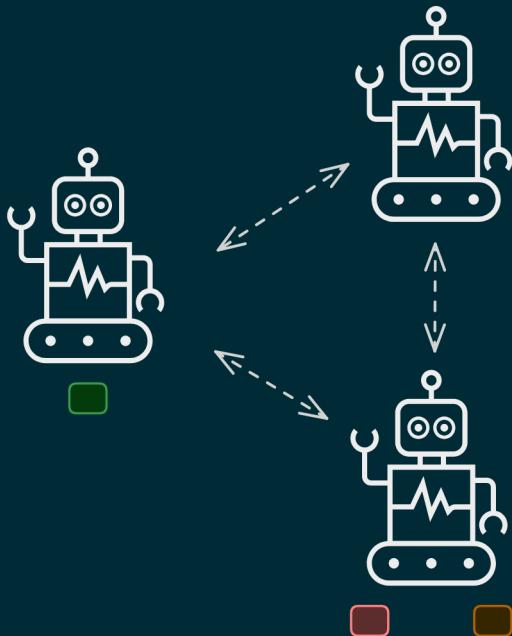
TYPES OF AGENTS: SINGLE AGENT SYSTEM



TYPES OF AGENTS: HIERARCHICAL MULTI-AGENT SYSTEM



TYPES OF AGENTS: NETWORK MULTI-AGENT SYSTEM



AI AGENTS FRAMEWORKS



LANGGRAPH



CREWAI



AUTOGEN



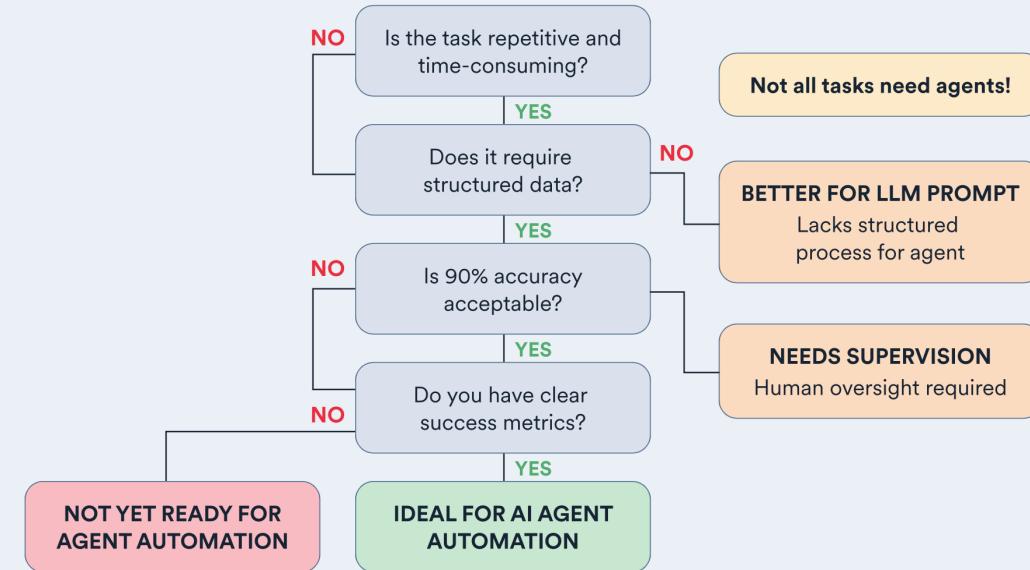
AGENTS SDK



AGENT DEVELOPMENT KIT

“Is This an Agent Job?” Decision Tree

A framework for determining which task to automate with AI agents



Source: AI Agents Unleashed Playbook 2025, Hubspot

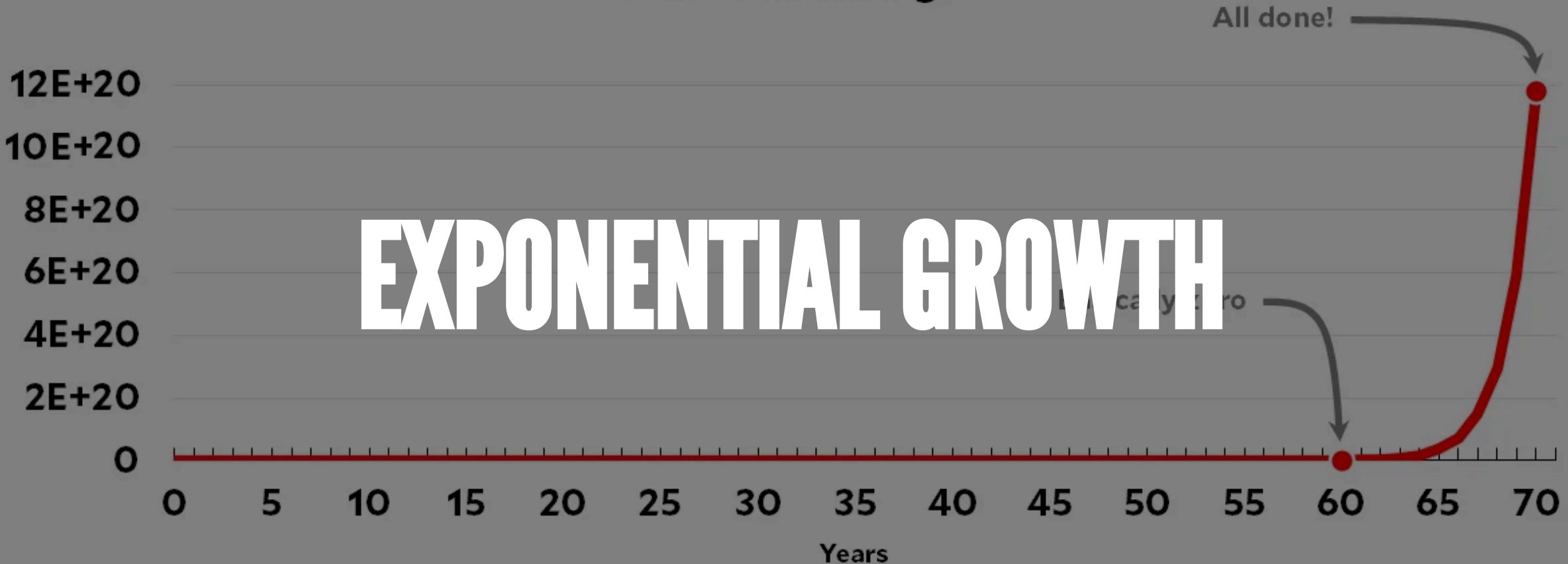


The Developers Guide to GenAI • Bauke Brenninkmeijer

BUILDING

Exponential Growth

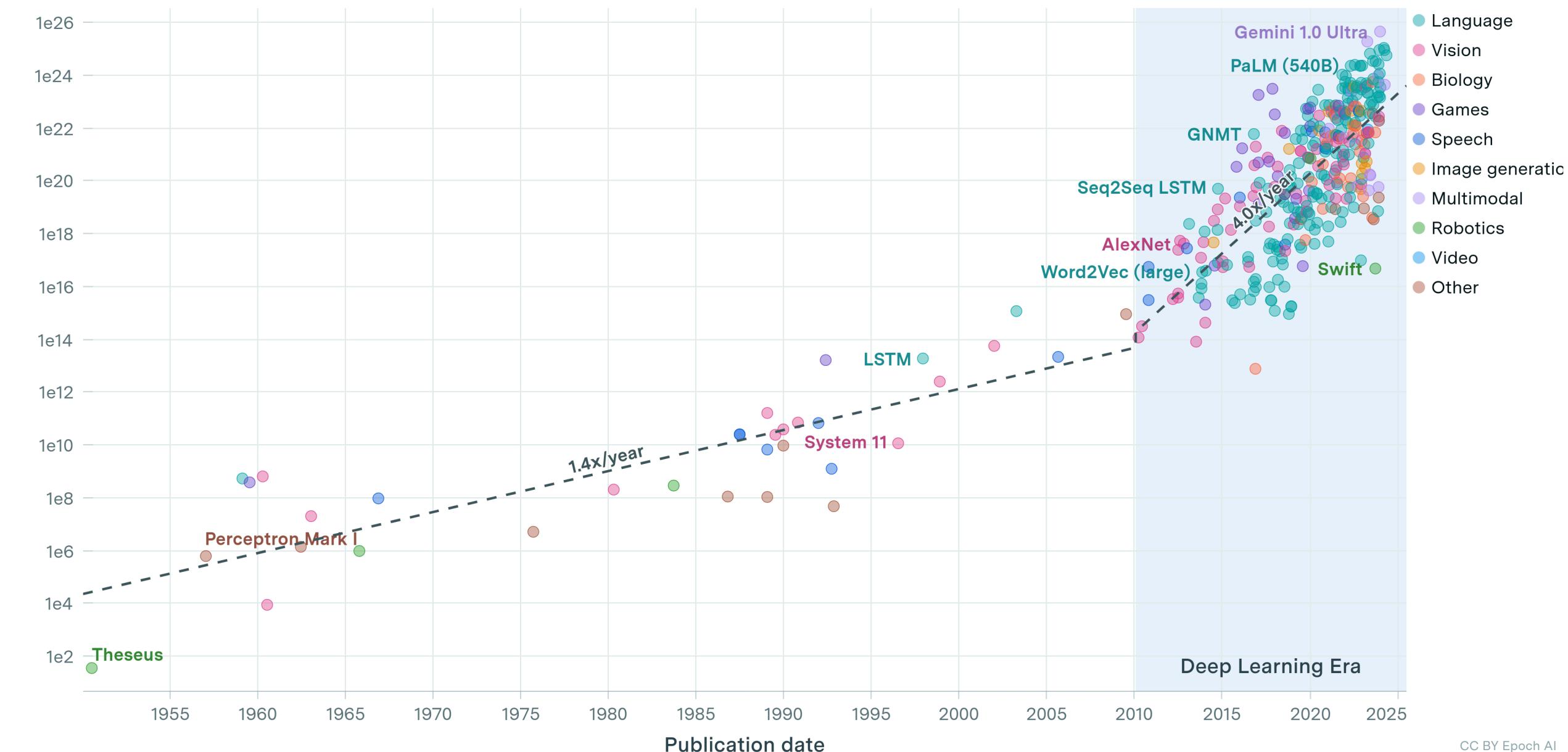
Isn't it amazing?

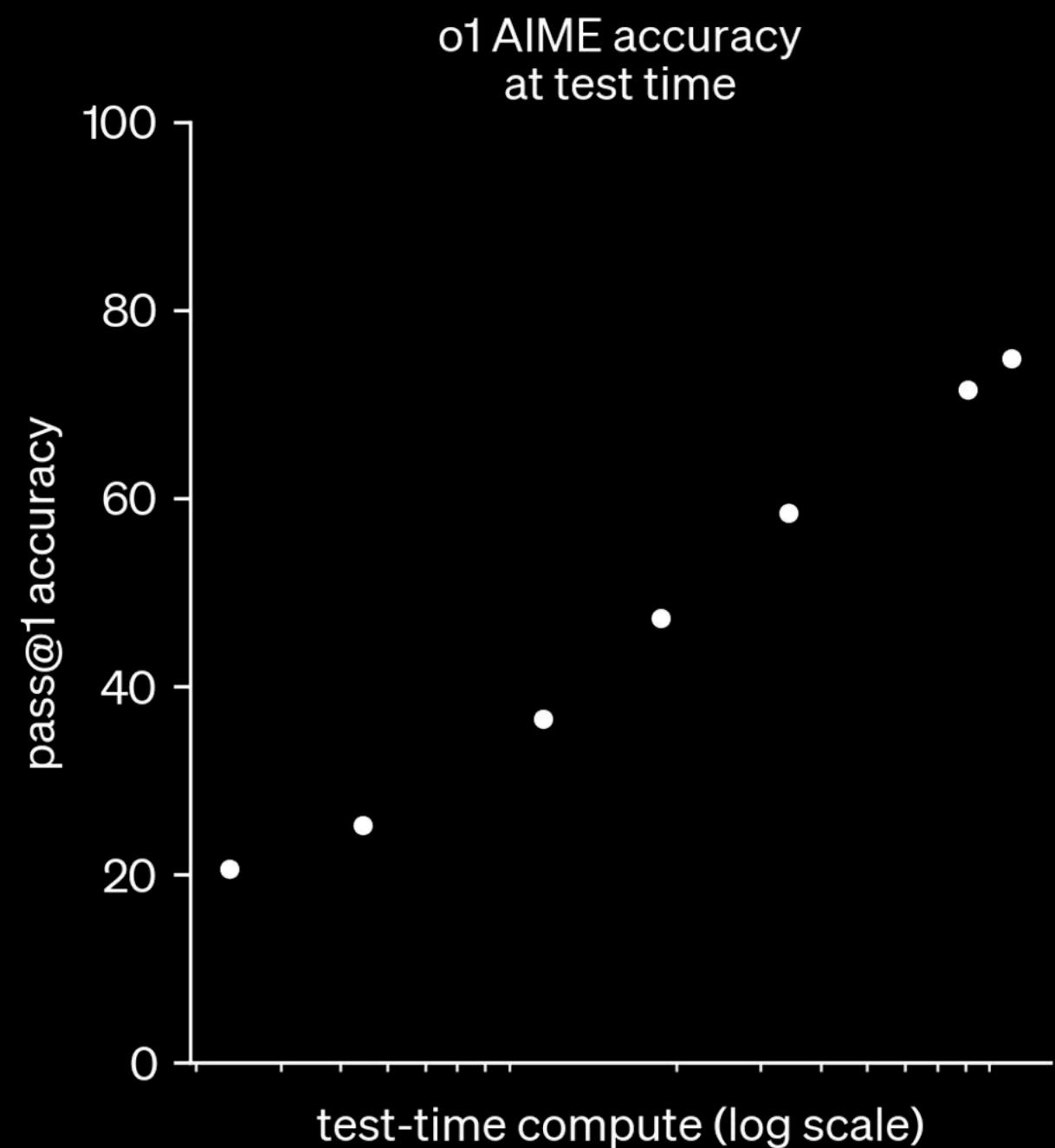
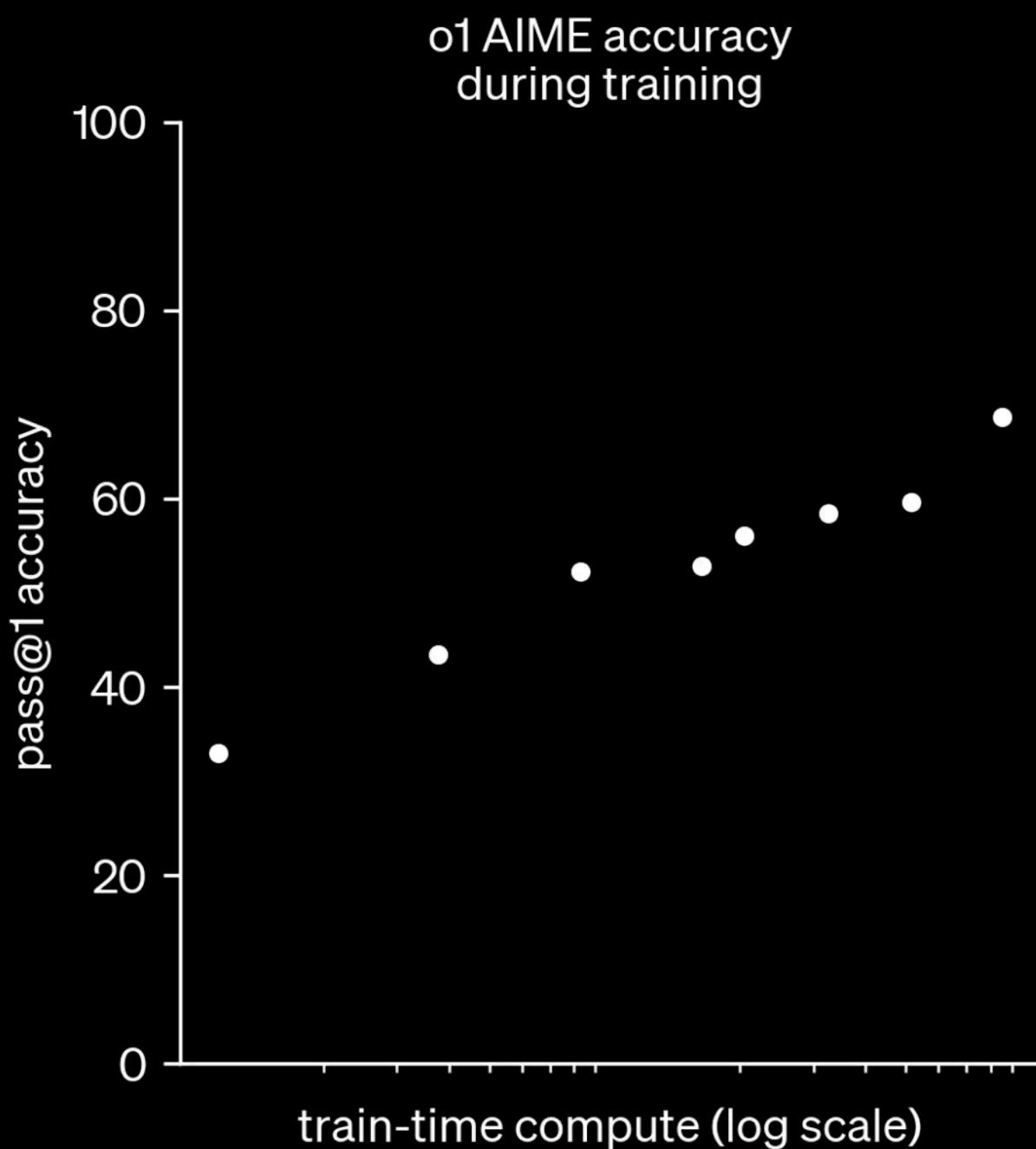


Source: Me

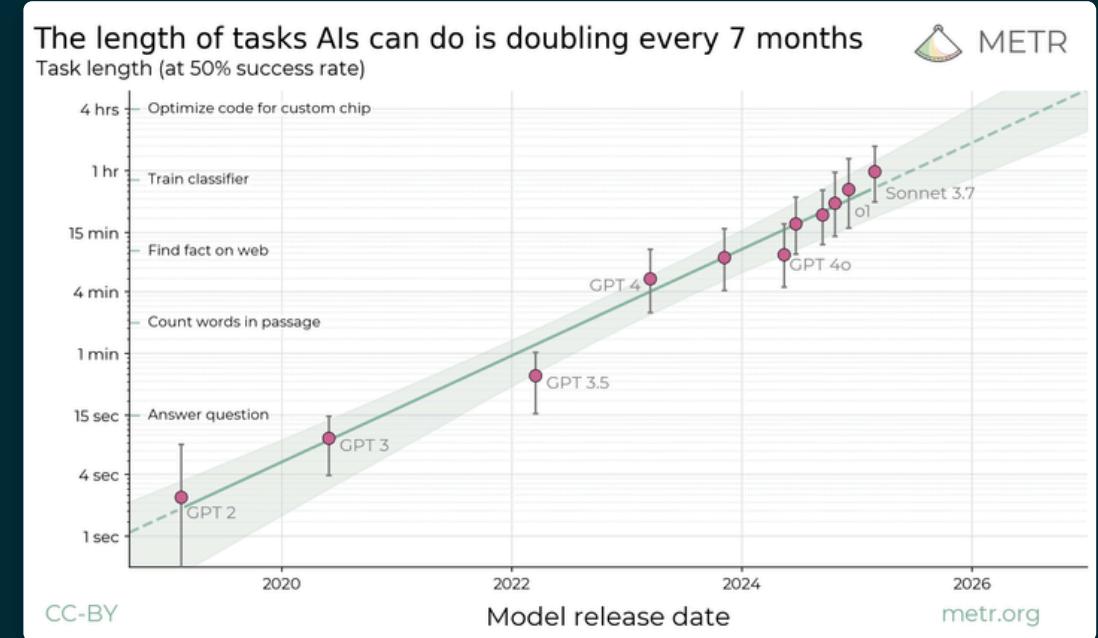
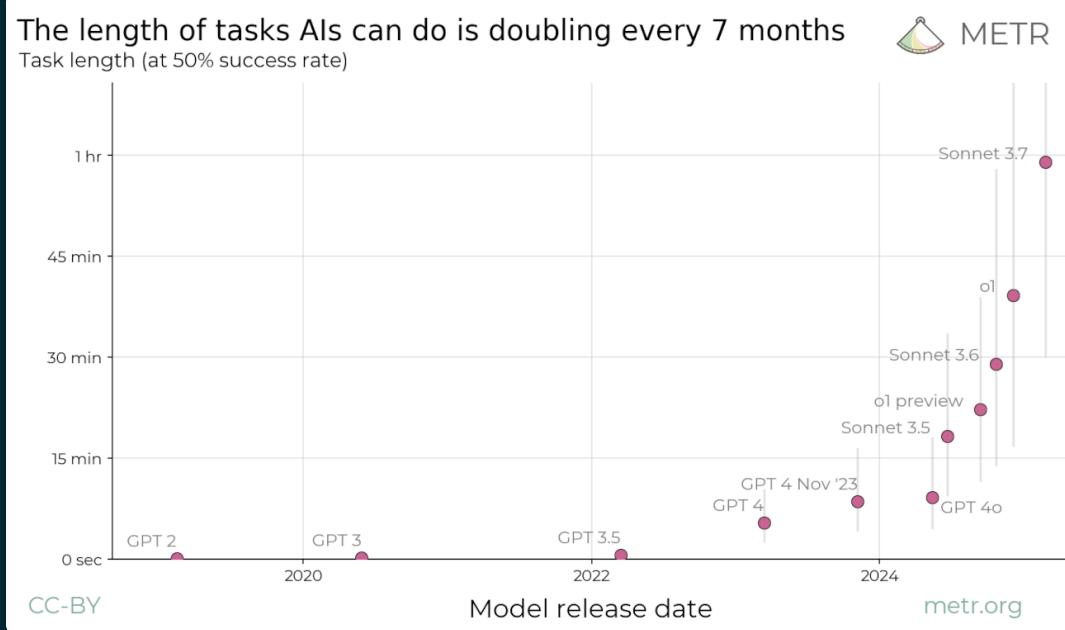


Training compute (FLOP)





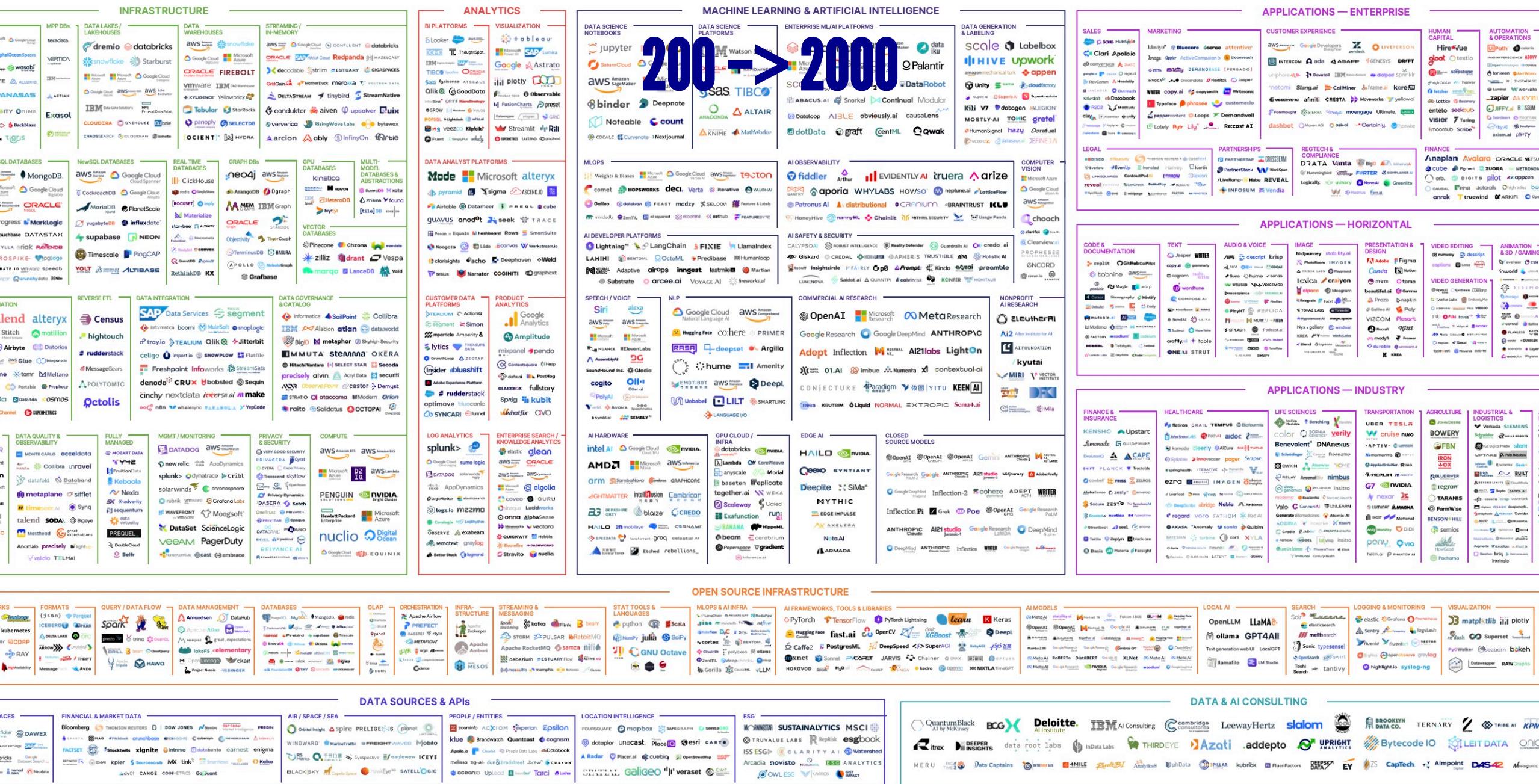
TASK LENGTH



"The more it reasons, the more unpredictable it becomes"

— Ilya Sutskever

THE 2024 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE





THANK YOU!

LET'S CONNECT

- Connect with me on LinkedIn
- Share your GenAI challenges
- Discuss implementation strategies

QUESTIONS?



Scan to connect



The Developers Guide to GenAI • Bauke Brenninkmeijer