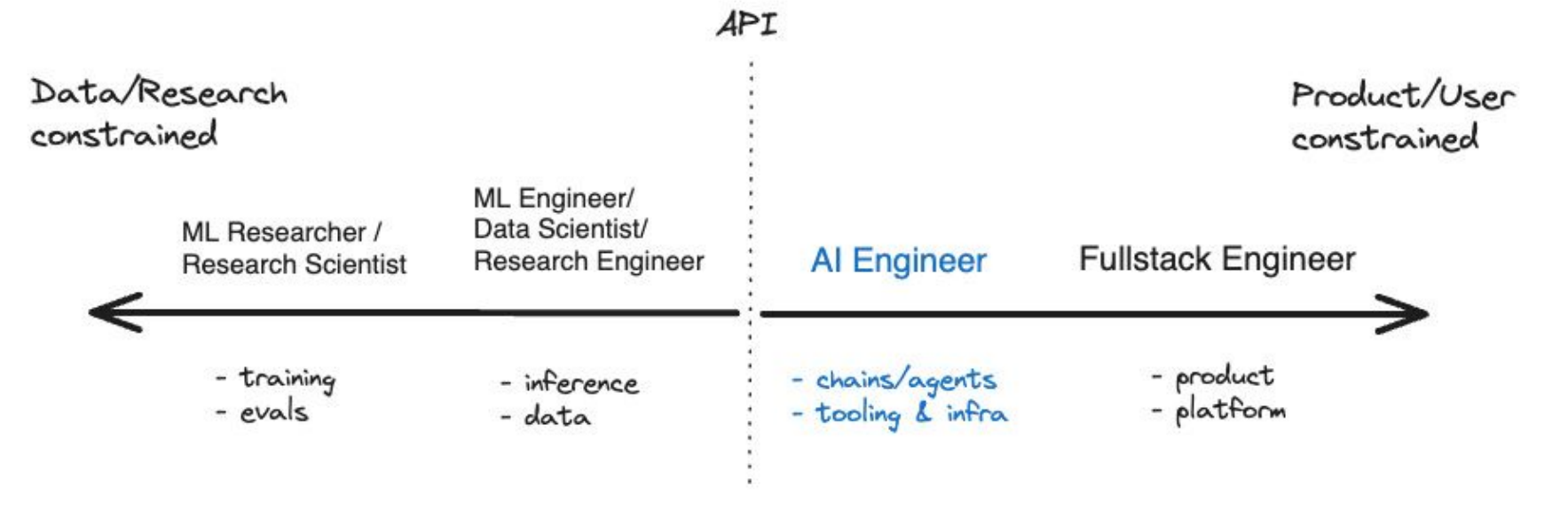
The Rise of the AI Engineer

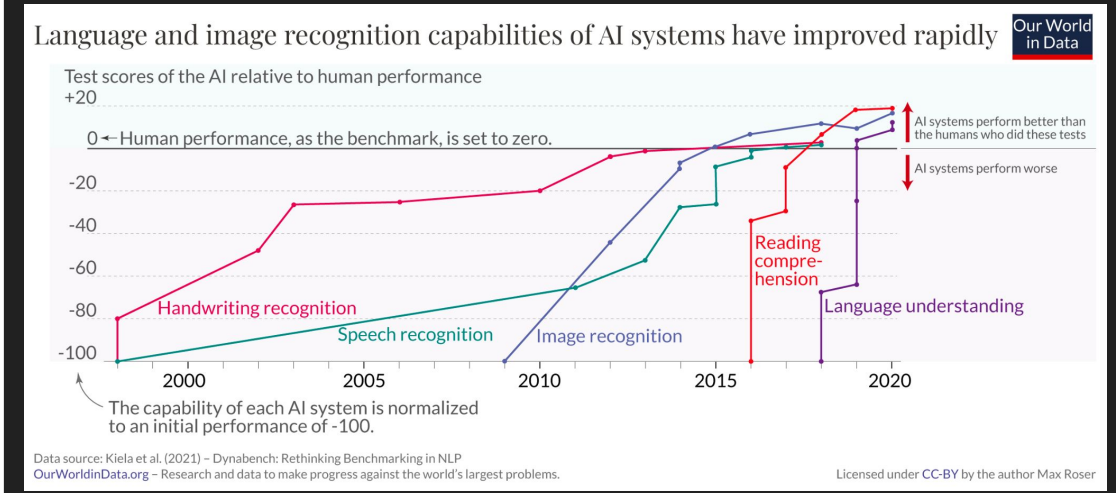
swyx.io/ai-landscape

The new role created



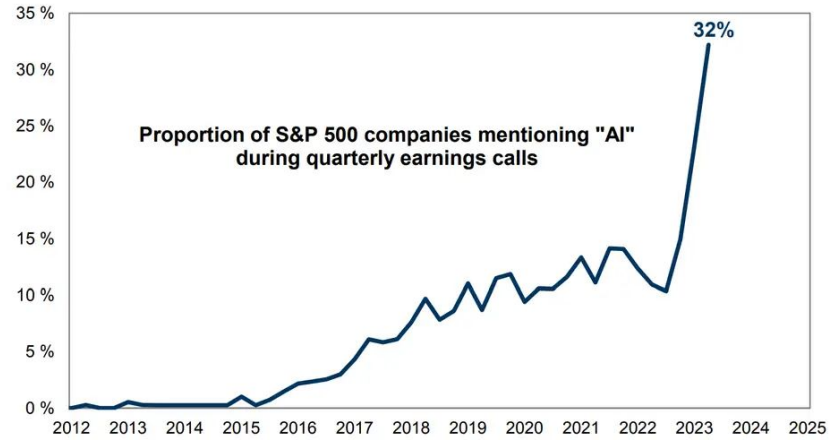
Part I: DATA

The Moore's law of our time



AI Madness

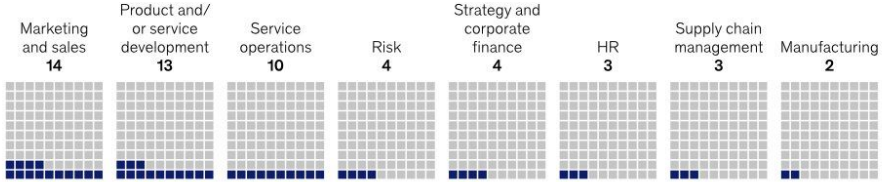
Exhibit 2: Mentions of "AI" on earnings calls have risen sharply this quarter
as of 2Q 2023



Source: Goldman Sachs Global Investment Research

The most commonly reported uses of generative AI tools are in marketing and sales, product and service development, and service operations.

Share of respondents reporting that their organization is regularly using generative AI in given function, %[1]



Most regularly reported generative AI use cases within function, % of respondents

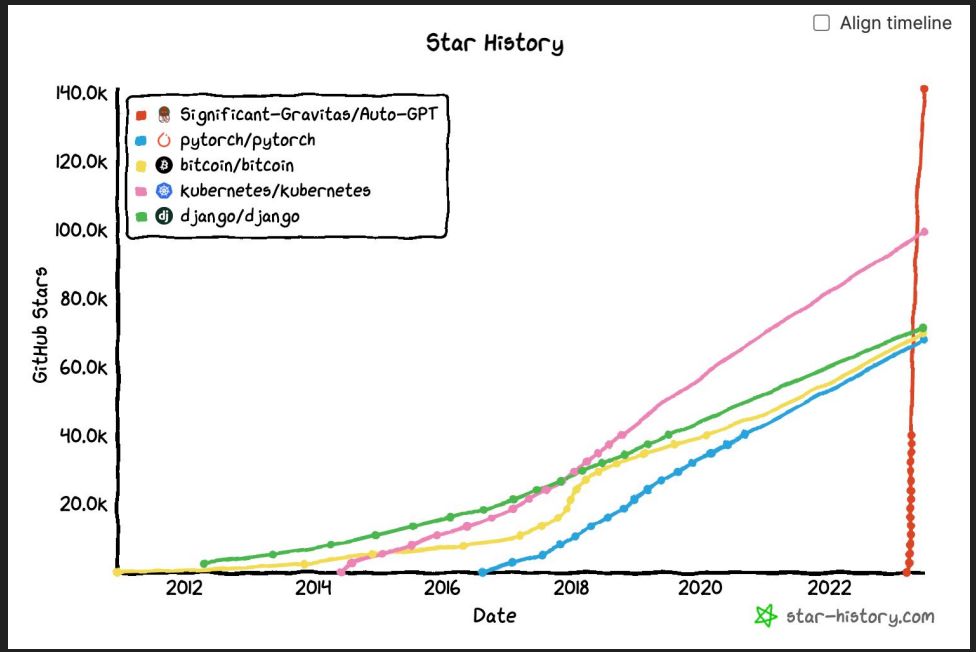
Marketing and sales	Product and/or service development	Service operations
Crafting first drafts of text documents	Identifying trends in customer needs	Use of chatbots (eg, for customer service)
9	7	6
Personalized marketing	Drafting technical documents	Forecasting service trends or anomalies
8	5	5
Summarizing text documents	Creating new product designs	Creating first drafts of documents
8	4	5

[1]Questions were asked of respondents who said their organizations have adopted AI in at least 1 business function. The data shown were rebased to represent all respondents.
Source: McKinsey Global Survey on AI, 1,684 participants at all levels of the organization, April 11–21, 2023

McKinsey & Company

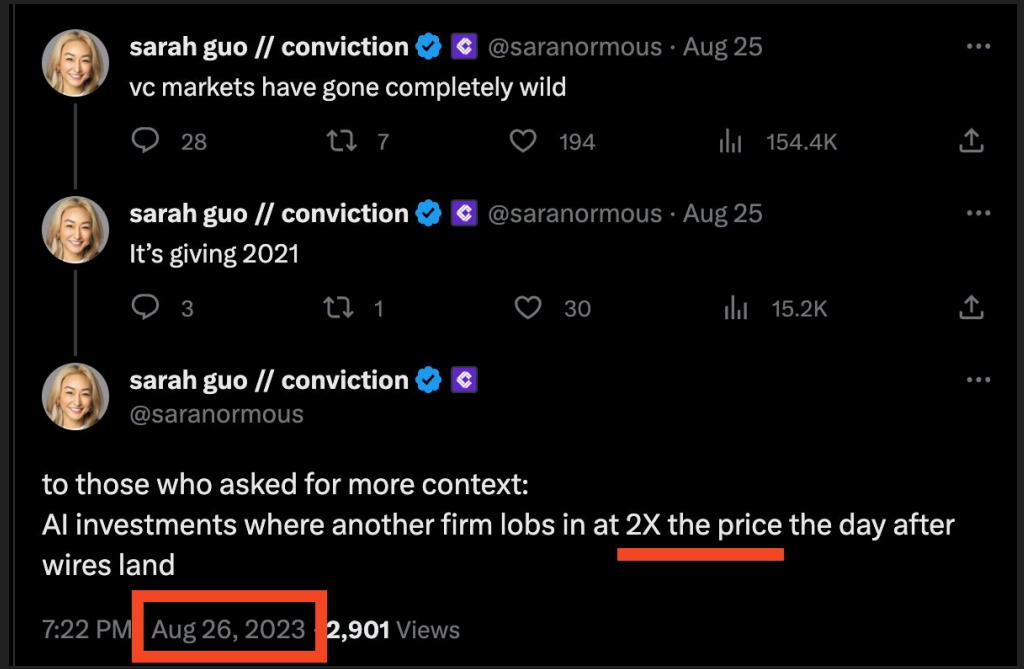
AI AI I/O

AI Madness

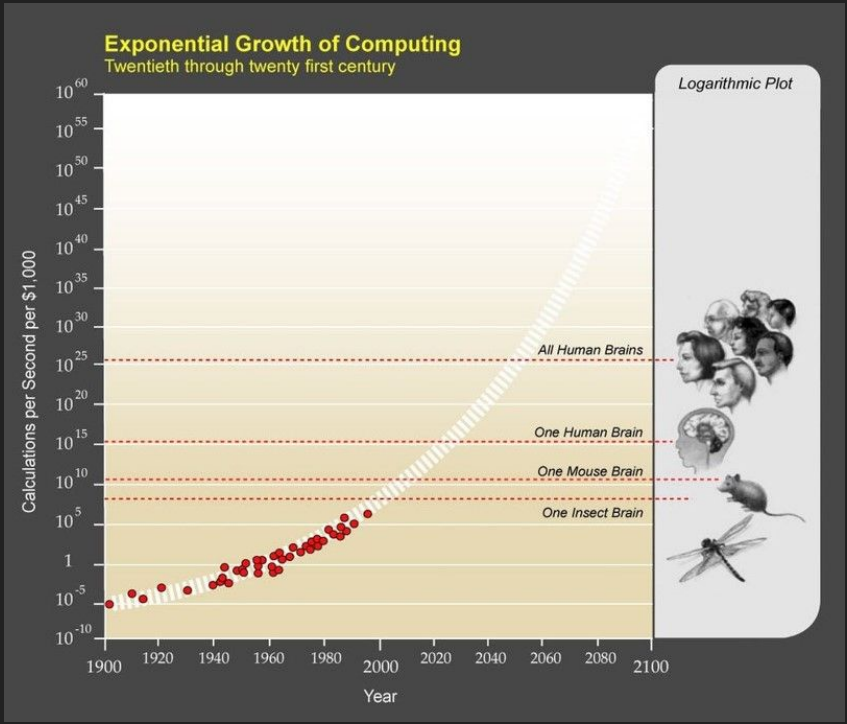

Star History
Align timeline
Significant-Gravitas/Auto-GPT
pytorch/pytorch
bitcoin/bitcoin
kubernetes/kubernetes
django/django
140.0k
120.0k
100.0k
80.0k
60.0k
40.0k
20.0k
GitHub Stars
2012
2014
2016
2018
2020
2022
Date
star-history.com

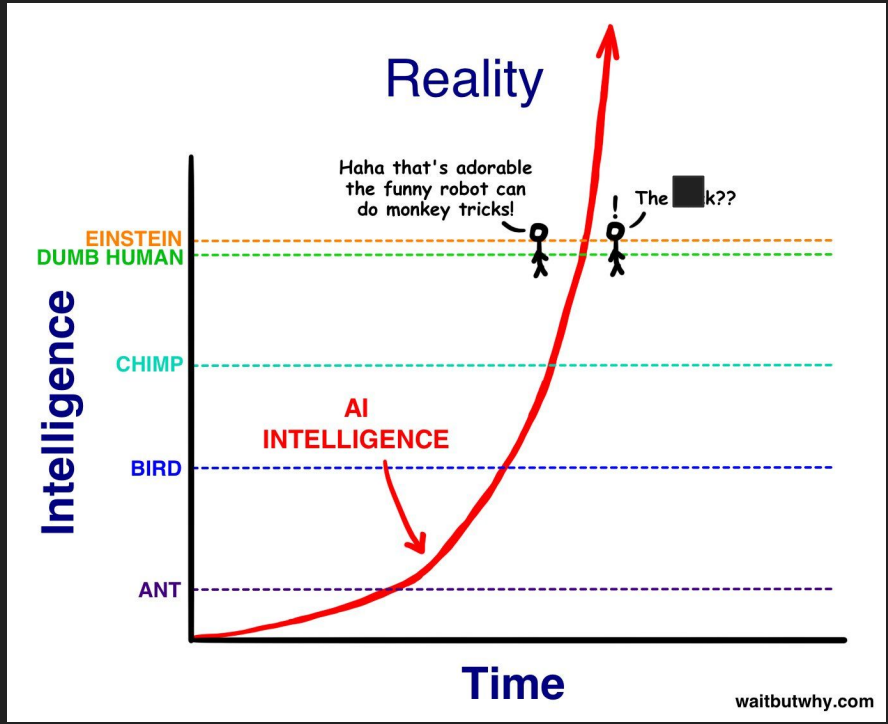

Yam Peleg
@Yampeleg
I refuse to believe a single person is using auto-gpt.
0. exactly 0.
5:43 PM · Jun 22, 2023 · 363.6K Views
23 Retweets
4 Quotes
762 Likes
152 Bookmarks

AI Manic Depression

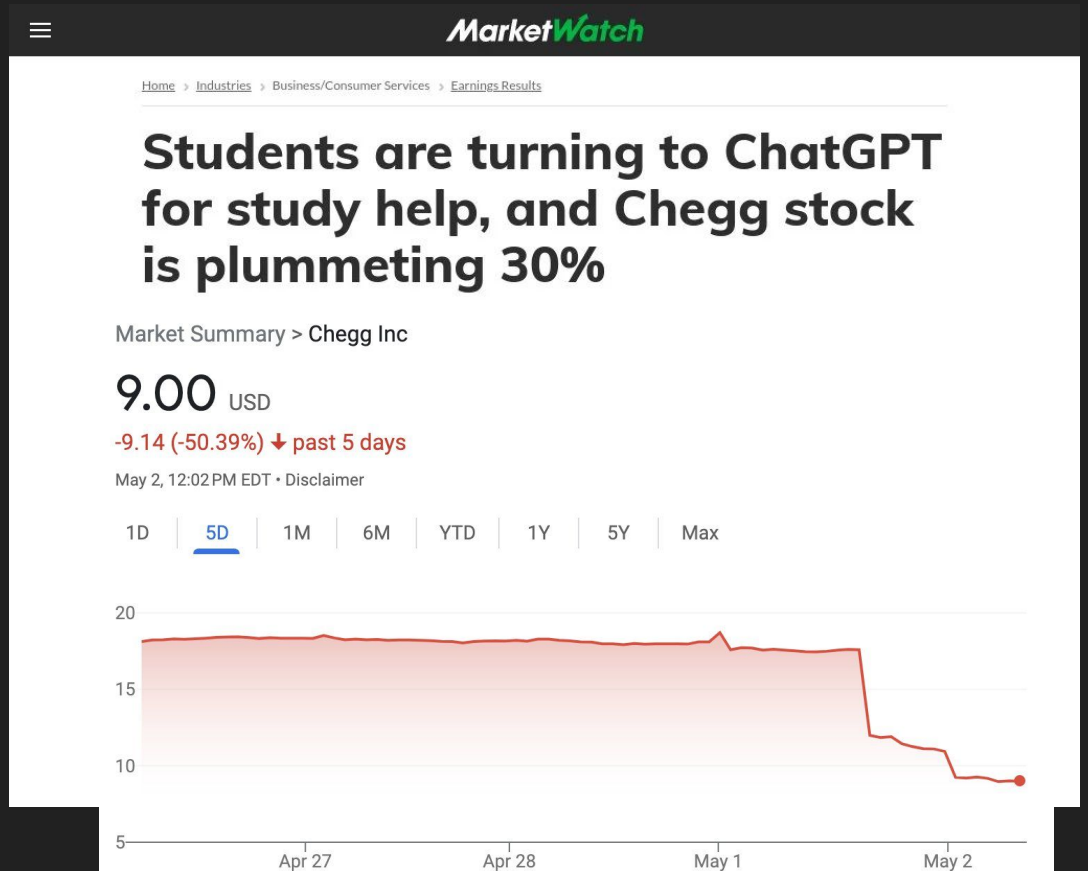


AI Fears


Exponential Growth of Computing
Twentieth through twenty first century
Logarithmic Plot
Calculations per Second per $1,000
All Human Brains
One Human Brain
One Mouse Brain
One Insect Brain
1900
1920
1940
1960
1980
2000
2020
2040
2060
2080
2100
Year


Reality
Haha that's adorable the funny robot can do monkey tricks!
The ■k??
EINSTEIN
DUMB HUMAN
CHIMP
AI INTELLIGENCE
BIRD
ANT
Intelligence
Time
waitbutwhy.com

AI Blame


MarketWatch
Home › Industries › Business/Consumer Services › Earnings Results
Students are turning to ChatGPT for study help, and Chegg stock is plummeting 30%
Market Summary > Chegg Inc
9.00 USD
-9.14 (-50.39%) ↓ past 5 days
May 2, 12:02 PM EDT • Disclaimer
1D 5D 1M 6M YTD 1Y 5Y Max
20
15
10
5
Apr 27
Apr 28
May 1
May 2


Market Summary > Chegg Inc
9.82 USD
Follow
-21.70 (-68.85%) ↓ past 5 years
Closed: Sep 15, 16:29 EDT • Disclaimer
After hours 9.82 0.00 (0.00%)
1D 5D 1M 6M YTD 1Y 5Y Max
150
100
50
0
2020
2021
2022
2023
Open 10.02
High 10.02
Low 9.80
Mkt cap 1.13B
P/E ratio 5.11
Div yield -
CDP score D
52-wk high 30.05
52-wk low 8.55

It’s Impossible to Keep Up

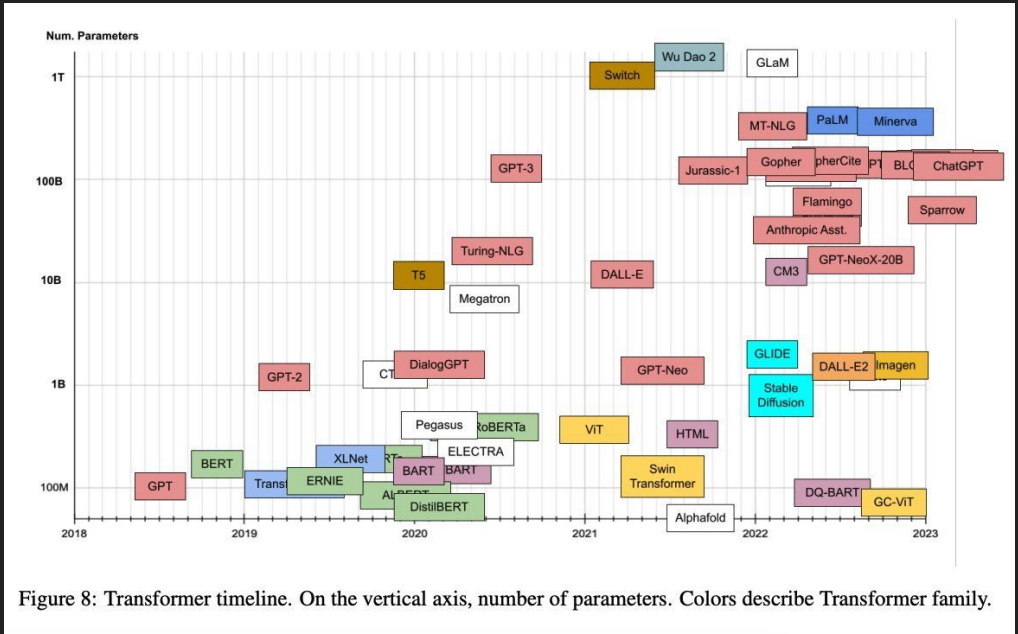


Figure 8: Transformer timeline. On the vertical axis, number of parameters. Colors describe Transformer family.

Latent Space

Fundraises and other Big Milestones

- HuggingFace $225m series D at 4b valuation
- WandB $50m series C+ at 1.25b valuation (tweet, led by Nat and Dan)
  - launched W&B Prompts
  - interview with Lukas
- Anthropic $100m round with SK Telecom with some unusual aggreement to develop a multilingual LLM for telcos (?)
- Modular raised $100m at $600m valuation (The Information, tweet thread)
  - upcoming guest!
- Ideogram launched with $16.5m seed from a16z/Index (Jim Fan, Ethan Mollick)
  - former Google Imagen team actually shipping! kinda good text-in-image!

AI Eng Tooling

- Langchain released an “expression language” - upcoming guest!
  - Harrison noted that Benchmarking Q&A over CSV (tweet, blog, video walkthru) was a big post for them

Who’s doing a good job?

- **Generative Text**

AI content platform Jasper raises $125M at a $1.5B valuation

- **Generative Images**

$200 million

According to media reports, with a mere 40-member team, the company is slated to rake in $200 million in revenue this year from its 15 million community members. 6 days ago

- **Coding Assistants**

inside.com
https://inside.com › tech › posts › microsoft-s-github-c...

3 Microsoft's GitHub Copilot now has over a million users

Jan 26, 2023 — More: Within a month, the AI pair programmer attracted 400,000 **subscribers** and now counts over 1 million users. **Copilot**, powered by OpenAI's ...

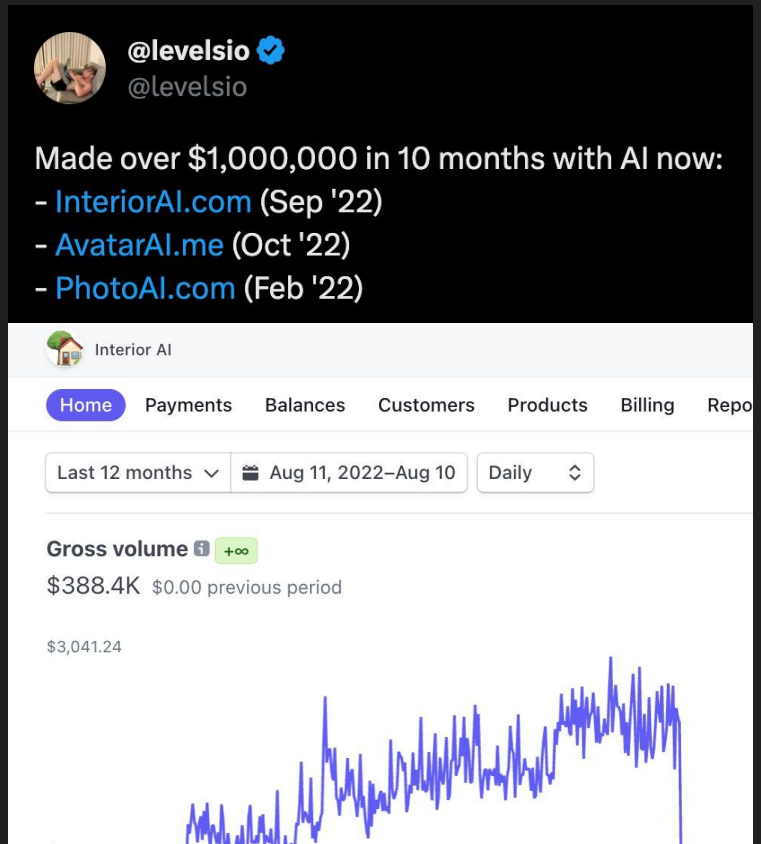
- **Chatbots**

08-30-23 | MOST INNOVATIVE COMPANIES

OpenAI reportedly nears $1 billion in annual sales

According to a new report, the maker of ChatGPT is growing even faster than the company projected as demand for AI technology increases.

Individual Developers too



Part I: DATA

PART II: PEOPLE

AI Engineer > AI Developer?

The Rise of the AI Engineer

Emergent capabilities are creating an emerging title: to wield them, we'll have to go beyond the Prompt Engineer and write *software*. Plus: Join 500 AI Engineers at our first summit, Oct 8-10 in SF!

SWYX

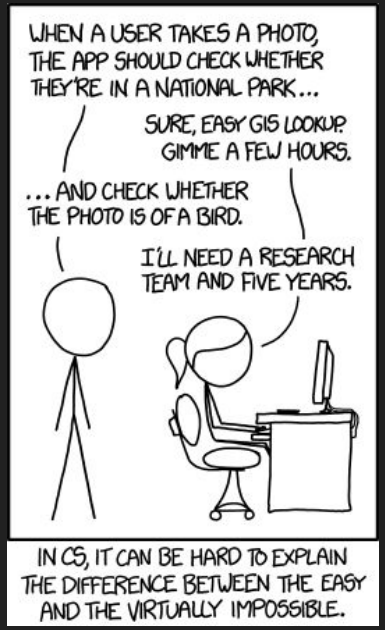
JUN 30, 2023

212 13 18 Share

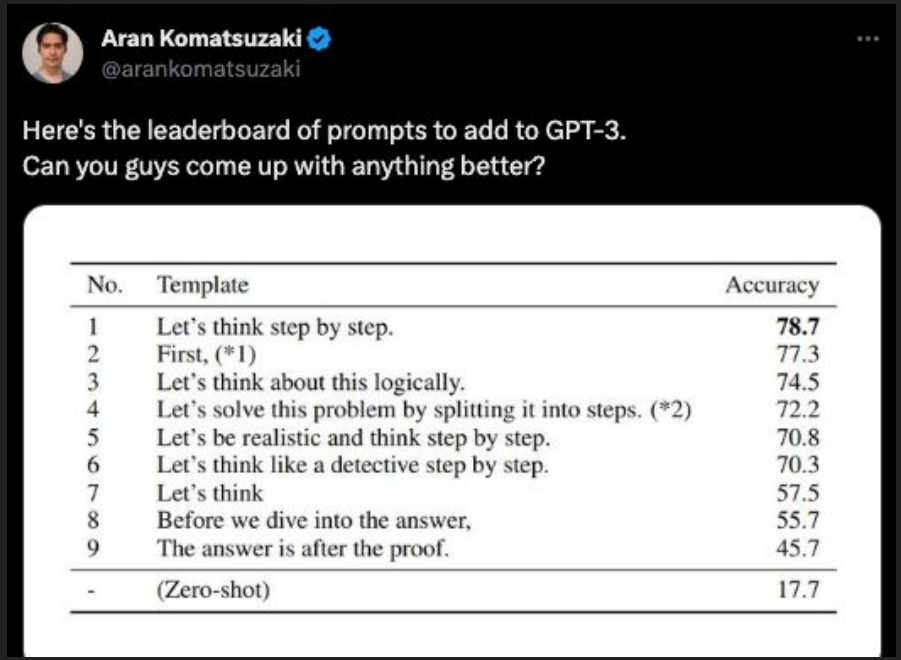
Thanks for the many comments and questions on HN and Twitter! We convened a snap Twitter Space to talk it through and >1,000 AI Engineers tuned in. Playback here!

Central Thesis - AI is “shifting right”

2013

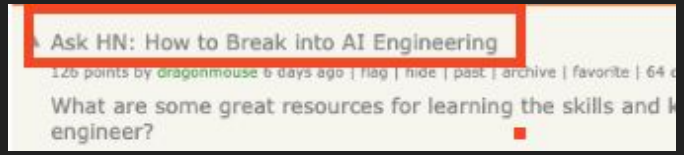


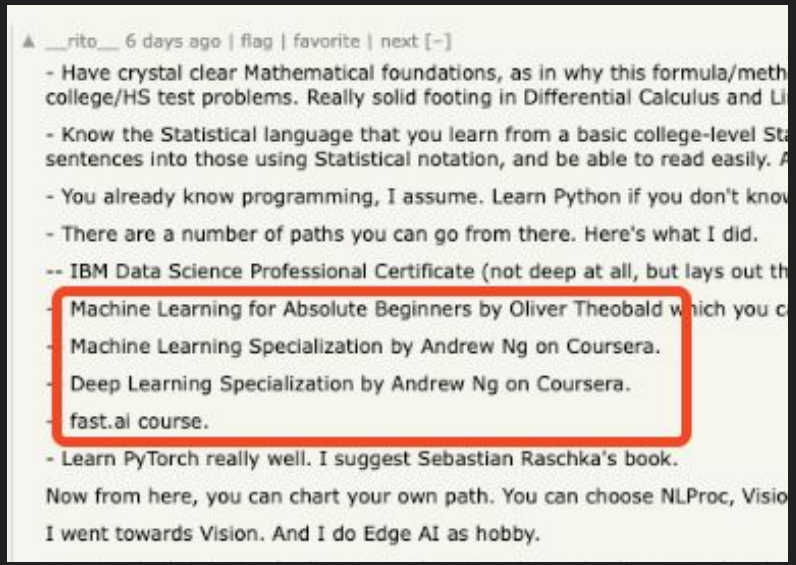
2023

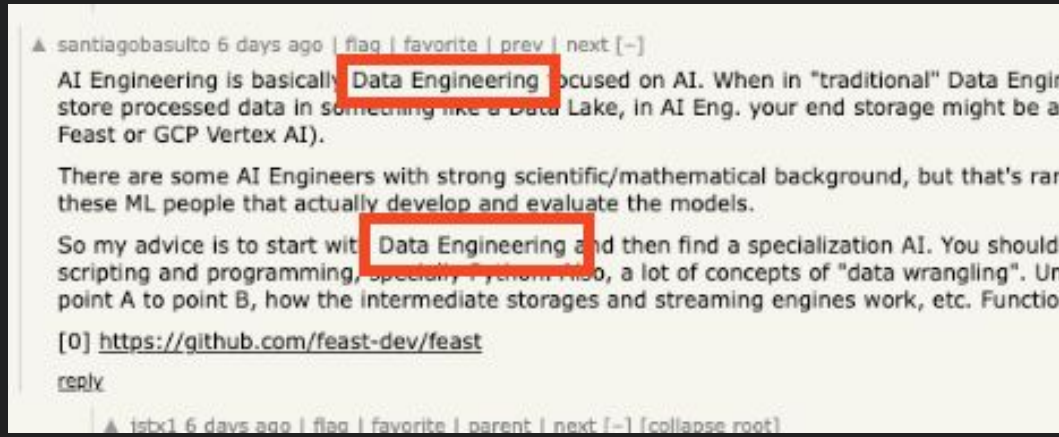


No.	Template	Accuracy
1	Let's think step by step.	**78.7**
2	First, (*1)	77.3
3	Let's think about this logically.	74.5
4	Let's solve this problem by splitting it into steps. (*2)	72.2
5	Let's be realistic and think step by step.	70.8
6	Let's think like a detective step by step.	70.3
7	Let's think	57.5
8	Before we dive into the answer,	55.7
9	The answer is after the proof.	45.7
-	(Zero-shot)	17.7

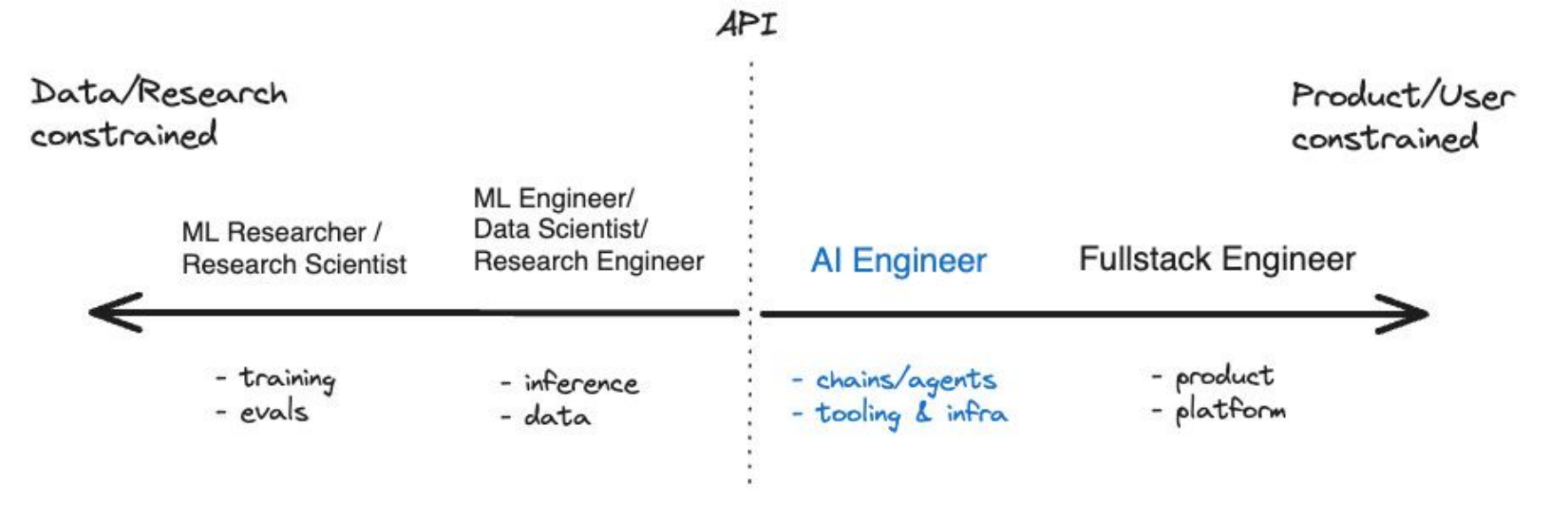
The Starting Point has moved







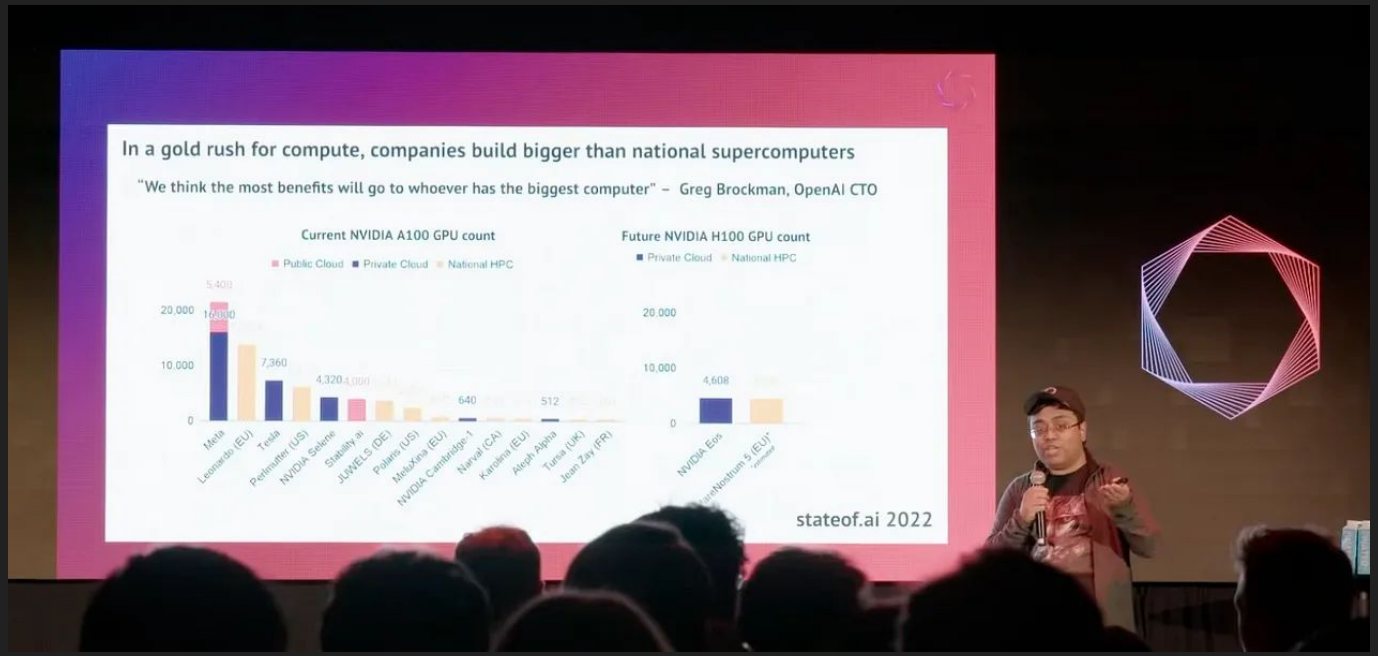
The new role created



5 reasons for the Rise of the AI Engineer

- Economics
- Sociology
- Tech
- Product
- Language

It’s not a tech issue - it’s economics



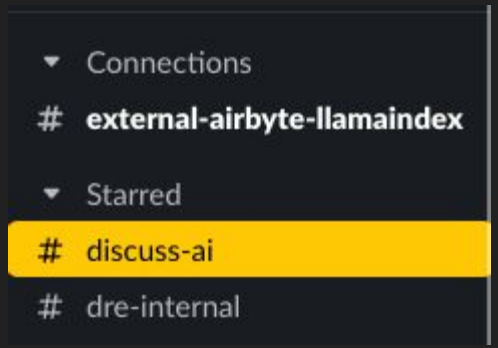
GPU Hoarding:

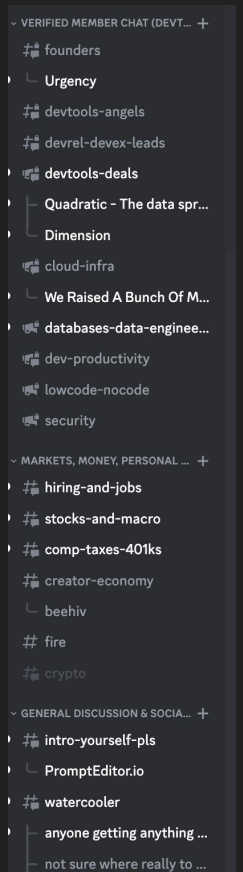
- Inflection ($1.3b)
- Mistral ($113m)
- Reka ($58m)
- Poolside ($26m)
- Contextual ($20m)

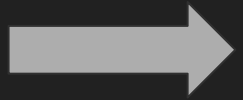
It’s not a tech issue - it’s economics

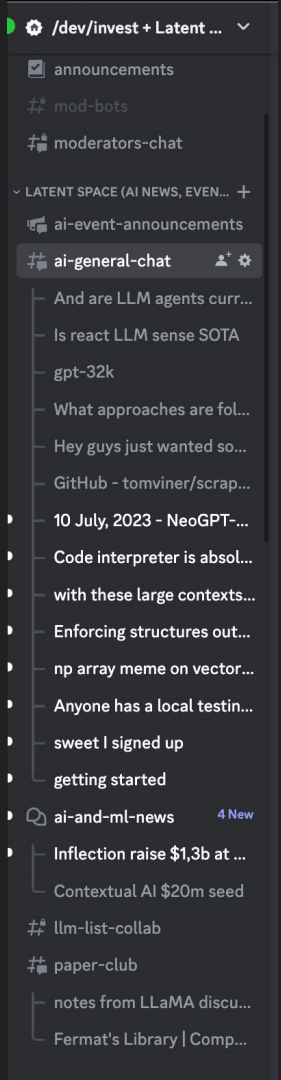
> “In numbers, there's probably going to be significantly more AI Engineers than there are ML engineers / LLM engineers. One can be quite successful in this role without ever training anything.” - Andrej Karpathy

It’s not a tech issue - it’s sociology

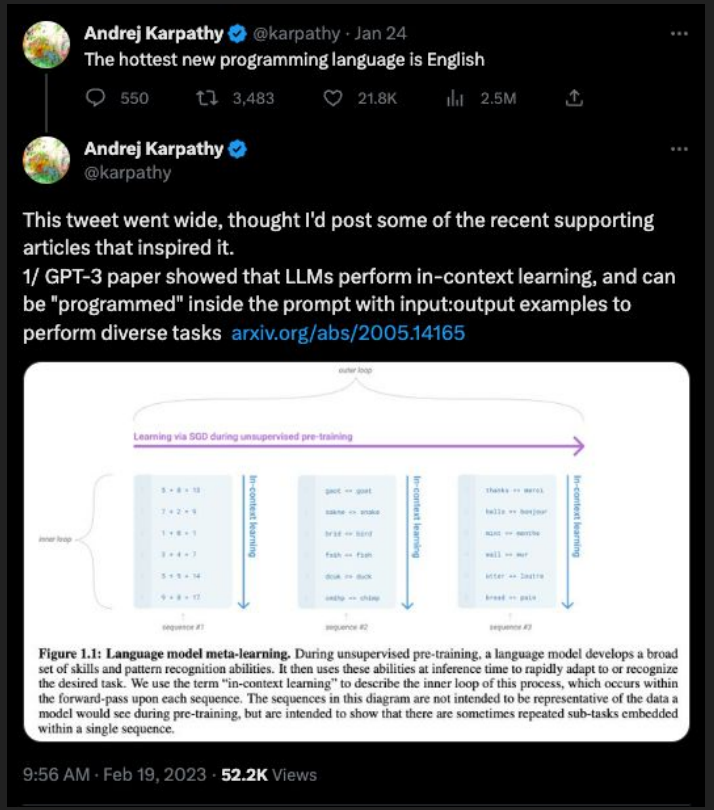


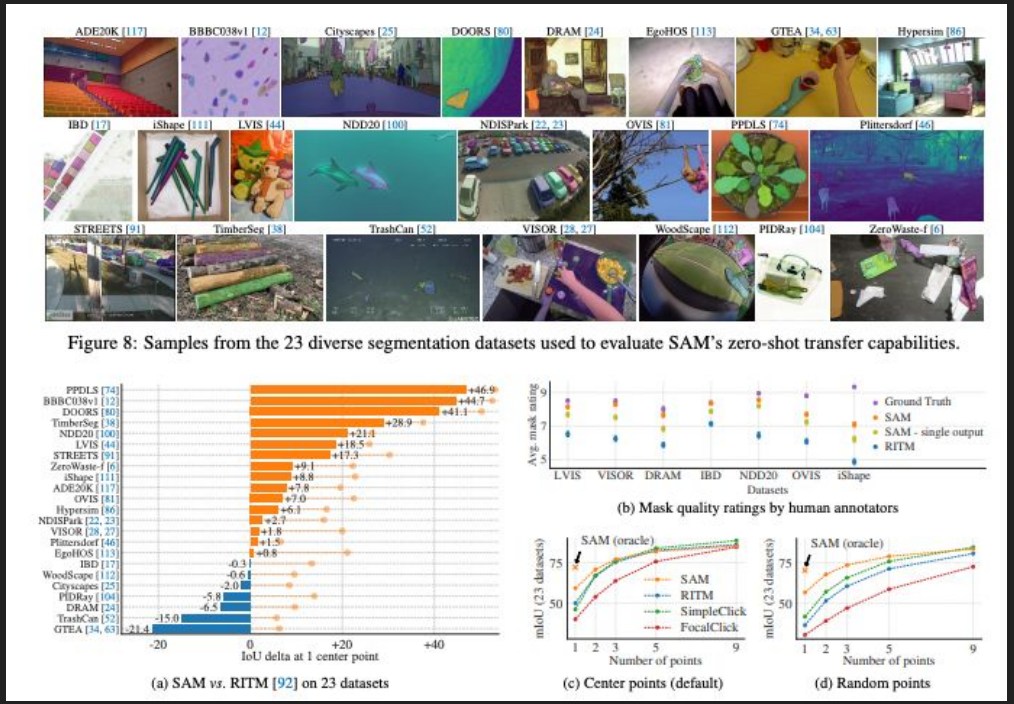




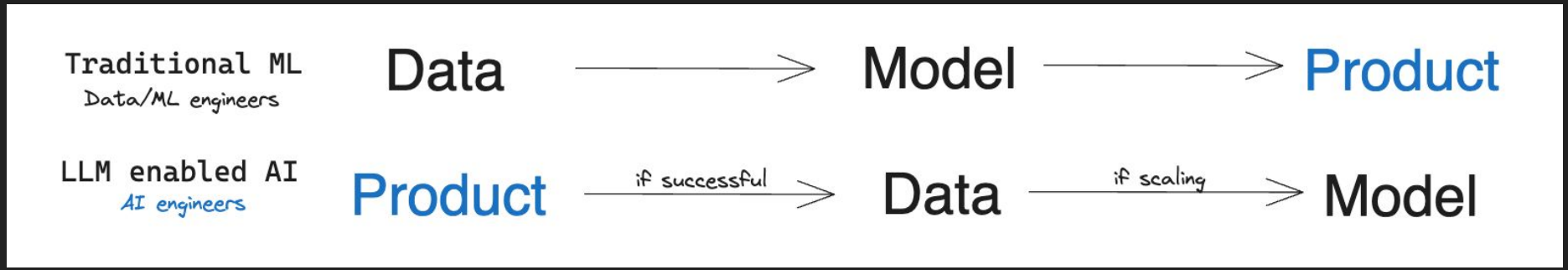


It’s also enabled by tech





Ready-Aim-Fire => Fire-Ready-Aim



Python -> JavaScript

Python → JavaScript. Data/AI is traditionally extremely Python centric, and the first AI Engineering tools like LangChain, LlamaIndex and Guardrails arose out of that same community. However, there are at least as many JavaScript developers as Python developers, so now tools are increasingly catering to this widely expanded audience, from LangChain.js and Transformers.js to Vercel's new AI SDK. The TAM expansion and opportunity is dramatic.

5 reasons for the Rise of the AI Engineer

- Economics: GPU and People
- Sociology: Intrinsic desire
- Tech: Zero Gradient / In Context Learning
- Product: Fire-Ready-Aim
- Language: Python -> JS

Software 3.0, aka *why AI **engineers**?*

Not prompt engineers?

Software 1.0 -> 2.0

1. AI = “A lot of Ifs”
2. AI = Learned Weights

Software 2.0

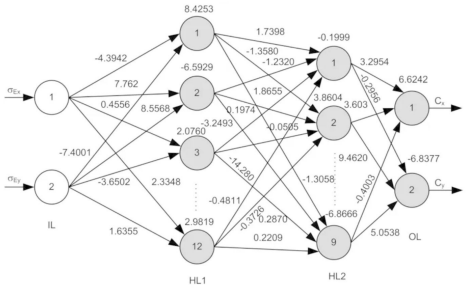
Andrej Karpathy · Follow

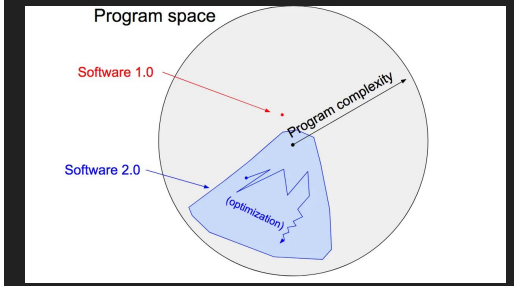
9 min read · Nov 11, 2017

50K 170

The “classical stack” of **Software 1.0** is what we’re all familiar with — it is written in languages such as Python, C++, etc. It consists of explicit instructions to the computer written by a programmer. By writing each line of code, the programmer identifies a specific point in program space with some desirable behavior.

In contrast, **Software 2.0** is written in much more abstract, human unfriendly language, such as the weights of a neural network. No human is involved in writing this code because there are a lot of weights (typical networks might have millions), and coding directly in weights is kind of hard (I tried).





Software 1.0 -> 2.0 -> 3.0

Chris Olah @ch402 · Jun 19, 2020

Software 1.0: Write an algorithm that has the right behavior.

@karpathy's Software 2.0: Optimize differentiable blob to have correct behavior. medium.com/@karpathy/soft...

Software 3.0: Figure out the right prompt to make your meta-learning language model have the right behavior? :P

Andrej Karpathy ✔
@karpathy

Love the idea for Software 3.0 😂. Programming moving from curating datasets to curating prompts to make the meta learner "get" the task it's supposed to be doing. LOL 🤣👌

3:14 AM · Jun 19, 2020

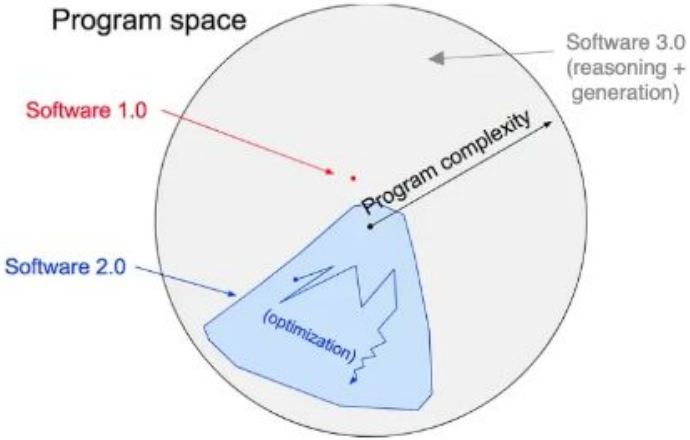
AI + Code >> AI + Language

Role of Codegen

1 + 2 = 3

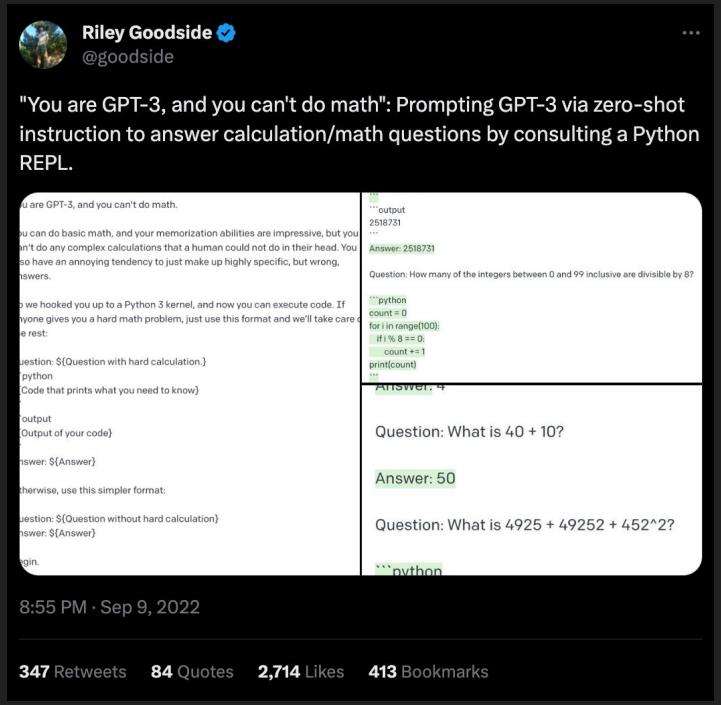
1+2=3: The Role of Code in the evolution from Software 2.0 to Software 3.0

6 years ago, Andrej Karpathy wrote a very influential essay describing Software 2.0 - contrasting the "classical stack" of hand-coded programming languages that precisely model logic against the new stack of "machine learned" neural networks that approximate logic, enabling software to solve a lot more problems than could humanly be modeled. He followed it up this year by noting that the hottest new programming language is English, finally filling out the gray area in his diagram that was left unlabeled in the original essay.



Brief History of Code in LLMs

Sept 2022



Riley Goodside
@goodside

"You are GPT-3, and you can't do math": Prompting GPT-3 via zero-shot instruction to answer calculation/math questions by consulting a Python REPL.

u are GPT-3, and you can't do math.

u can do basic math, and your memorization abilities are impressive, but you in't do any complex calculations that a human could not do in their head. You so have an annoying tendency to just make up highly specific, but wrong, nswers.

o we hooked you up to a Python 3 kernel, and now you can execute code. If yone gives you a hard math problem, just use this format and we'll take care e rest:

uestion: ${Question with hard calculation.}
python
{Code that prints what you need to know}

output
{Output of your code}

nswer: ${Answer}

therwise, use this simpler format:

uestion: ${Question without hard calculation}
nswer: ${Answer}

gin.

```
output
2518731
---
Answer: 2518731

Question: How many of the integers between 0 and 99 inclusive are divisible by 8?

```python
count = 0
for i in range(100):
    if i % 8 == 0:
        count += 1
print(count)
```

Answer: 4

Question: What is 40 + 10?

Answer: 50

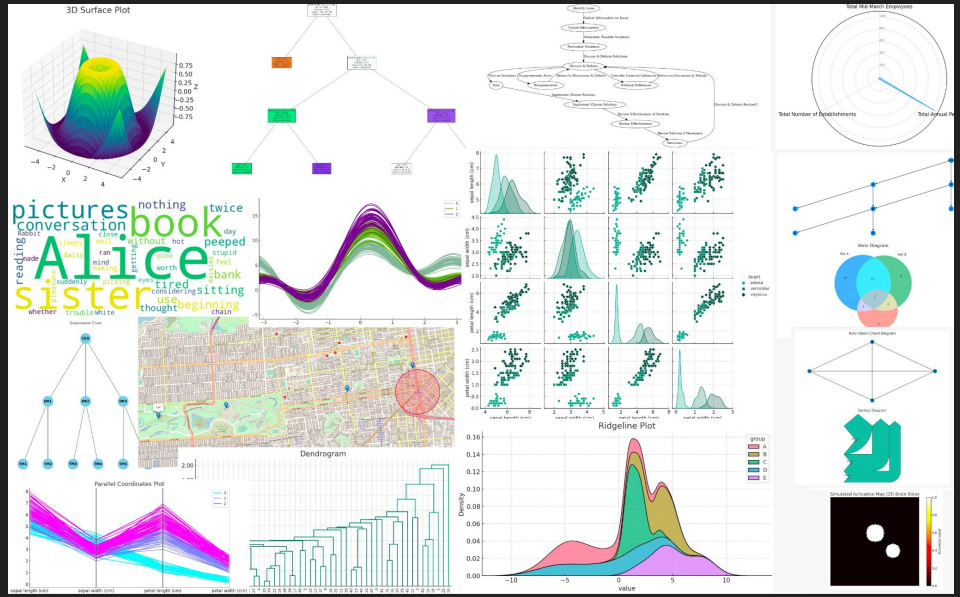
Question: What is 4925 + 49252 + 452^2?

```python
```

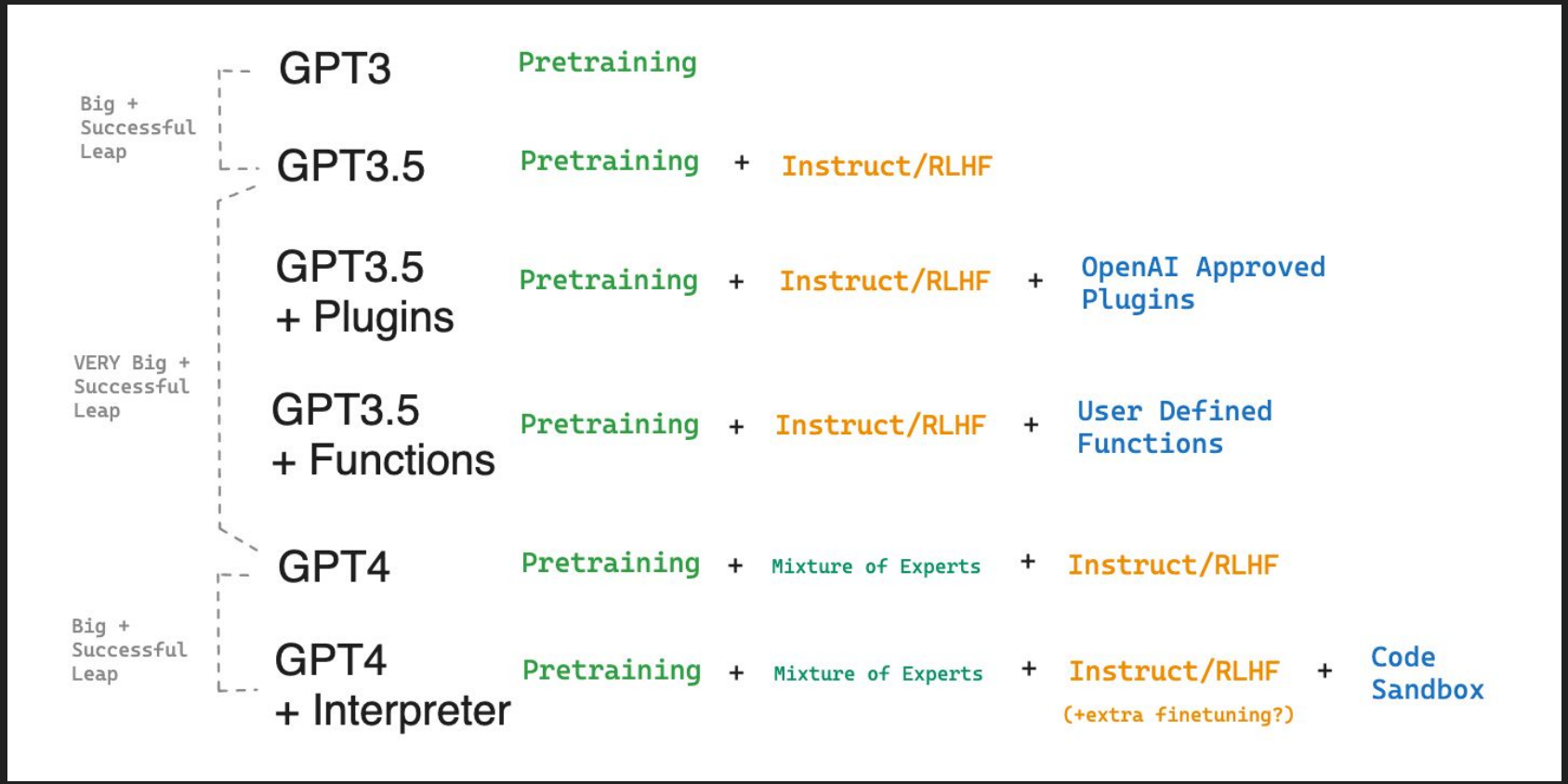
8:55 PM · Sep 9, 2022

347 Retweets **84** Quotes **2,714** Likes **413** Bookmarks

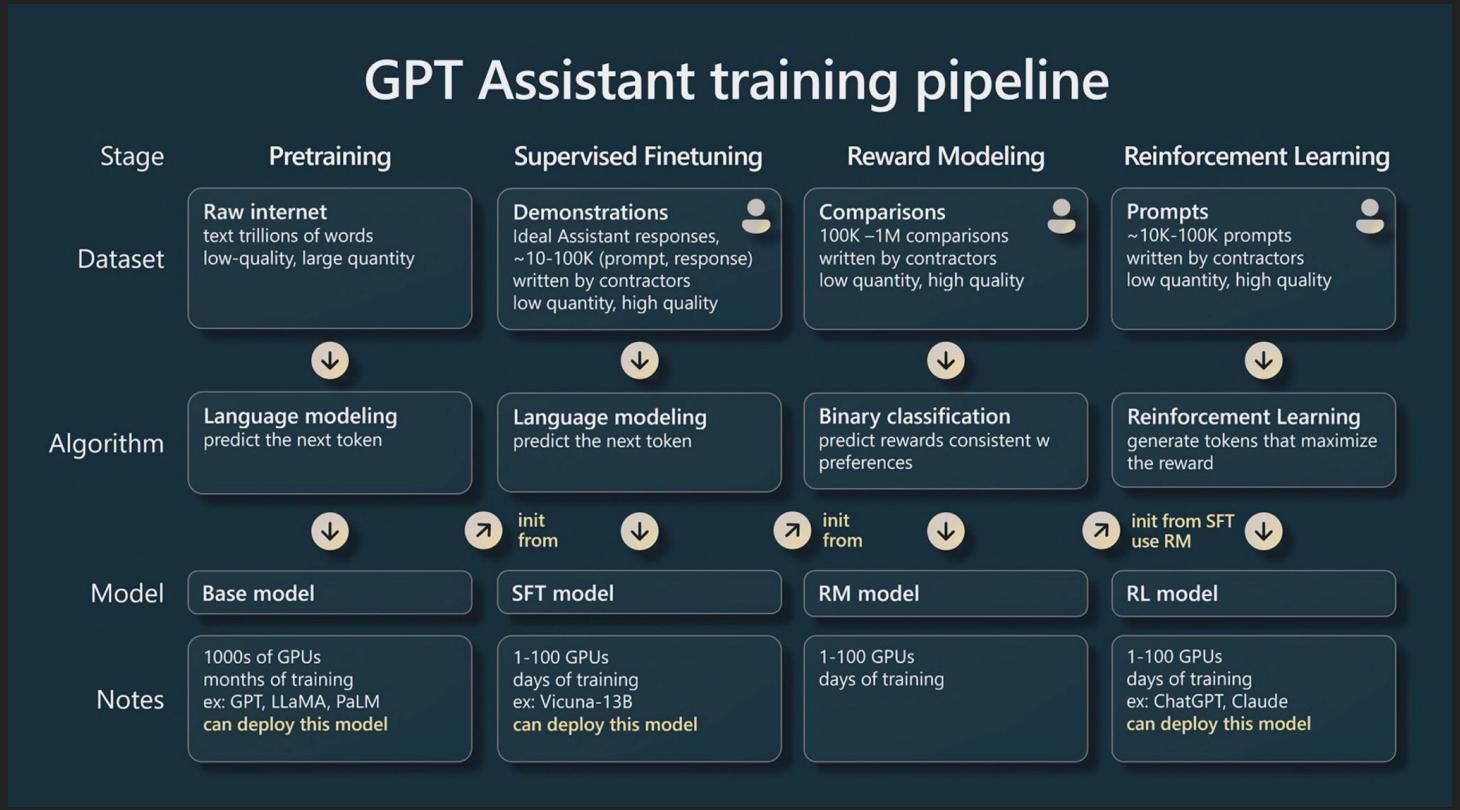
June 2023


```

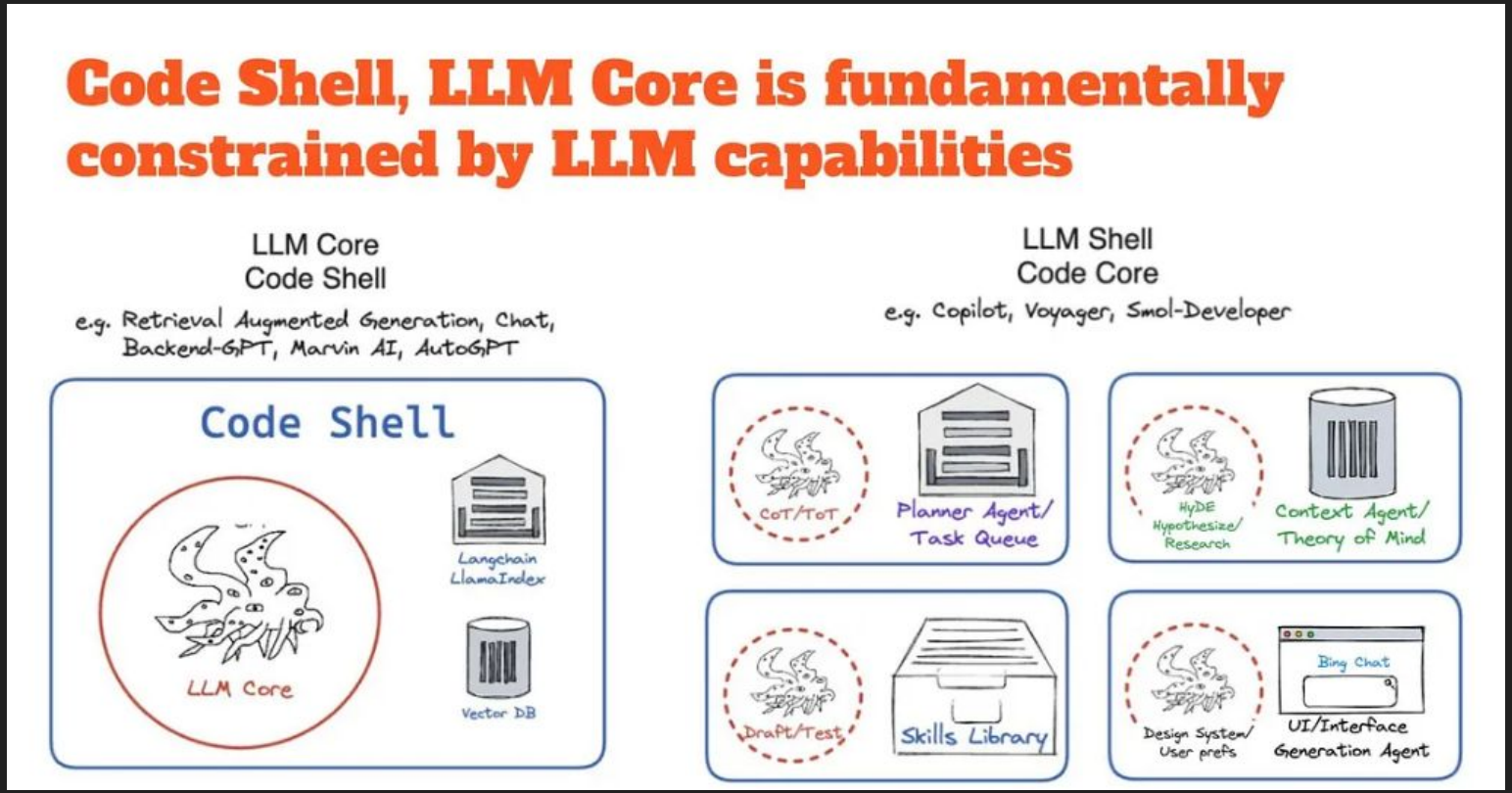
GPT4.5: Code Augmented Inference



GPT3->4 training per OpenAI

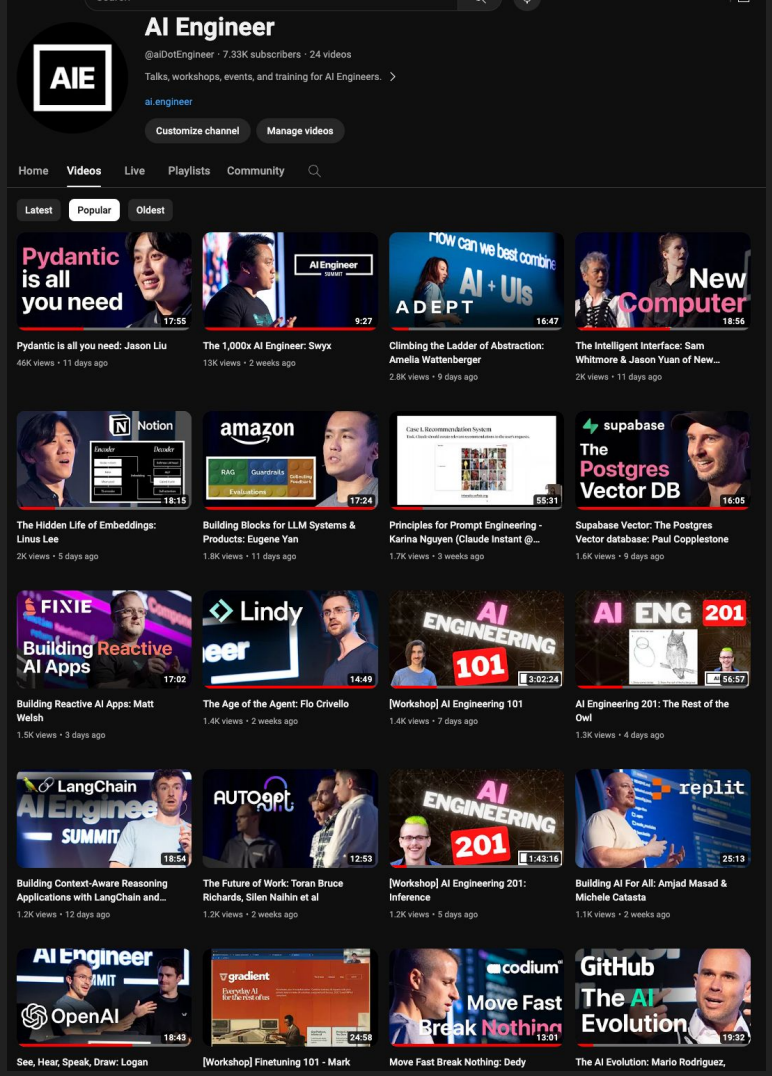


LLM Core vs Code Core



Central Problems of the AI Engineer

- **AI UX**
- **AI Eng Tooling**
  - Prompt Engineering
  - Structured Responses
  - Vector DBs
  - Evals
- **AI Productivity Devtools**
- **OSS Hosting & Infra**
- **Finetuning & Post-Training**
- **AI Agents**

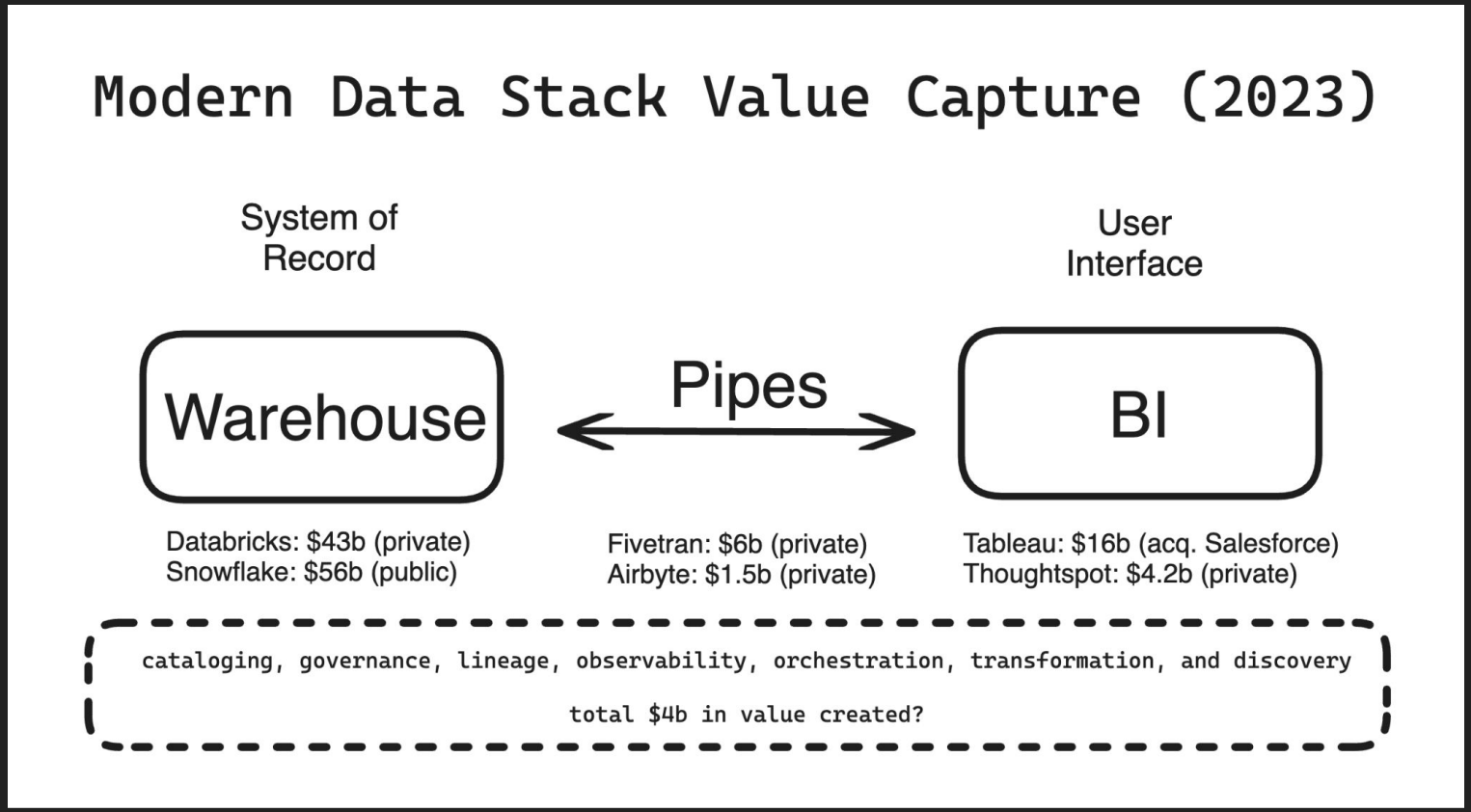


Part I: DATA

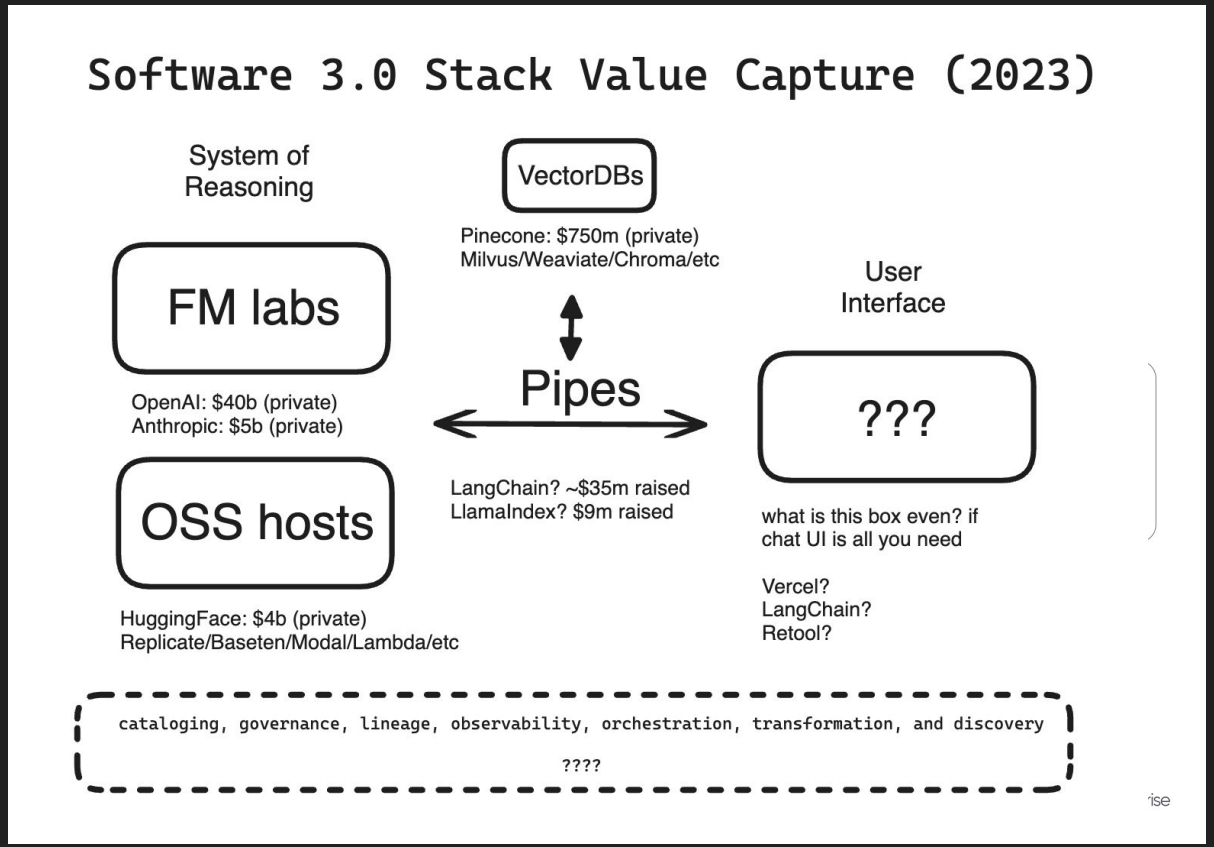
PART II: PEOPLE

Part III: TOOLS

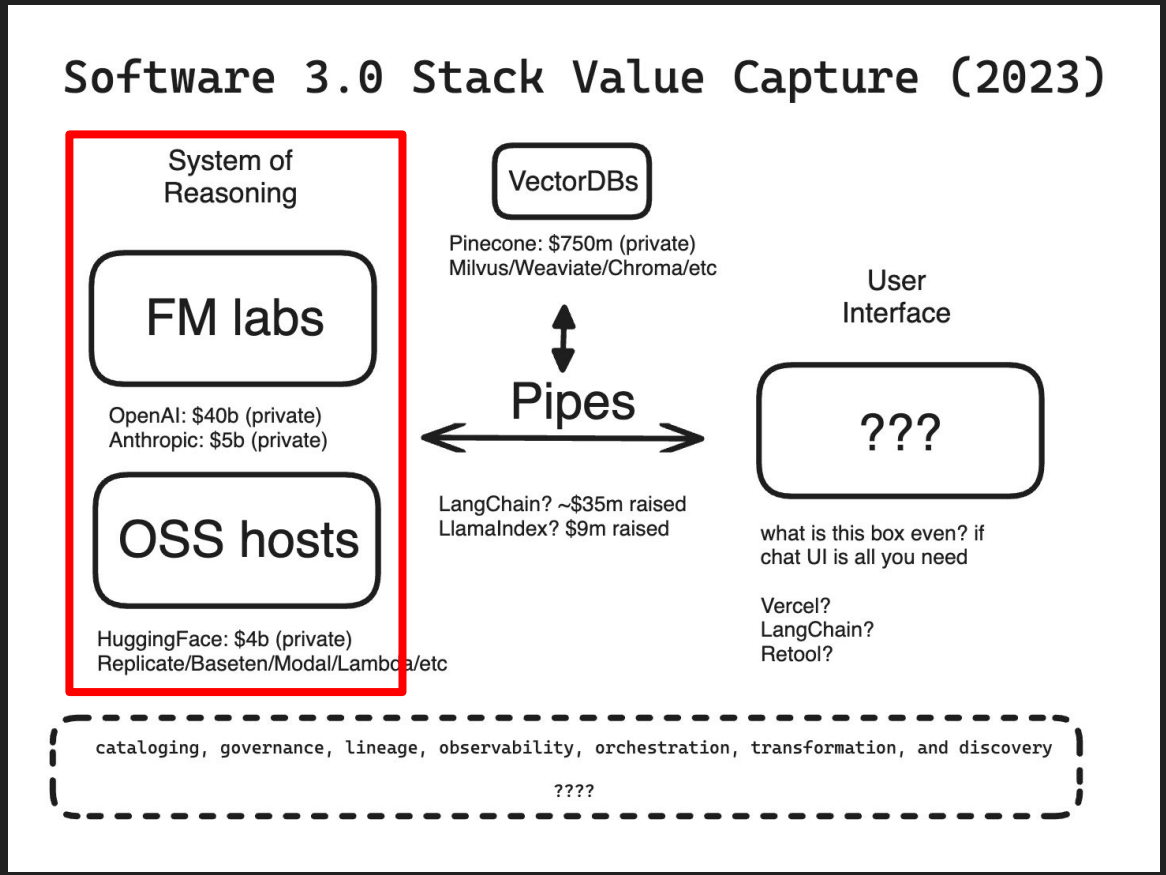
The Real Modern Data Stack



The Software 3.0 Stack



The Software 3.0 Stack



Foundation Model Labs

- OpenAI
- Anthropic
- Inflection
- Cohere
- Google/Deepmind
- HuggingFace
- Meta
- AI21/Falcon/Dolly/etc
- (others, Adept, Bloomberg…)

OSS Hosting/Training

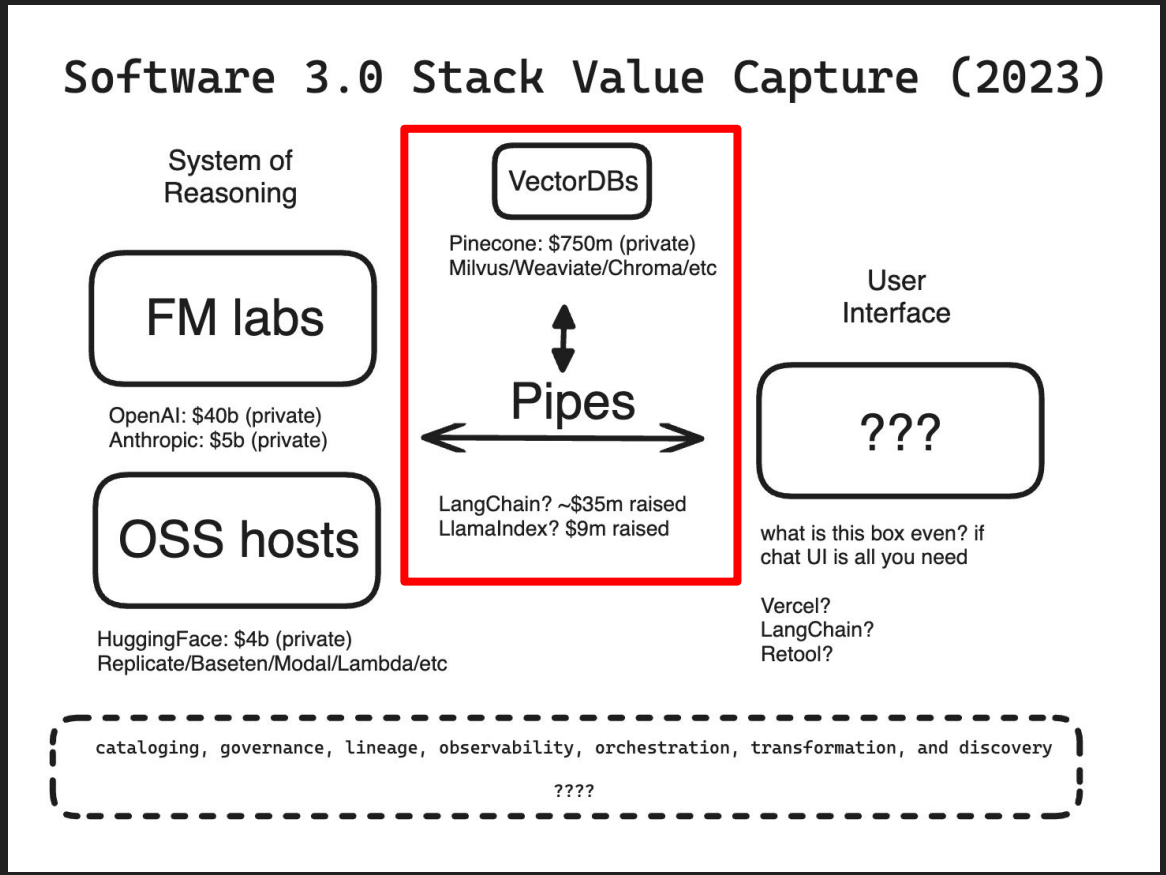
- GCP/AWS/Azure
- Replicate
- Baseten
- OctoML
- Runpod
- Anyscale
- Modal
- Fireworks.ai
- MosaicML
- PyTorch Lightning

Know the Comparisons

	ChatGPT	Claude	Bing	Bard	Perplexity	Phind	You.com	Poe	Pi
Model	✓✓✓	✓✓	✓✓	✓	✓			✓	✓
Search	x	x	✓✓	✓	✓✓	✓	✓✓	✓	x
Context	✓✓	✓✓✓	✓✓	✓					✓✓
CodeGen	✓✓✓	✓✓	✓		✓✓	✓		✓✓	
Sandbox	✓✓	x	x	✓	x	x	x	x	x

LLM Arena: https://chat.lmsys.org/

The Software 3.0 Stack



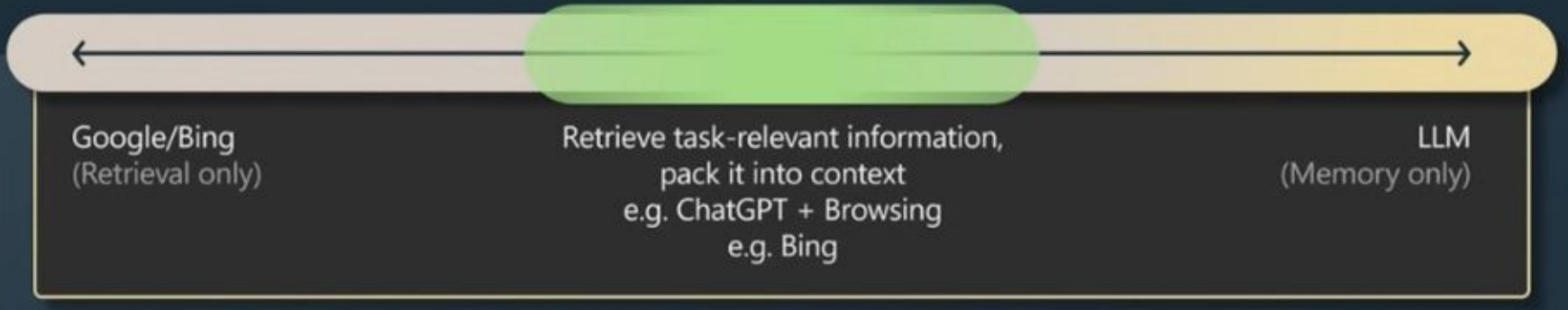
Vector DBs

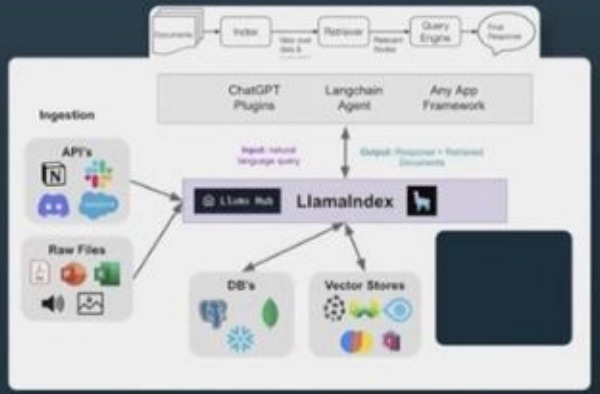
- Pinecone
- Weaviate
- Chroma
- Qdrant
- LanceDB
- Addons
  - MongoDB
  - Postgres (pgvector et al)
  - Elastic/Redis/Cassandra/etc

Frameworks/Libraries

- LangChain
- LlamaIndex
- Semantic Kernel
- Deepset Haystack
- Microsoft TypeChat
- Vercel AI SDK
- Guardails
- Guidance
- (many more….)

Retrieval Augmented Generation

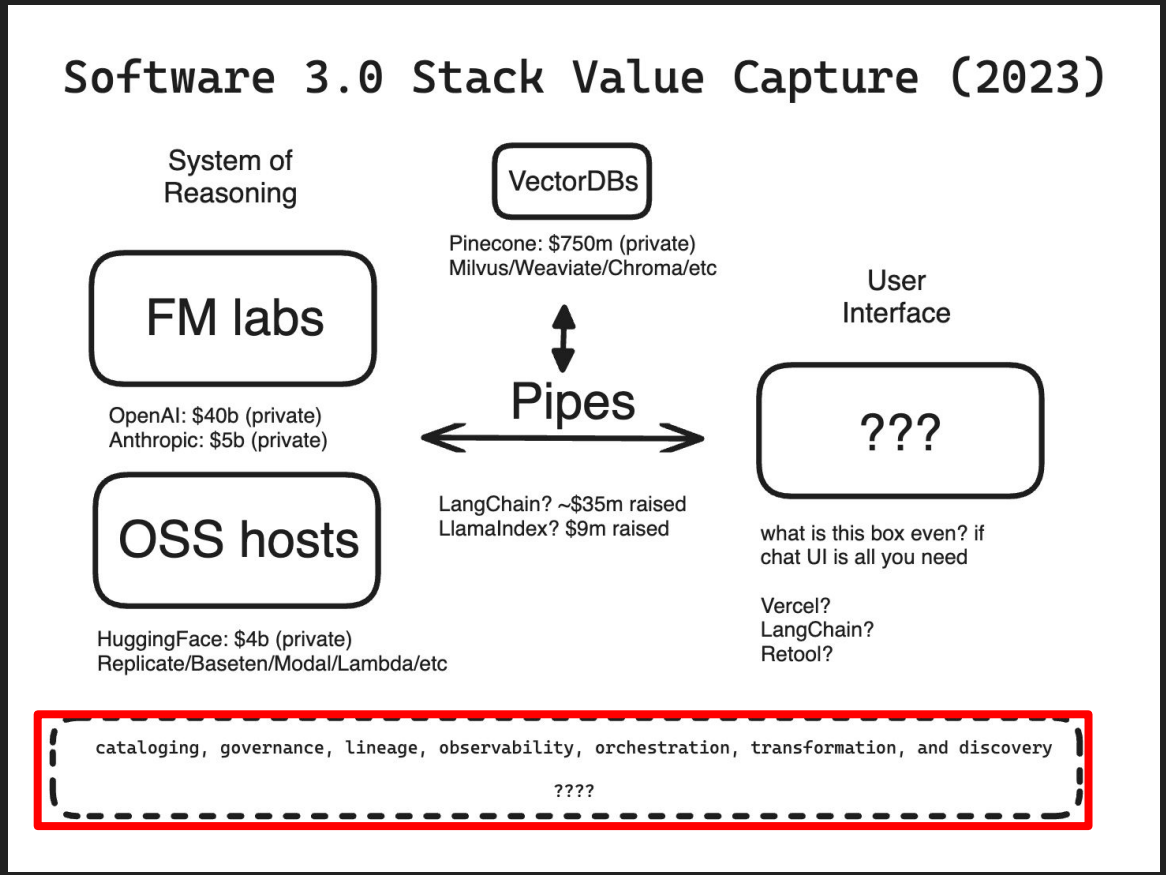




Emerging recipe:

- Break up relevant documents into chunks
- Use embedding APIs to index chunks into a vector store
- Given a test-time query, retrieve related information
- Organize the information into the prompt

The Software 3.0 Stack



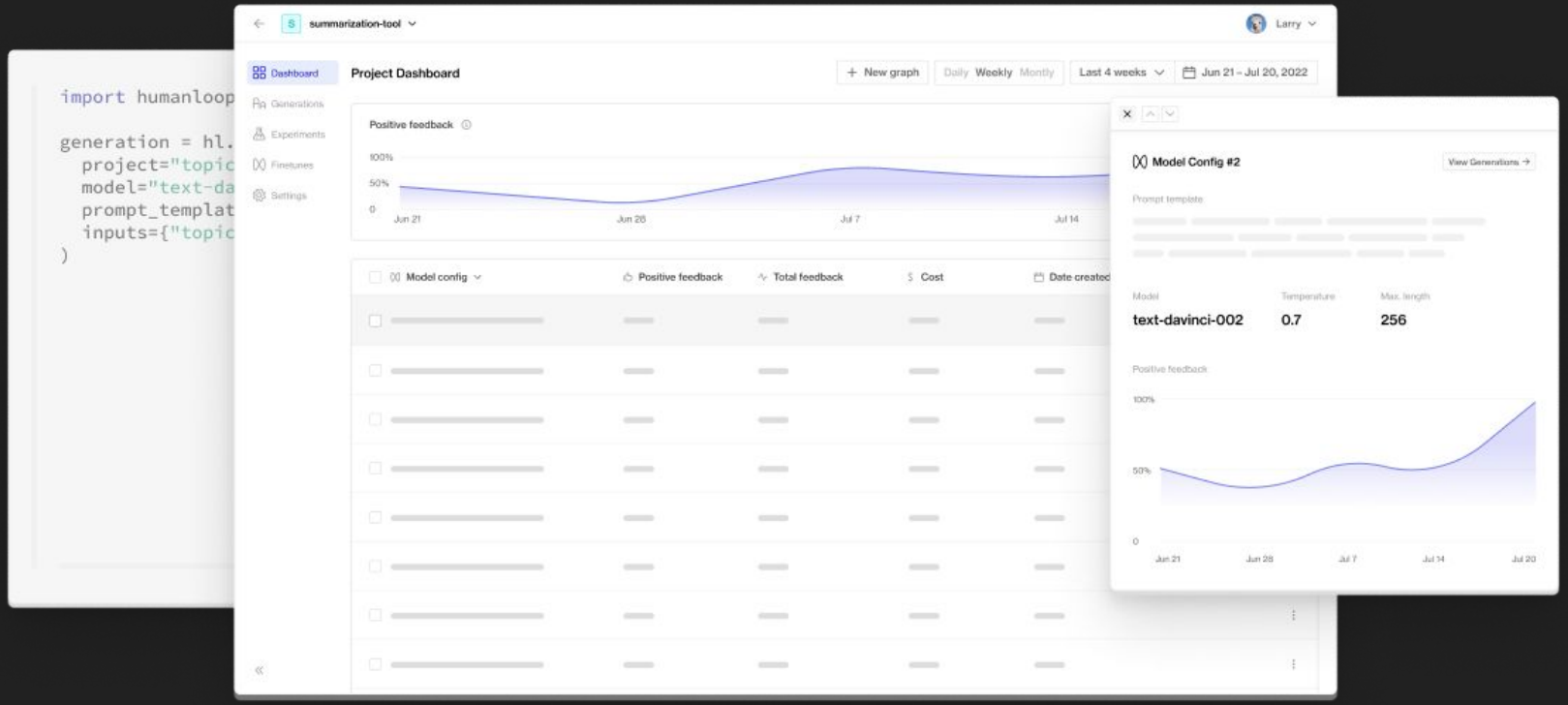
Misc - PromptOps

- Hegel.ai
- Honeyhive
- Weights & Biases
- Scale.ai Spellbook
- LangSmith Hub
- PromptLayer
- Vellum
- HumanLoop

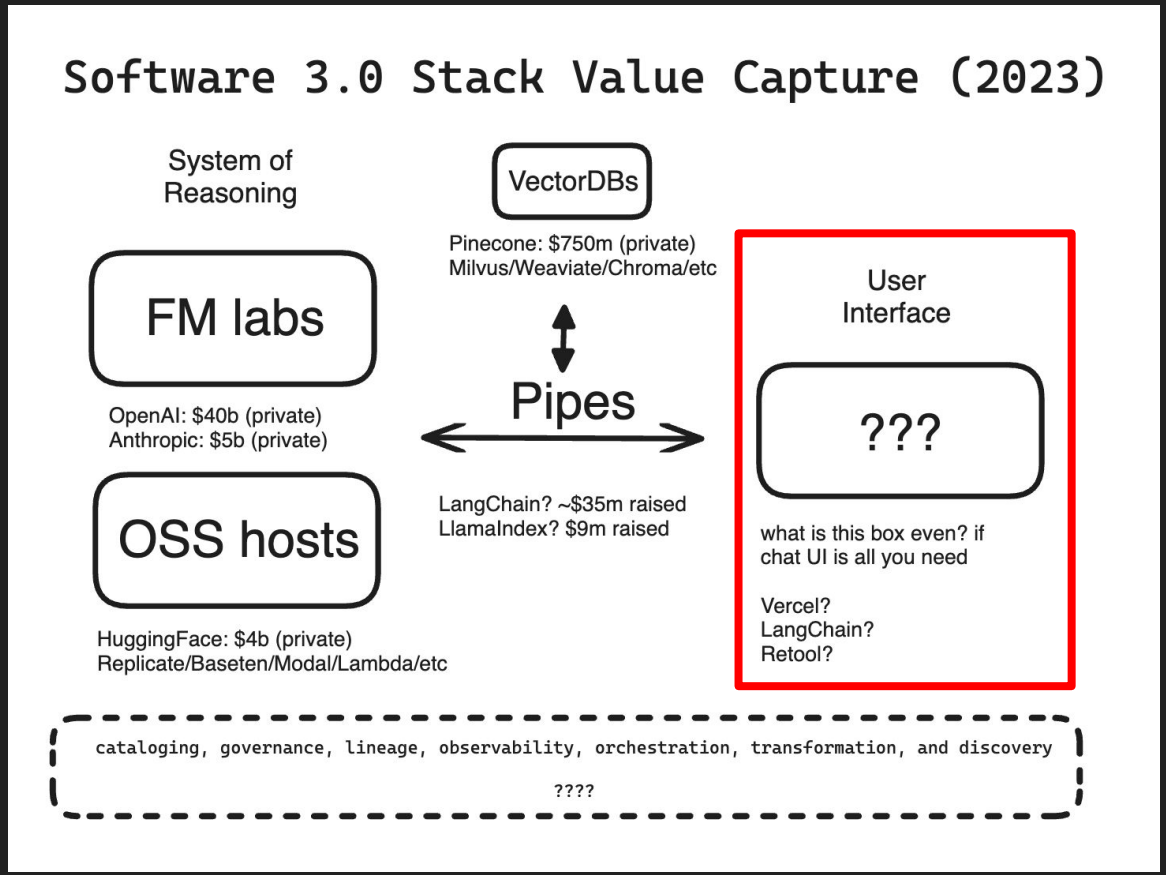
Misc - LLMOps

- Arthur
- Arize
- Fiddler
- Gantry
- Helicone
- LangSmith
- Datadog/Honeycomb/ etc

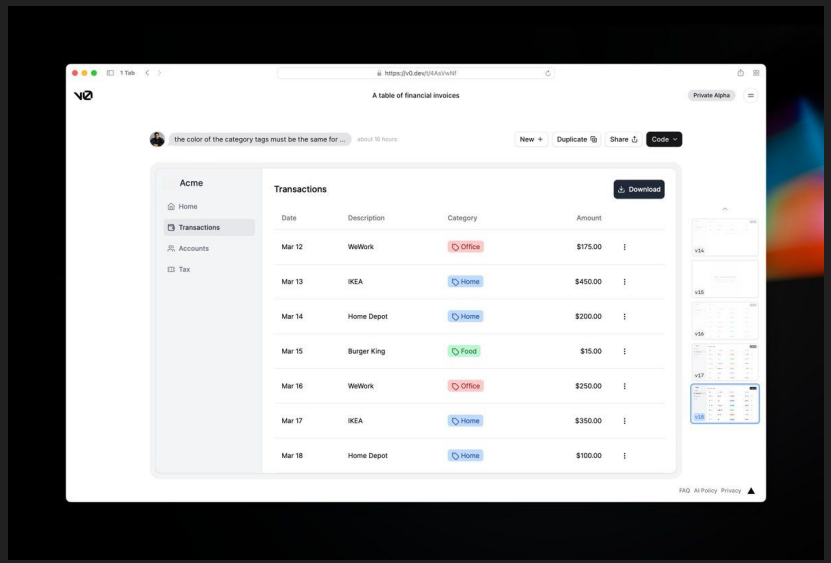
Typical PromptOps/LLMOps Landing Page



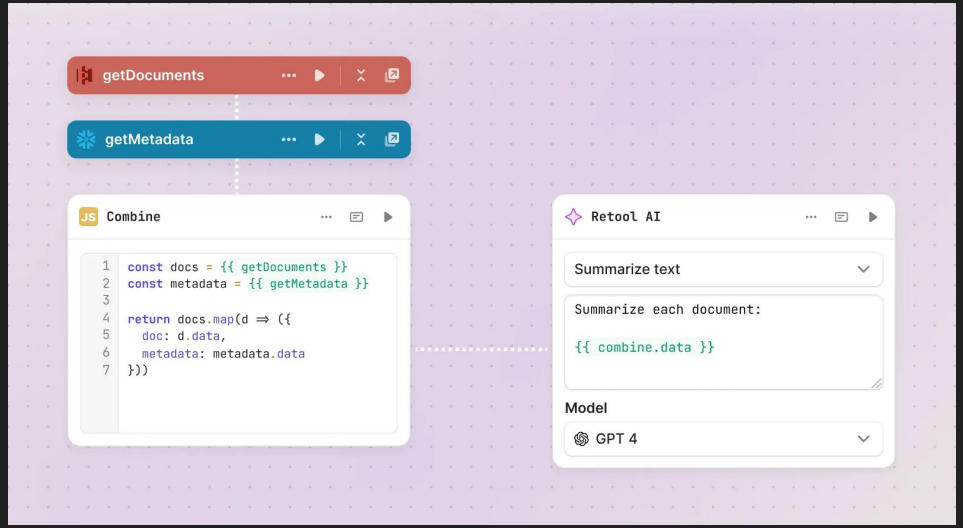
The Software 3.0 Stack



Vercel v0



Retool AI

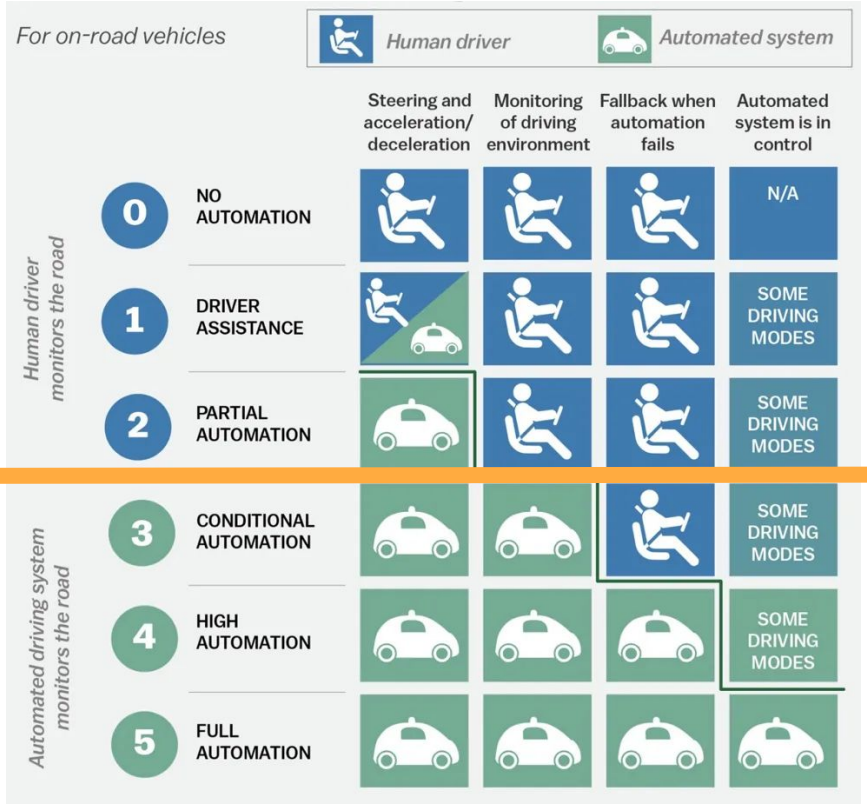


Central Problems of the AI Engineer

- **AI UX**
- **AI Tooling**
  - Prompt Engineering
  - Structured Responses
  - Vector DBs
- **AI Productivity Devtools**
- **OSS Hosting & Infra**
- **Finetuning & Evals**
- **AI Agents**

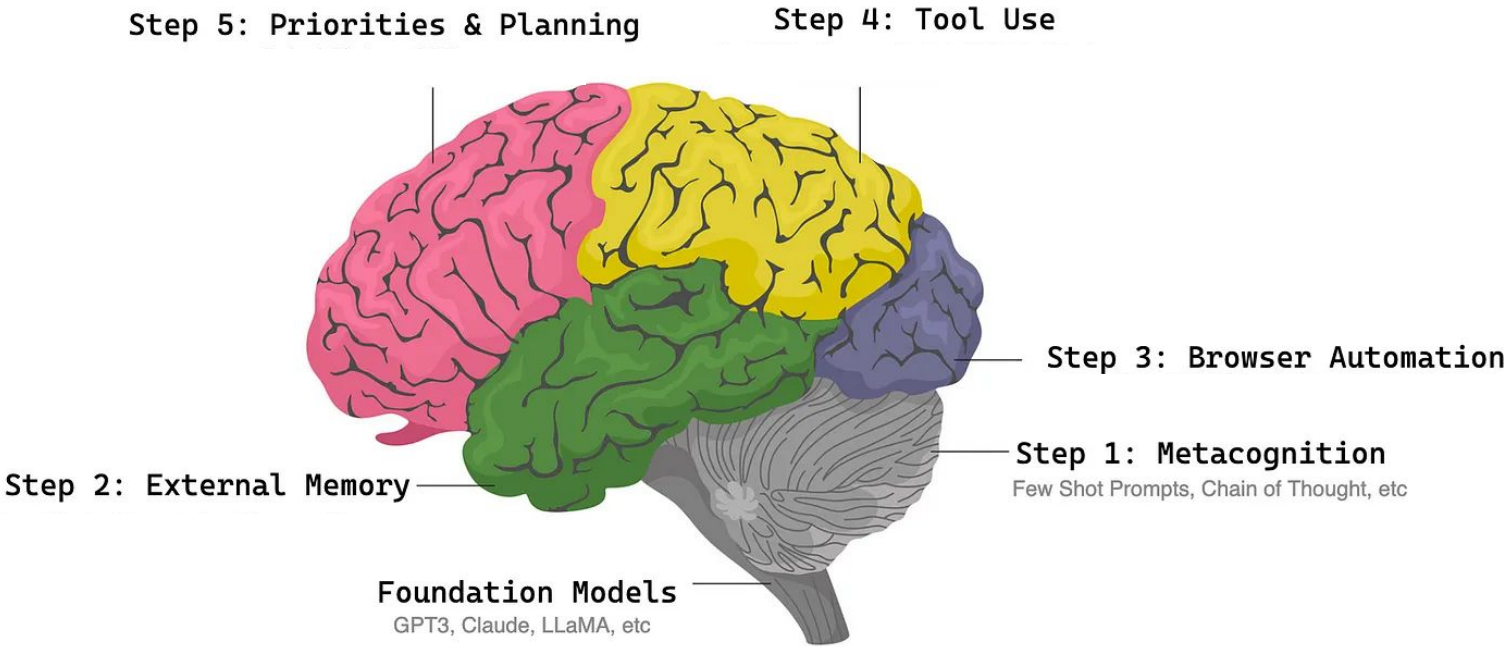
Human driver

AI driver



Anatomy of Autonomy

https://latent.space/p/agents

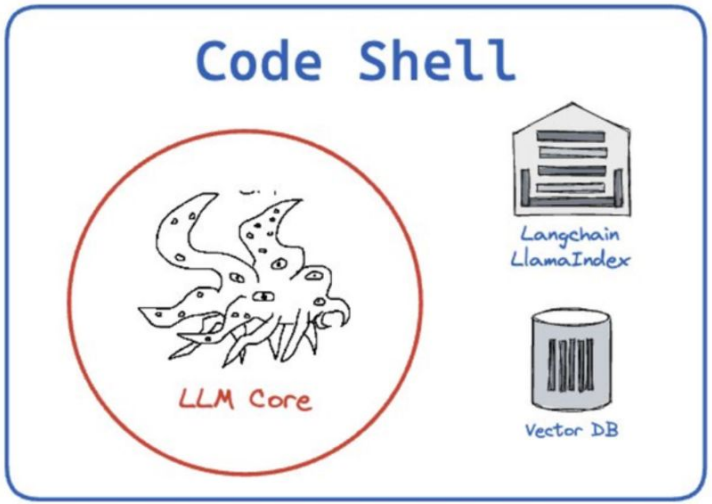


Agents = LLM + memory + planning + skills

“LLM Core” apps are fundamentally constrained by LLM capabilities

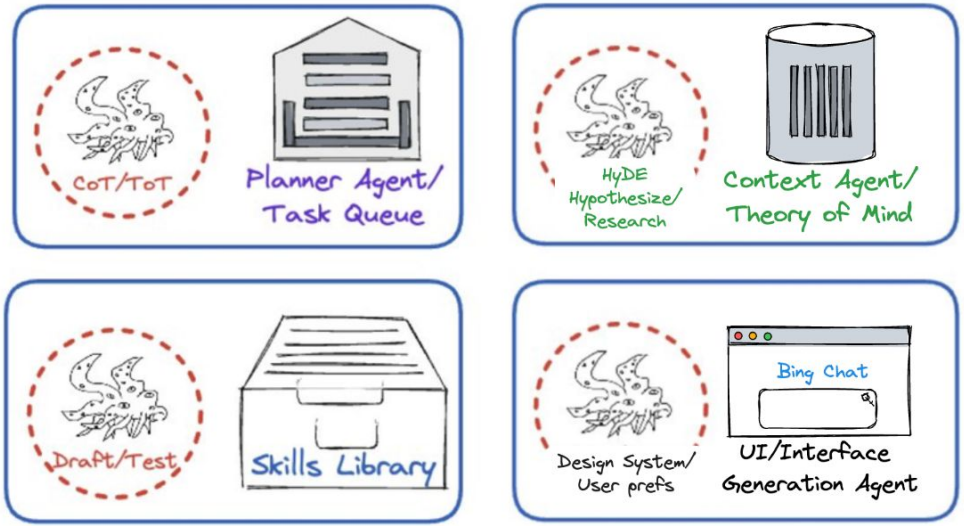
LLM Core
Code Shell

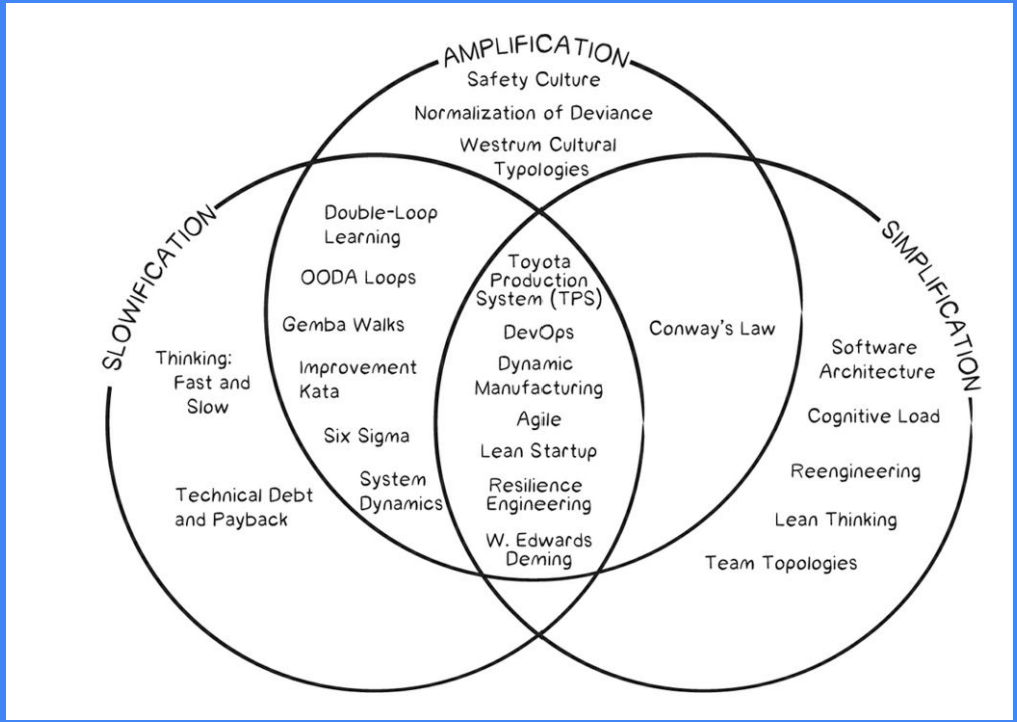
e.g. Retrieval Augmented Generation, Chat, Backend-GPT, Marvin AI, AutoGPT



LLM Shell
Code Core

e.g. Copilot, Voyager, Smol-Developer



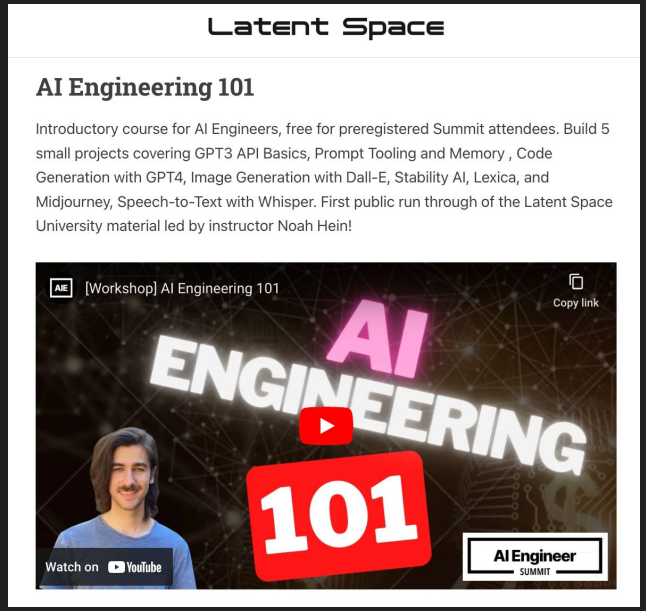


Where to start?

Slowify

https://latent.space/s/university

1. **GPT3 API** Basics
2. **Prompt Tooling** and Memory
3. **Code Generation** with GPT4
4. **Image Generation** with Dall-E, et al
5. **Speech-to-Text** with Whisper
6. Finetuning & Running **Open Source Models**
7. Build your own **AI Agents**



Simplify - Long Timeline

- 2017: **Attention** is All You Need
- 2018: **GPT-1** - 117m model
- 2019: **GPT-2** staged release of 124m - 1.5B models
- 2020: **GPT-3** - 175b model
- 2021: **Codex**, GitHub **Copilot** launch
- 2022: **InstructGPT**
- 2022: **GPT-3.5** - ? size
- 2022: **GPT4** training finished in May,
    - demoed to Bill Gates in August
- 2022: **ChatGPT**

You Are Not Too Old (To Pivot Into AI)

Everything important in AI happened in the last 5 years, and you can catch up

SWYX
31 MAR 2023

78 8 4 Share

Translated into Chinese on InfoQ.

A developer friend recently said "If I was 20 now I'd drop everything and jump into AI." But he's spent over a decade building expertise, network, and reputation to be at the very top of his field. So he's staying put for now.

Another, older, college friend is a high flying exec at a now-publicly-listed tech startup. He's good at what he does, has the perfect resume, the rest of his career could be easily extrapolated toward enviable positions. Yet he's pivoting because, as he told me, "life is short" and he doesn't want to end it wondering "what if?"

I've had many similar conversations with both technical and non-technical friends in recent days. As much as I'd like this newsletter to be about concrete technical developments and soaring SOTA advancements, I think it's worth spending one issue on the fuzzy wuzzy topic of **career pivots**, because this is one topic that I'm coincidentally uniquely qualified to offer commentary.

Pivoting in my thirties

I remember how scary my first career pivot felt, also at age 30. I was 6-7 years into the finance career I had wanted since I was 16, jet-setting around the world, grilling CEOs and helping to run a billion dollars at one of the top hedge funds in the world. Externally I was hot shit but I knew deep down it was unsatisfying and not my endgame. Making some endowments and pensions a bit richer paled in comparison to making something from nothing. I decided to pivot from finance to software engineering (and devrel). The rest is history.

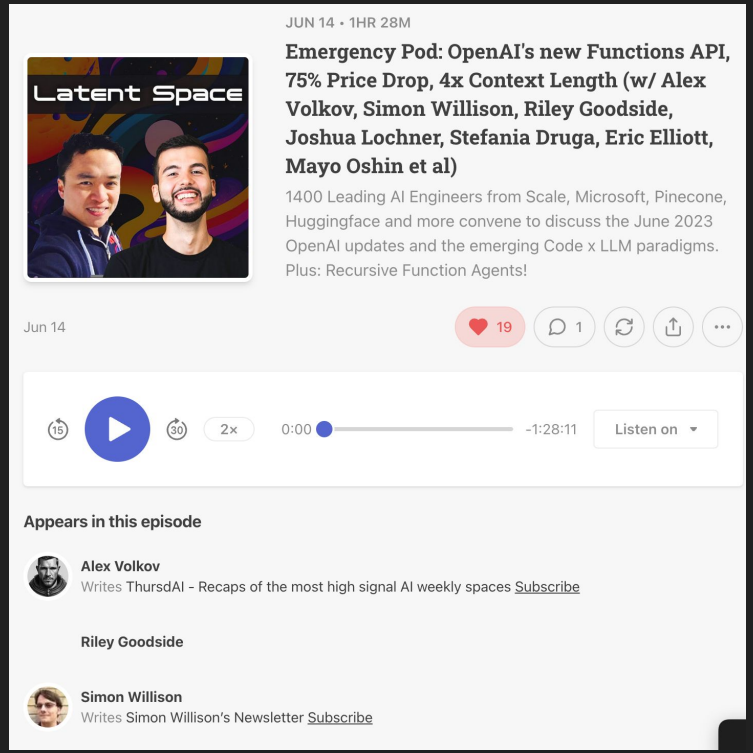
6-7 years later, I am again pivoting my career. I think a SWE → AI pivot is almost as much of a pivot as going from Finance → SWE, just in terms of superficial similarity while also requiring tremendous amount of new knowledge and practical experience in order to get reasonably productive. My pivot strategy follows the same playbook as last time; study nights and weekends as much as possible for 6 months to get confidence that this is a lasting interest [1] where I can make meaningful progress, then cut ties/burn bridges/go all in and learn it in public [2].

But that's just what works for me; your situation will be different. I trust that you can figure out the *how* if you wished; I write for the people who are looking to get enough confidence about their *why* that they actually decide to take the leap.

I think there's a lot of internalized ageism and sunk cost fallacy in tech career decisionmaking. So here's a quick list of reasons why you are not *too old* to pivot.

Simplify - ChatGPT timeline

- Nov 30 2022: The Day the AGI Was Born
- Feb 2023: Latent Space pod with OpenAI!
- Mar 2023: GPT4 demo
- Mar 2023: ChatGPT Plugins + 16k tokens
- May 2023: Google “No Moats” memo
- Jun 2023: Functions API
- Jul 2023: Code Interpreter (GPT 4.5?)
- Oct 2023: GPT3.5 Finetuning
- Oct 2023: ChatGPT Voice/TTS/Vision/DallE3
- Nov 2023: Custom GPTs and Assistants API



Simplify - Open Models

- Feb 2023 - **Meta** LLaMA 1
- Feb 2023 - **Bing** Chat
- Mar 2023 - **Falcon** 40B
- May 2023 - **Anthropic** 100k Tokens
- May 2023 - **Inflection** Pi
- May 2023 - **Mosaic** MPT-7B
- May 2023 - **Google** PaLM 2
- Jul 2023 - **Anthropic** Claude 2
- Jul 2023 - **Meta** LLaMA 2 + CodeLlaMA
- Sep 2023 - **Falcon** 180B
- Sep 2023 - **Mistral** 7B

The New Kings of Open Source AI (Oct 2023 Recap)

Mistral is the new open source unicorn in town, top takes from the AI Engineer Summit, and our usual highest-signal recap of top items for the AI Engineer from Oct 2023

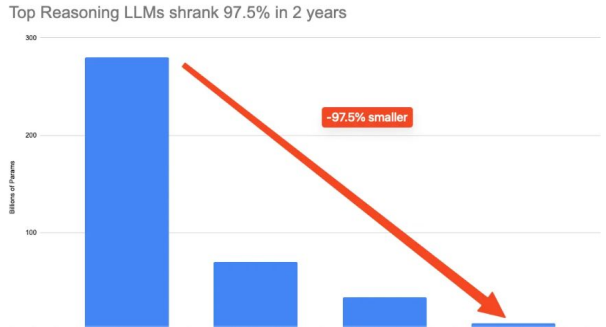
SWYX AND NOAH HEIN
NOV 12, 2023

28 2 1 Share

We're sorry that this monthly recap is delayed - it feels futile to cover >1 month old news in AI but we're still committed to recapping things monthly, so as to provide useful historical perspective for future readers of this newsletter. This work is as much a part of our process for keeping up to date as it is for you to read.

Join us to celebrate the One Year Anniversary of ChatGPT and at Modular ModCon!

Move over, Meta, there are new open source kings in town. **Mistral 7B,** released at the tail end of Sept 2023, is both **Apache 2.0** and **smaller but better** than Llama 2, and is now rumored to be raising **$400m at $2.5b valuation** from a16z:



Amplify

https://ai.engineer

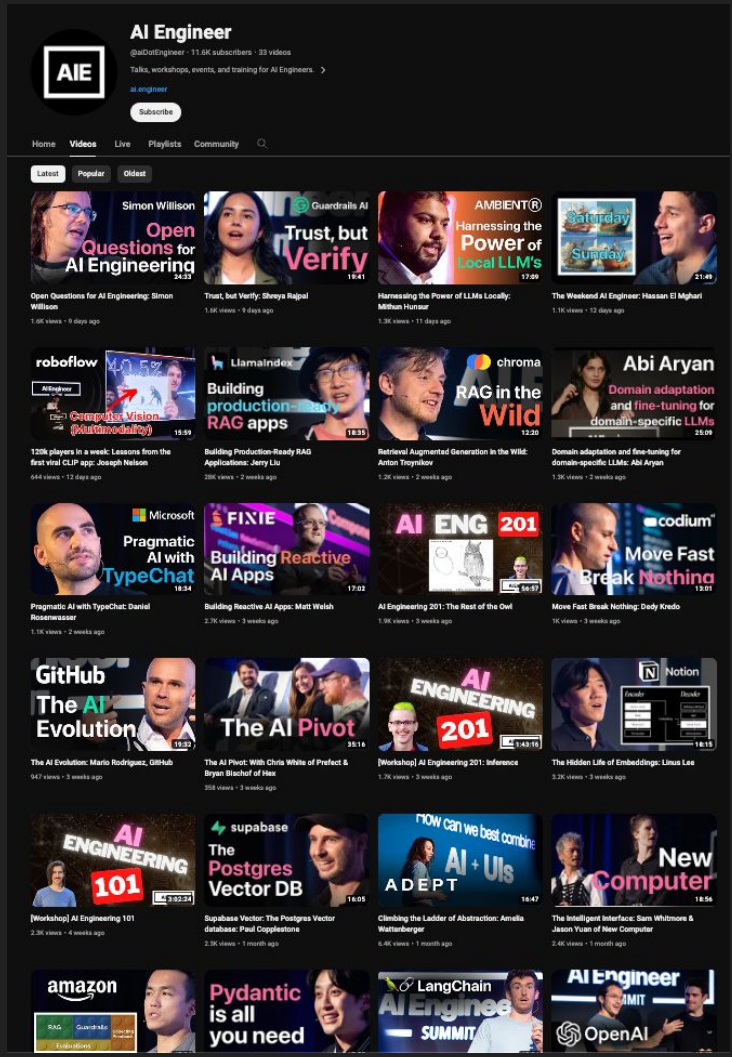
Featuring talks from speakers representing leading AI companies & technologies

GitHub · OpenAI · FIXIE · amazon

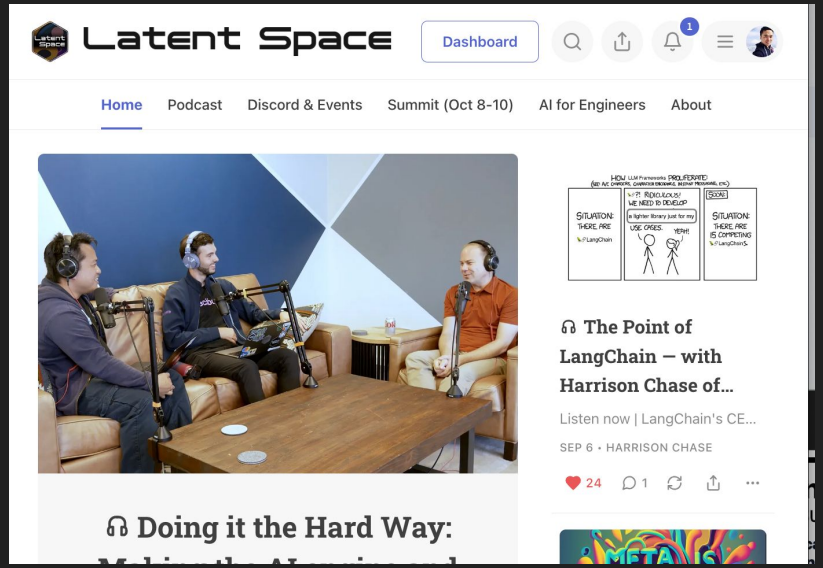
LangChain · LlamaIndex · Vercel · Notion

replit · ANTHROP\C · HEX · ADEPT

Full speaker schedule and talks available now!



Thank You!



latent.space

KIV

Tech's Two Philosophies

Tech's Two Philosophies

Wednesday, May 9, 2018

🤖 The Zuck School

Tech tracks you,
feeds you,
decides for you

🚲 The Jobs School

Tech is bicycle for the mind

BabyAGI - Interrupt based Level 3 Agent

replit Features Blog Pricing Teams Pro Careers Shop Sign Up Log In

Files
- main.py

Packager files
- poetry.lock
- pyproject.toml

```
def task_creation_agent(objective: str, result: Dict, task_description: str, task_list: List[str]):
    prompt = f"You are an task creation AI that uses the result of an execution agent to create new tasks with the following objective: {objective}, The last completed task has the result: {result}. This result was based on this task description: {task_description}. These are incomplete tasks: {', '.join(task_list)}. Based on the result, create new tasks to be completed by the AI system that do not overlap with incomplete tasks. Return the tasks as an array."
    response = openai.Completion.create(engine="text-davinci-003",prompt=prompt,temperature=0.5,max_tokens=100,top_p=1,frequency_penalty=0,presence_penalty=0)
    new_tasks = response.choices[0].text.strip().split('\n')
    return [{"task_name": task_name} for task_name in new_tasks]

def prioritization_agent(this_task_id:int):
    global task_list
    task_names = [t["task_name"] for t in task_list]
    next_task_id = int(this_task_id)+1
    prompt = f"""You are an task prioritization AI tasked with cleaning the formatting of and reprioritizing the following
```

Fork 240 Run 48 Hide code ···

babyagi

YoheiNakajima 185 followers + Follow

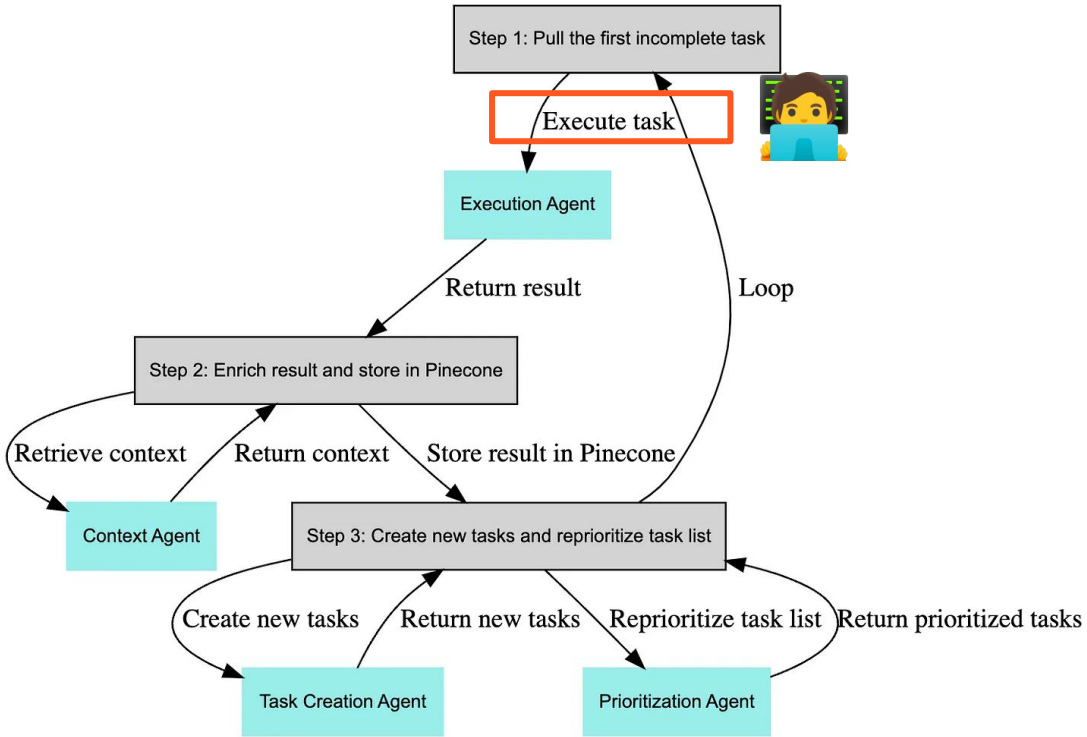
Apr 10, 2023 · 1.6K runs · Made with Python

The original commit of Baby AGI at 105 lines of code + comments.

See evolved BabyAGI on Github here: https://github.com/yoheinakajima/babyagi

poetry.lock

HOW TO USE



AI UX = Presence + Experience + Utility

How to Make AI UX Your Moat

Design great AI Products that go beyond "just LLM Wrappers": make AI more present, more practical, and then more powerful.