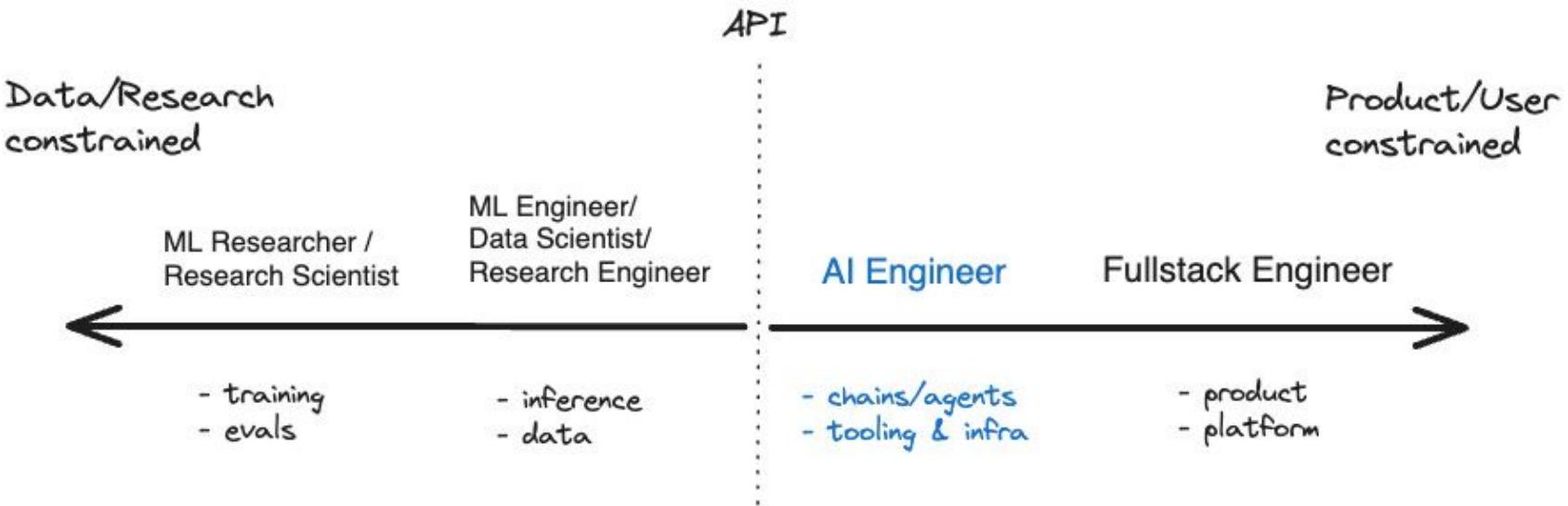




The Rise of the AI Engineer

swyx.io/ai-landscape

The new role created

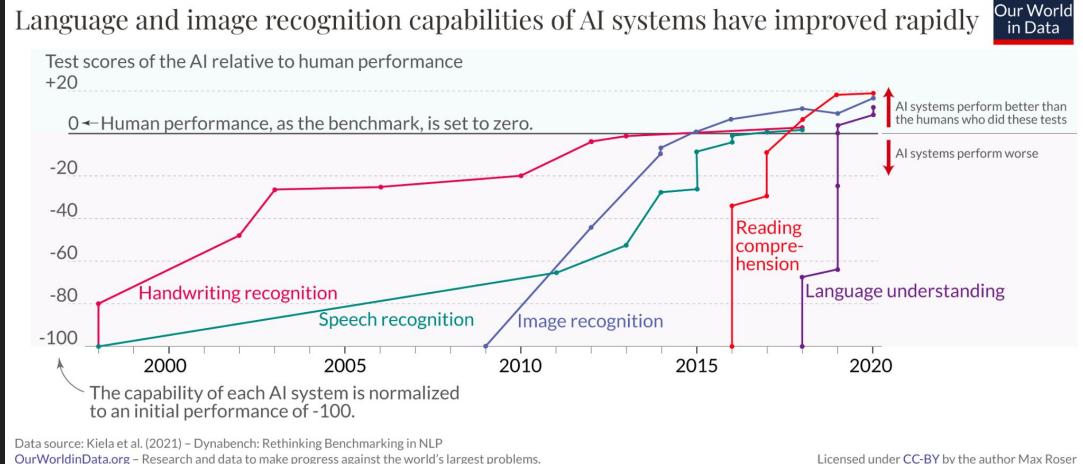


Part I: DATA

The Moore's law of our time

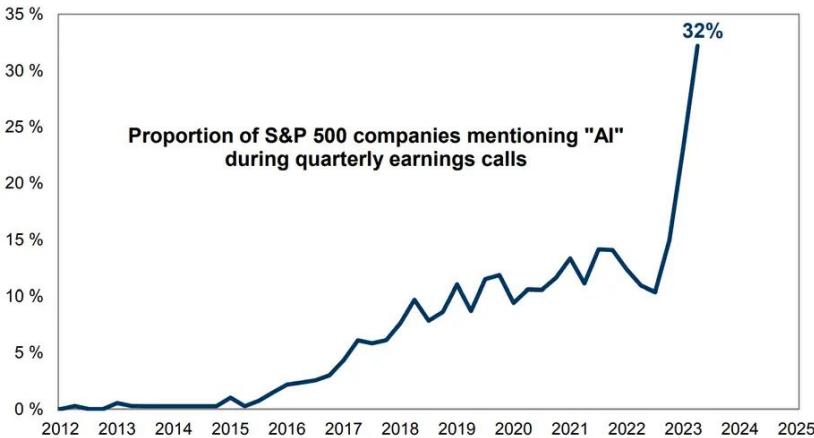
Timeline of images generated by artificial intelligence
These people don't exist. All images were generated by artificial intelligence.

Our World
in Data



AI Madness

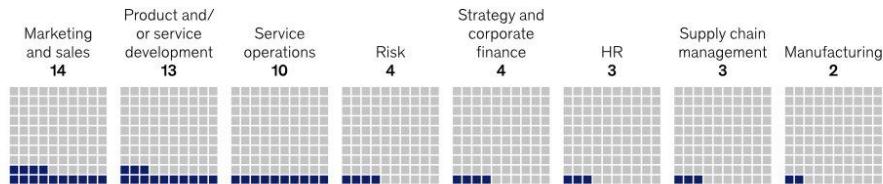
Exhibit 2: Mentions of "AI" on earnings calls have risen sharply this quarter
as of 2Q 2023



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 %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

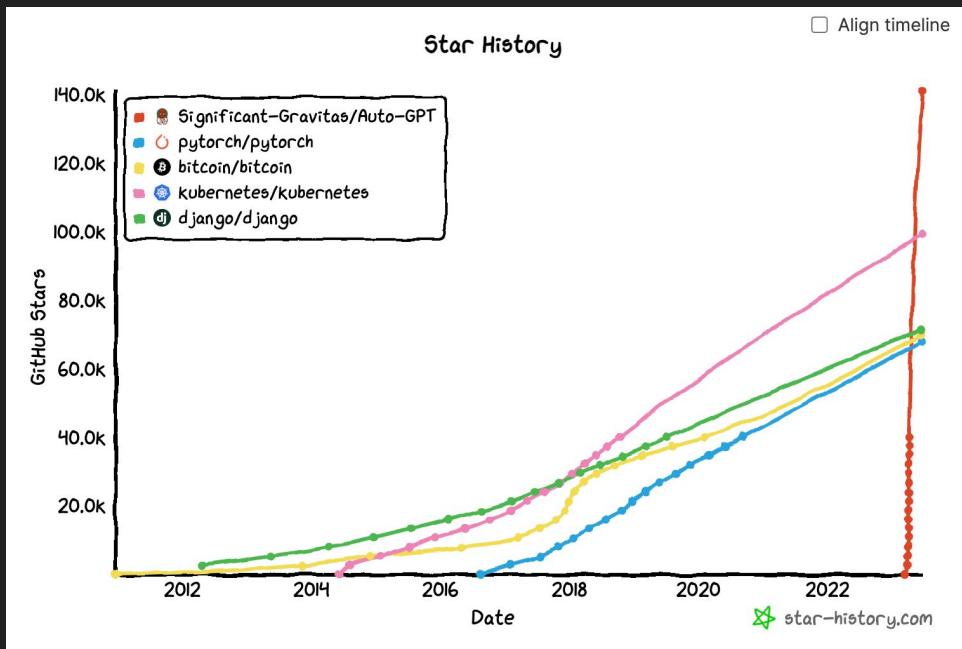
¹Questions were asked of respondents who said their organizations have adopted AI in at least 1 business function. The data shown were rebased to represent all respondents.

Source: McKinsey Global Survey on AI, 1,684 participants at all levels of the organization, April 11–21, 2023

AI AI I/O



AI Madness



AI Manic Depression



sarah guo // conviction  
@saranormous

late 2022-23 AI interest has peaked?

3:12 PM · Aug 12, 2023 · 78K Views

39 comments, 8 retweets, 96 likes



sarah guo // conviction   @saranormous · Aug 25
vc markets have gone completely wild

28 comments, 7 retweets, 194 likes, 154.4K views



sarah guo // conviction   @saranormous · Aug 25
It's giving 2021

3 comments, 1 retweet, 30 likes, 15.2K views

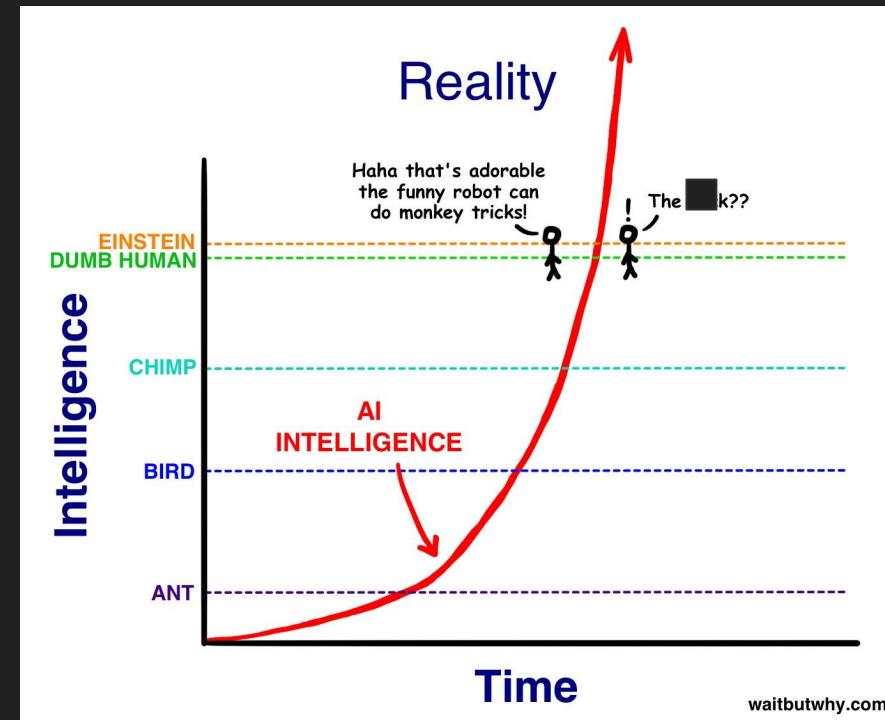
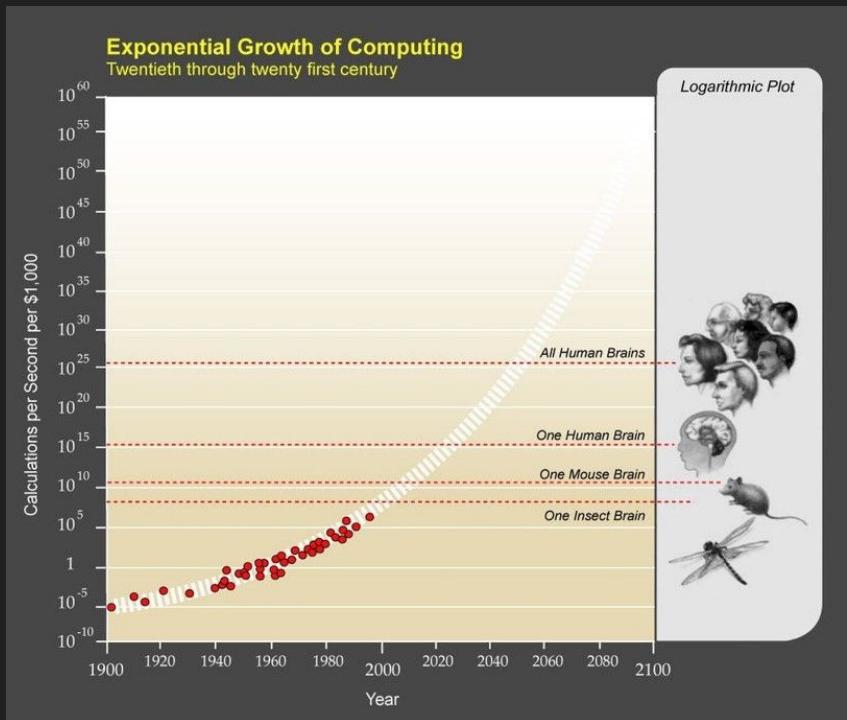


sarah guo // conviction   @saranormous

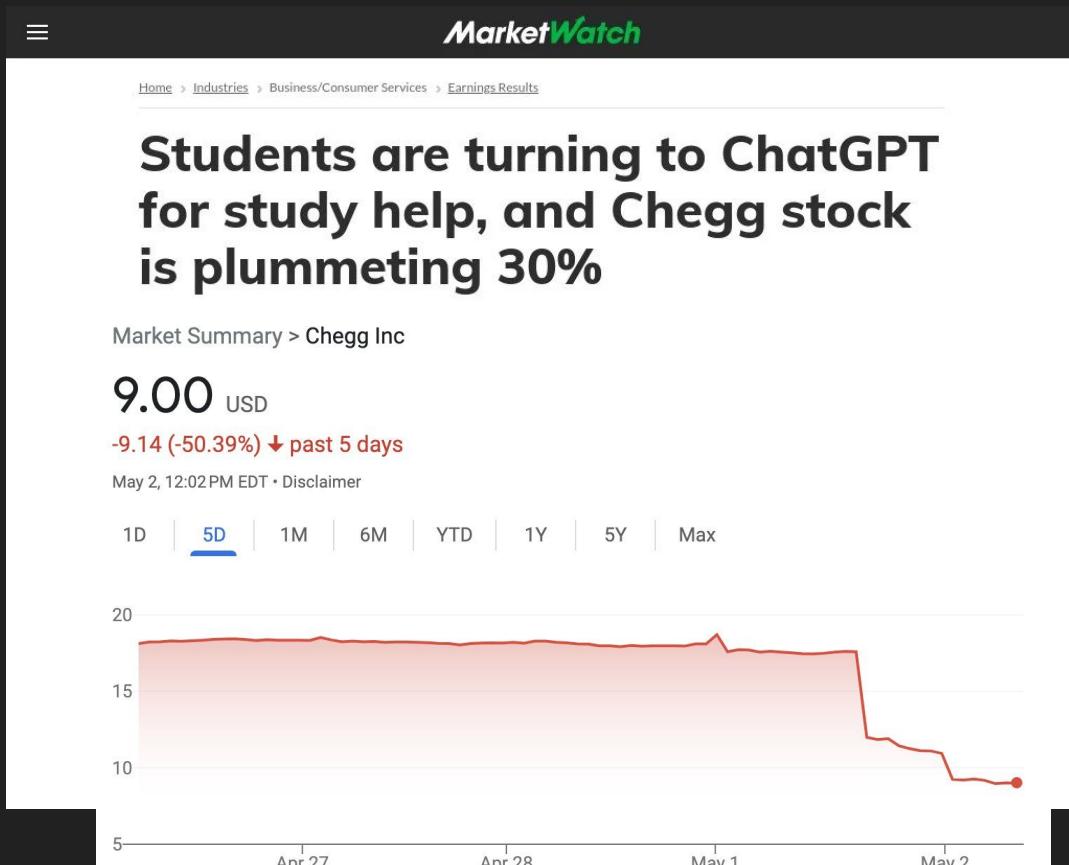
to those who asked for more context:
AI investments where another firm lobs in at 2X the price the day after wires land

7:22 PM · Aug 26, 2023 · 2,901 Views

AI Fears



AI Blame



It's Impossible to Keep Up

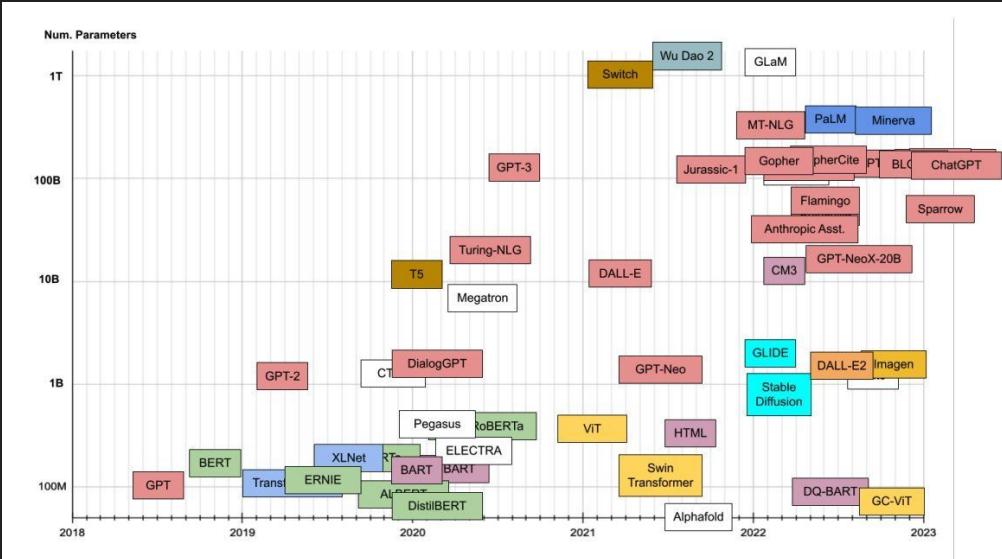


Figure 8: Transformer timeline. On the vertical axis, number of parameters. Colors describe Transformer family.

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 agreement 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](#), [walkthrough](#)) was a big pest 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



@levelsio ✅
@levelsio

Made over \$1,000,000 in 10 months with AI now:

- [InteriorAI.com](#) (Sep '22)
- [AvatarAI.me](#) (Oct '22)
- [PhotoAI.com](#) (Feb '22)



Interior AI
Home Payments Balances Customers Products Billing Reports

Last 12 months ▾ Aug 11, 2022–Aug 10

Daily ▾

Gross volume ⓘ +∞

\$388.4K \$0.00 previous period

\$3,041.24



Building & Selling a SaaS for +\$1M in 8 Months



Danny Postma

January 18, 2022



Marketing



Indonesia



\$10k-\$25k/mo

Danny founded Headline, an AI-powered copywriting SaaS that writes marketing copy for you automatically. It reached \$20K MRR in February 2021, and finally got acquired in March 2021 for a 7 figure sum by Jarvis.ai.

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

Aran Komatsuzaki  @arankomatsuzaki ...

Here's the leaderboard of prompts to add to GPT-3.
Can you guys come up with anything better?

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

Ask HN: How to Break into AI Engineering

126 points by dragonmouse 6 days ago | flag | hide | past | archive | favorite | 64 c

What are some great resources for learning the skills and k
engineer?

▲ __rito__ 6 days ago | flag | favorite | next [-]

- Have crystal clear Mathematical foundations, as in why this formula/methodology works. Be able to solve college/HS test problems. Really solid footing in Differential Calculus and Linear Algebra.
- Know the Statistical language that you learn from a basic college-level Statistics course. Be able to convert English sentences into those using Statistical notation, and be able to read easily. A good book for this is "Statistics for Data Science".
- You already know programming, I assume. Learn Python if you don't know it. It's the most common language used in AI.
- There are a number of paths you can go from there. Here's what I did.
- IBM Data Science Professional Certificate (not deep at all, but lays out the basics)
- Machine Learning for Absolute Beginners by Oliver Theobald which you can find for free online.
- Machine Learning Specialization by Andrew Ng on Coursera.
- Deep Learning Specialization by Andrew Ng on Coursera.
- fast.ai course.

- Learn PyTorch really well. I suggest Sebastian Raschka's book.

Now from here, you can chart your own path. You can choose NLP/Proc, Vision, Reinforcement Learning, etc. I went towards Vision. And I do Edge AI as hobby.

▲ santiagobasulto 6 days ago | flag | favorite | prev | next [-]

AI Engineering is basically **Data Engineering** focused on AI. When in "traditional" Data Engineering you store processed data in something like a Data Lake, in AI Eng. your end storage might be a database (MySQL, PostgreSQL, etc) or a data store (Feast or GCP Vertex AI).

There are some AI Engineers with strong scientific/mathematical background, but that's rare. Most of them have a background in Computer Science and work on the infrastructure that these ML people that actually develop and evaluate the models.

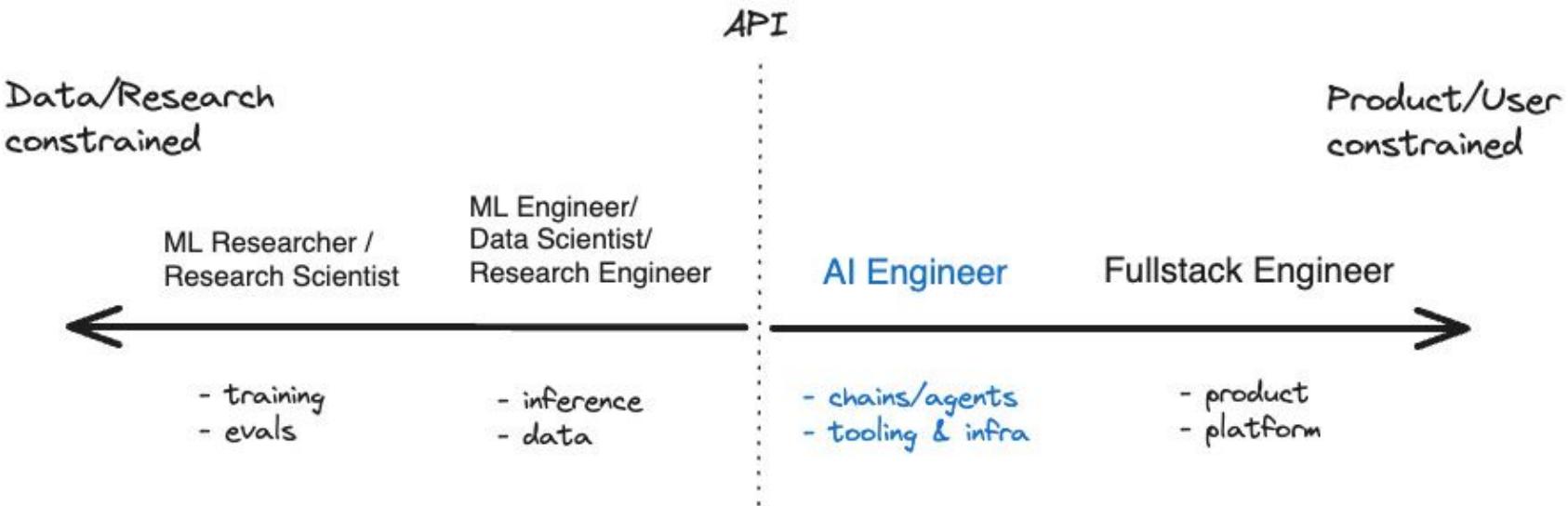
So my advice is to start with **Data Engineering** and then find a specialization AI. You should learn about databases, data processing, data pipelines, etc. Also, a lot of concepts of "data wrangling". Understanding how data flows from point A to point B, how the intermediate storages and streaming engines work, etc. Functions, etc.

[0] <https://github.com/feast-dev/feast>

reply

▲ isbx1 6 days ago | flag | favorite | parent | next [-] [collapse root]

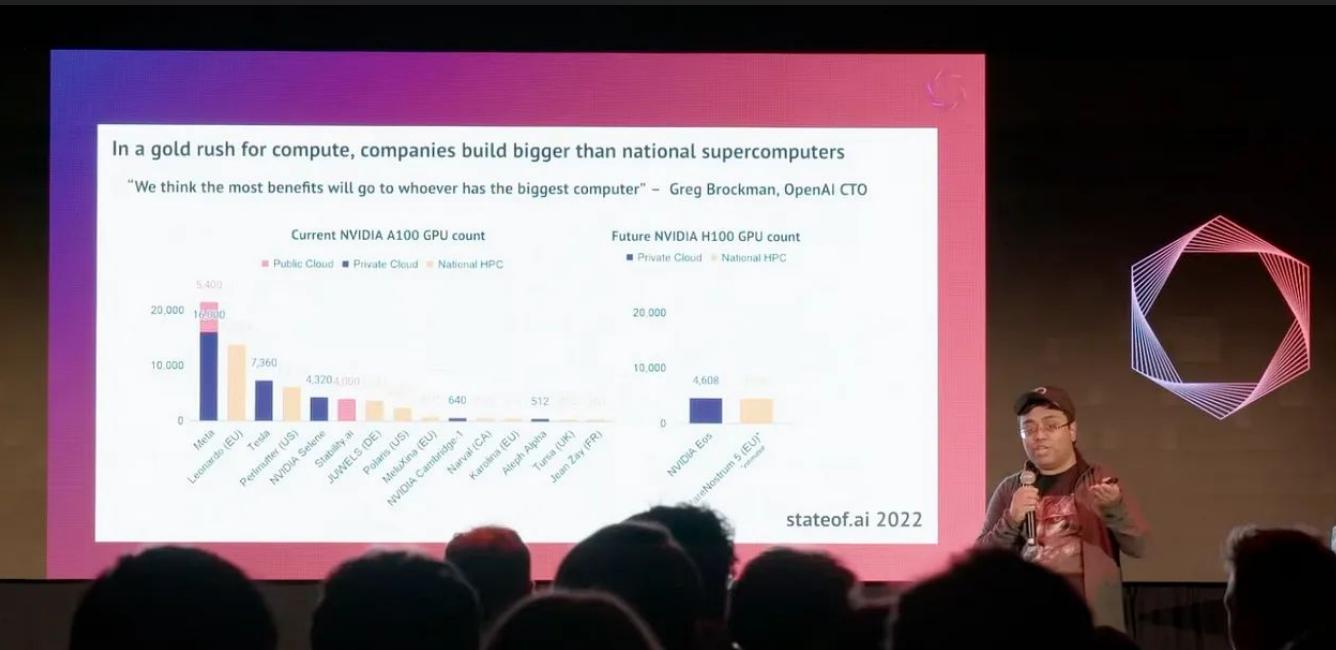
The new role created



5 reasons for the Rise of the AI Engineer

- Economics
- Sociology
- Tech
- Product
- Language

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



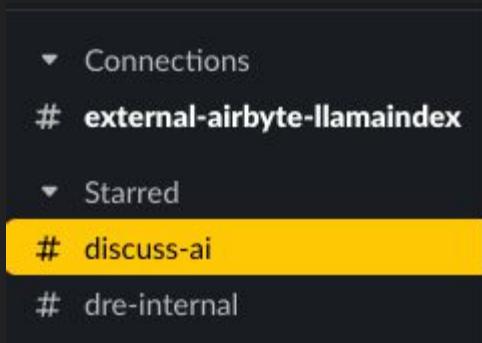
GPU Hoarding:

- 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



A large white arrow points from the left side of the image towards the right, indicating a transition or flow from the left-hand sidebar to the main channel list.

The main channel list includes:

- VERIFIED MEMBER CHAT (DEVT... +)
 - # founders
 - ↳ Urgency
 - # devtools-angels
 - ↳ devrel-devex-leads
 - # devtools-deals
 - ↳ Quadratic - The data spr...
 - # Dimension
 - ↳ cloud-infra
 - ↳ We Raised A Bunch Of M...
 - # databases-data-engineer...
 - # dev-productivity
 - ↳ lowcode-nocode
 - # security
- MARKETS, MONEY, PERSONAL ... +
 - # hiring-and-jobs
 - # stocks-and-macro
 - # comp-taxes-401ks
 - # creator-economy
 - ↳ beehiv
 - # fire
 - # crypto
- GENERAL DISCUSSION & SOCIA... +
 - # intro-yourself-pls
 - ↳ PromptEditor.io
 - # watercooler
 - ↳ anyone getting anything ...
 - ↳ not sure where really to ...

It's also enabled by tech

Andrej Karpathy @karpathy · Jan 24

The hottest new programming language is English

550 3,483 21.8K 2.5M

Andrej Karpathy @karpathy

This tweet went wide, thought I'd post some of the recent supporting articles that inspired it.

1/ GPT-3 paper showed that LLMs perform in-context learning, and can be "programmed" inside the prompt with input:output examples to perform diverse tasks arxiv.org/abs/2005.14165

Figure 1.1: Language model meta-learning. During unsupervised pre-training, a language model develops a broad set of skills and pattern recognition abilities. It then uses these abilities at inference time to rapidly adapt to or recognize the desired task. We use the term "in-context learning" to describe the inner loop of this process, which occurs within the forward-pass upon each sequence. The sequences in this diagram are not intended to be representative of the data a model would see during pre-training, but are intended to show that there are sometimes repeated sub-tasks embedded within a single sequence.

9:56 AM · Feb 19, 2023 · 52.2K Views

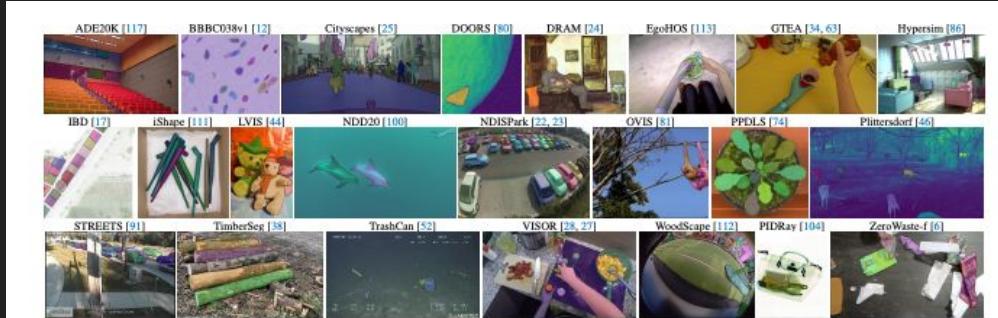
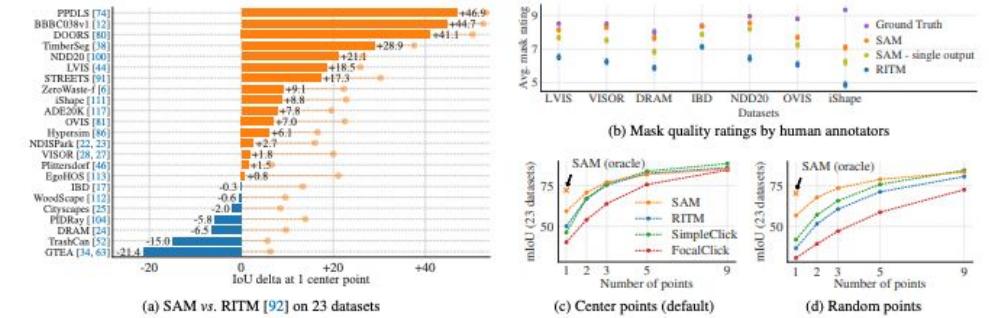


Figure 8: Samples from the 23 diverse segmentation datasets used to evaluate SAM's zero-shot transfer capabilities.



arXiv > cs > arXiv:2005.14165

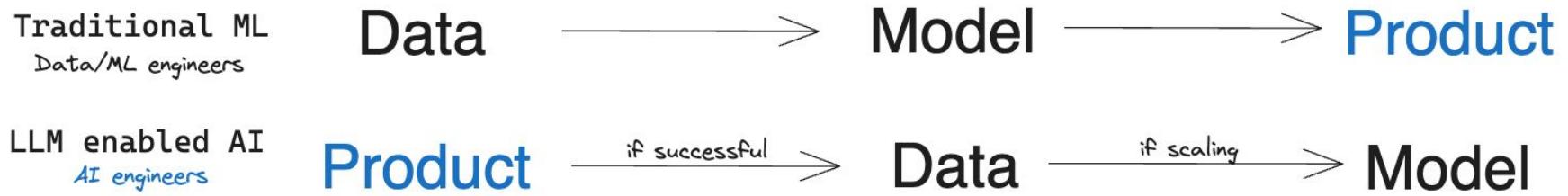
Computer Science > Computation and Language

[Submitted on 28 May 2020 (v1), last revised 22 Jul 2020 (this version, v4)]

Language Models are Few-Shot Learners

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini

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

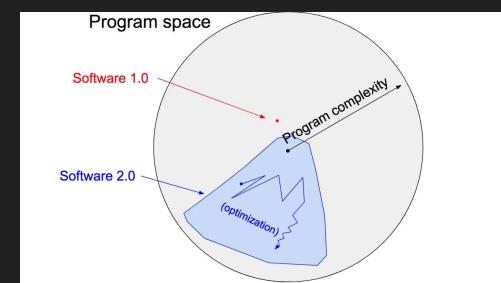
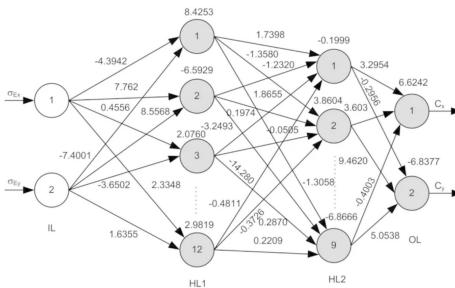
50K

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 💣🔥

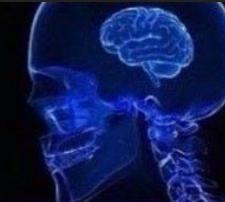
3:14 AM · Jun 19, 2020

**PRE-SOFTWARE:
SPECIAL-PURPOSE
COMPUTER**

**SOFTWARE 1.0:
DESIGN
THE ALGORITHM**

**SOFTWARE 2.0:
DESIGN
THE DATASET**

**SOFTWARE 3.0:
DESIGN
THE PROMPT**



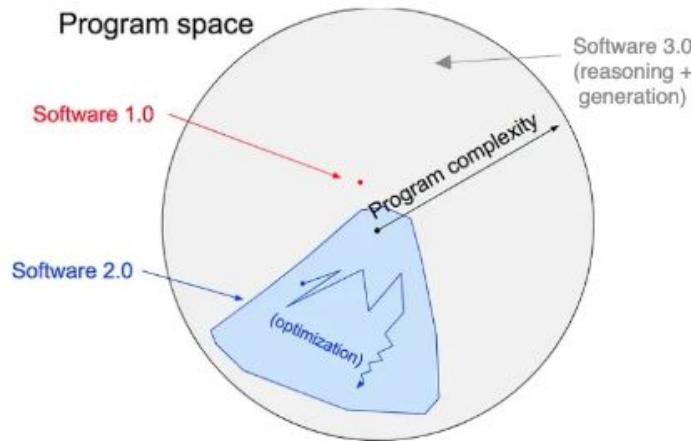
AI + Code >> AI + Language

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

Role of Codegen

$$1 + 2 = 3$$

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

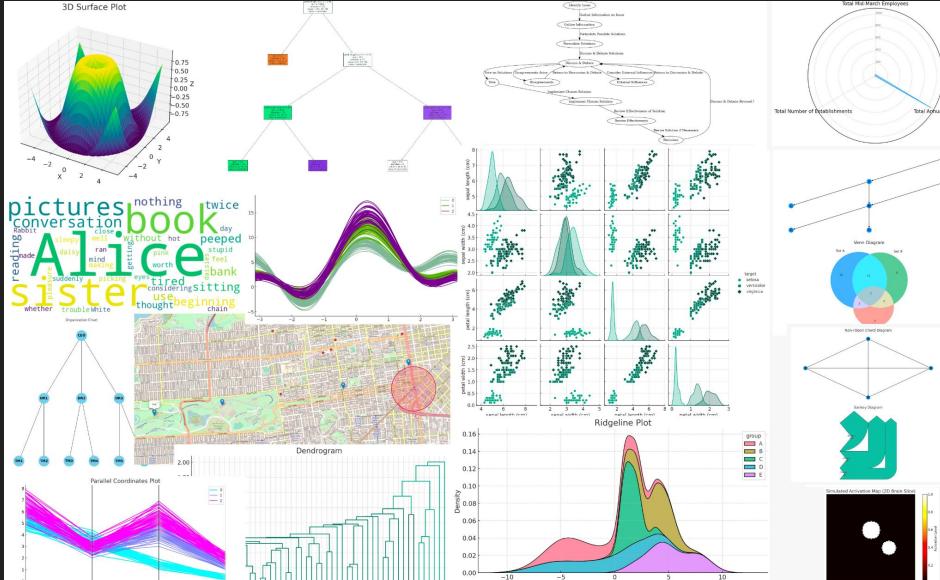
"You are GPT-3, and you can't do math": Prompting GPT-3 via zero-shot instruction to answer calculation/math questions by consulting a Python REPL.

You are GPT-3, and you can't do math.
You can do basic math, and your memorization abilities are impressive, but you don't do any complex calculations that a human could not do in their head. You also have an annoying tendency to just make up highly specific, but wrong, answers.
We hooked you up to a Python 3 kernel, and now you can execute code. If anyone gives you a hard math problem, just use this format and we'll take care of the rest:
Question: \${Question with hard calculation}
python
Code that prints what you need to know

Output
Output of your code)
Answer: \${Answer}
Otherwise, use this simpler format:
Question: \${Question without hard calculation}
Answer: \${Answer}
Margin.
8:55 PM · Sep 9, 2022

347 Retweets 84 Quotes 2,714 Likes 413 Bookmarks

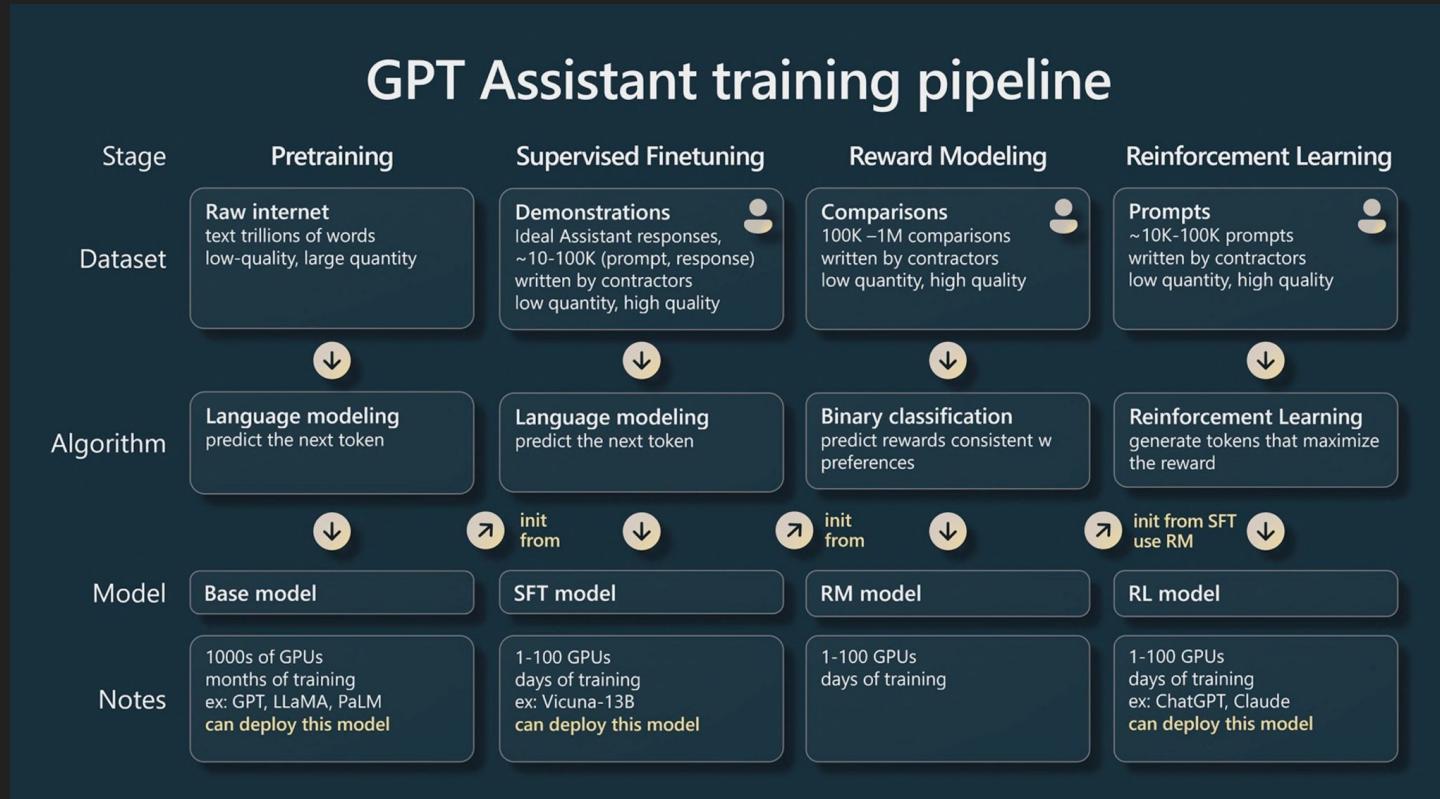
June 2023



GPT4.5: Code Augmented Inference



GPT3->4 training per OpenAI



LLM Core vs Code Core

Code Shell, LLM Core is fundamentally constrained by LLM capabilities

LLM Core
Code Shell

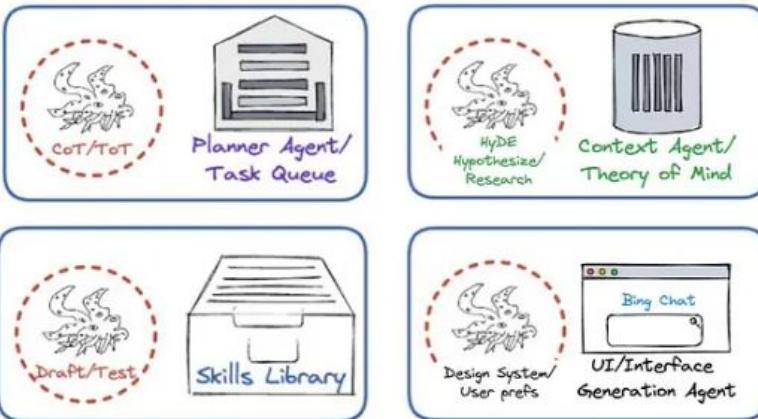
e.g. Retrieval Augmented Generation, Chat,
Backend-GPT, Marvin AI, AutoGPT

Code Shell



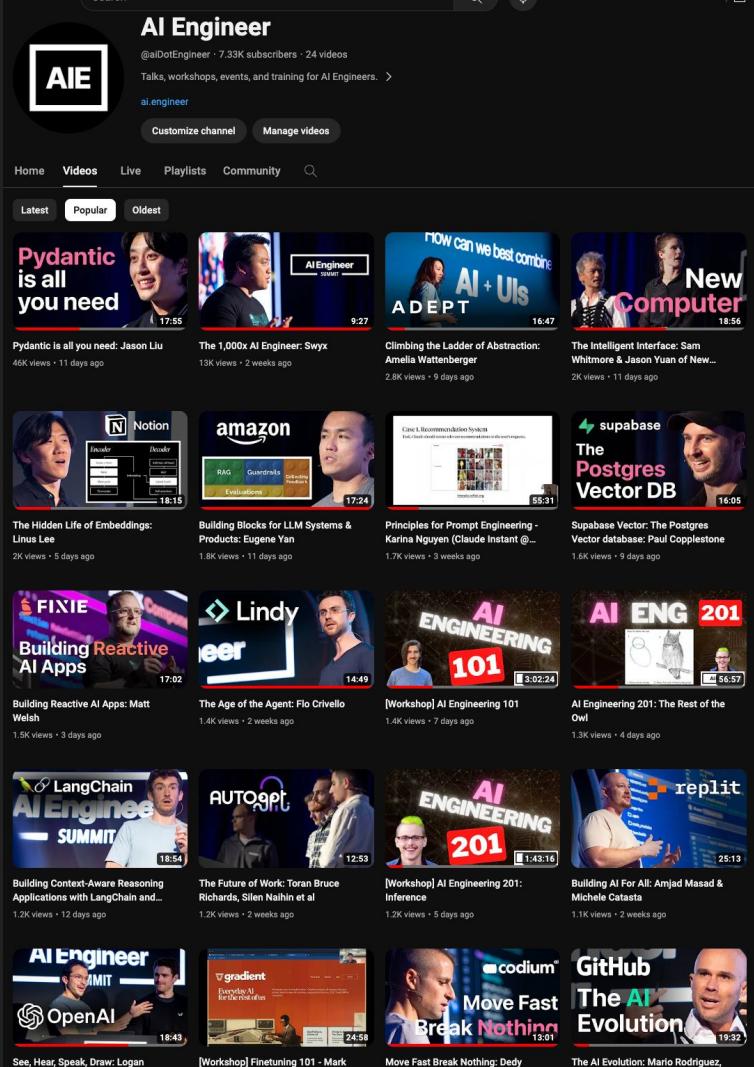
LLM Shell
Code Core

e.g. Copilot, Voyager, Smol-Developer



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

Modern Data Stack Value Capture (2023)

System of
Record

Warehouse

Databricks: \$43b (private)
Snowflake: \$56b (public)

User
Interface

BI

Pipes

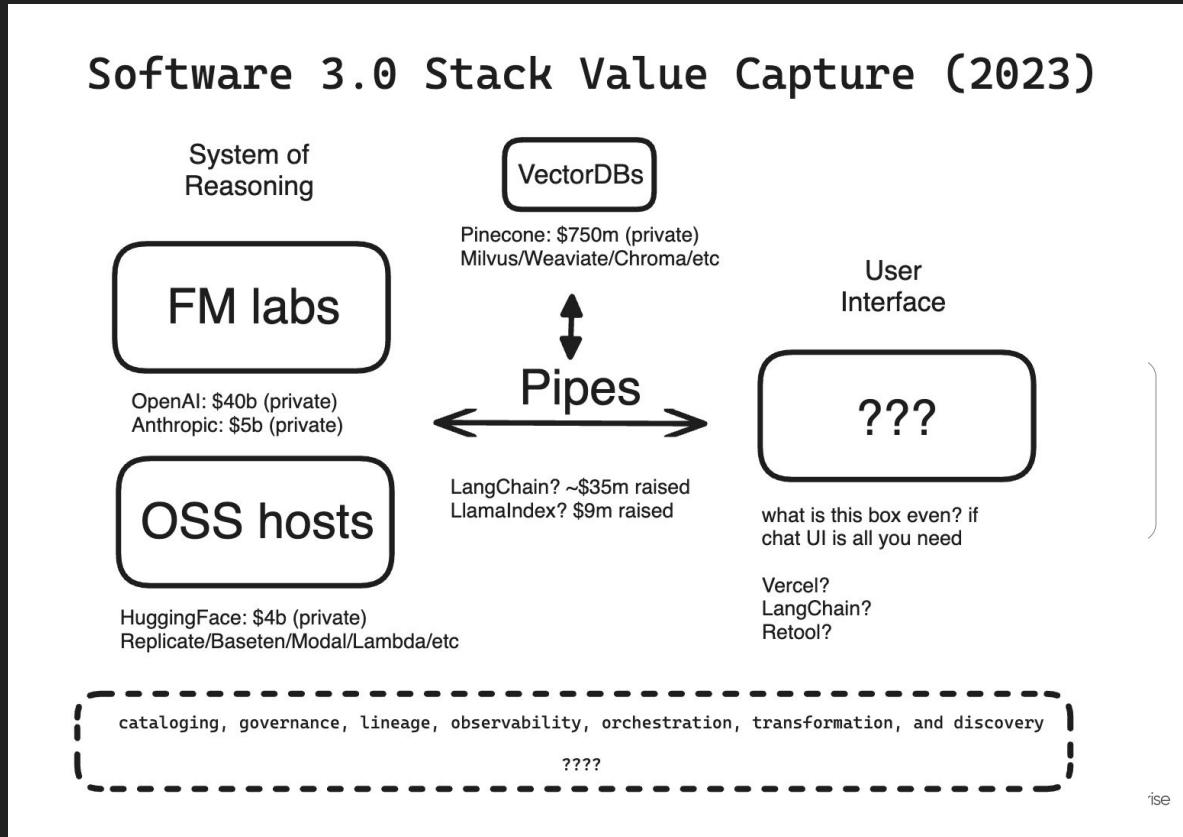
Fivetran: \$6b (private)
Airbyte: \$1.5b (private)

Tableau: \$16b (acq. Salesforce)
Thoughtspot: \$4.2b (private)

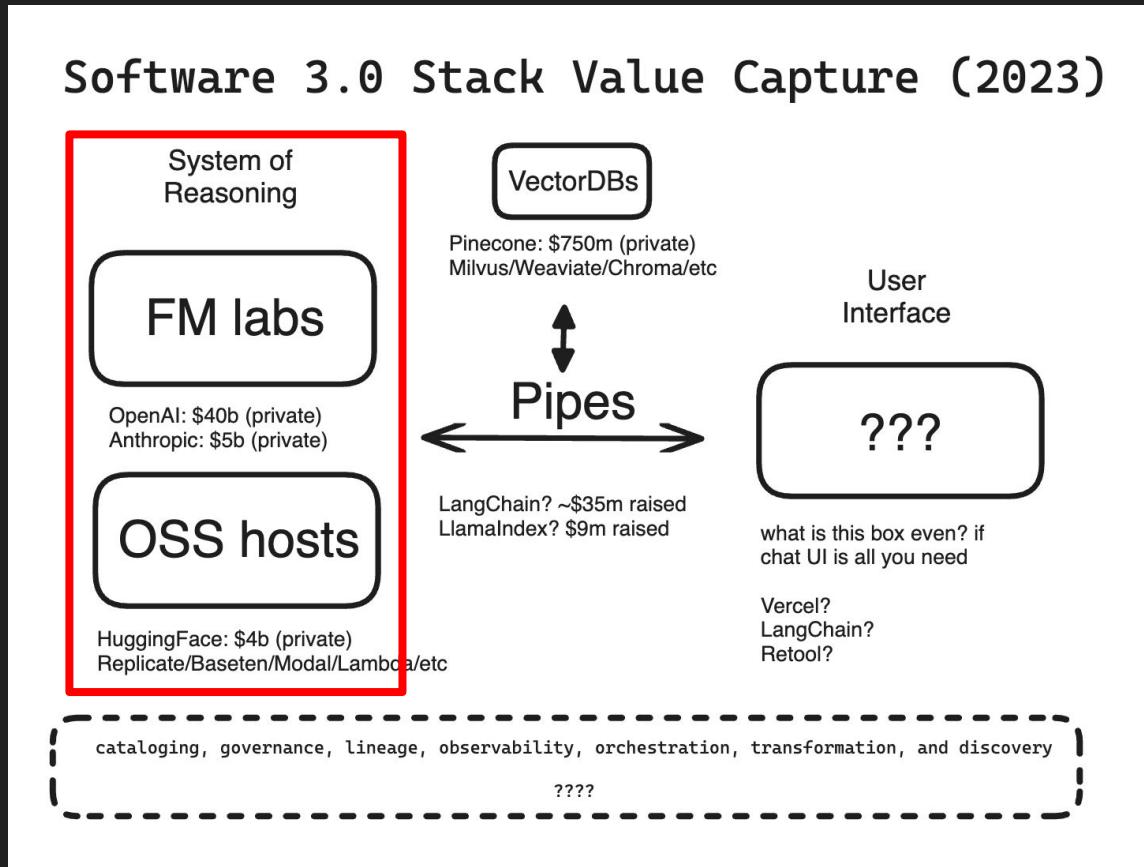
cataloging, governance, lineage, observability, orchestration, transformation, and discovery

total \$4b in value created?

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

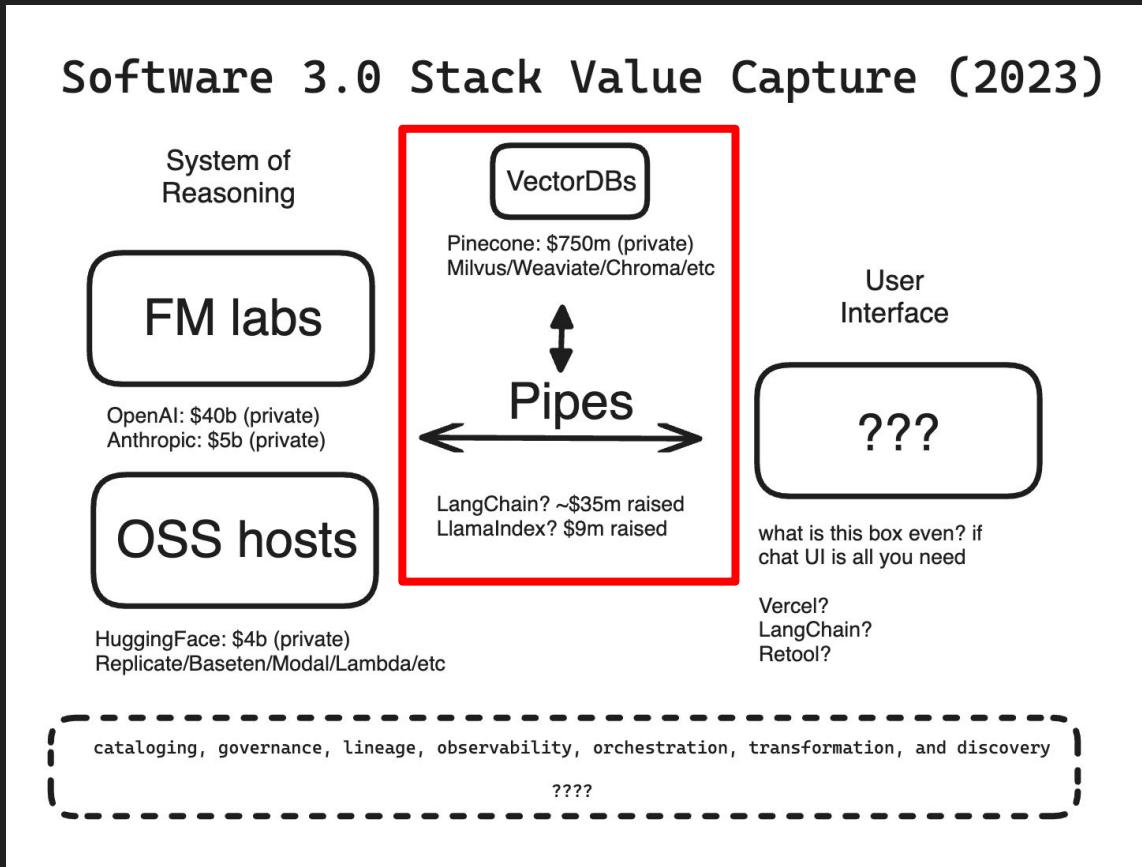
- GCP/AWS/Azure
- Replicate
- Baseten
- OctoML
- Runpod
- Anyscale
- Modal
- Fireworks.ai
- MosaicML
- PyTorch Lightning

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



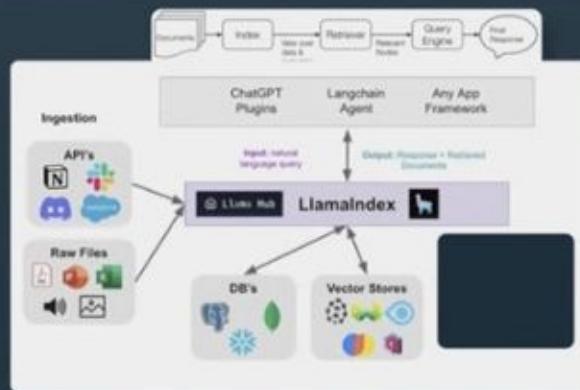
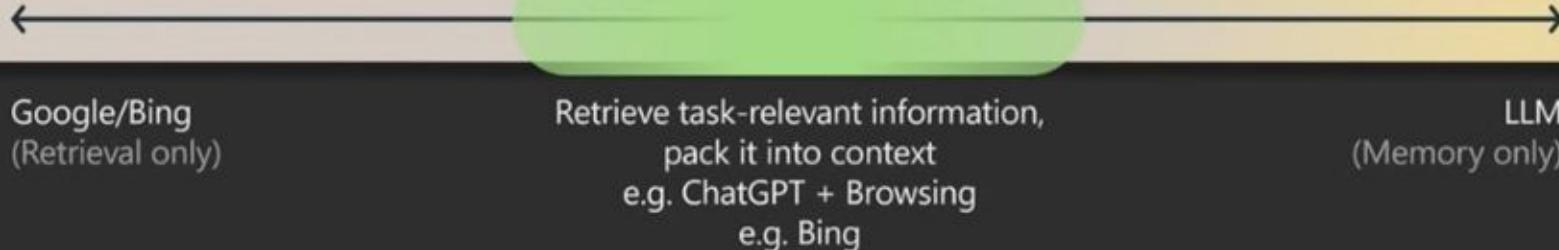
Vector DBs

- 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....)

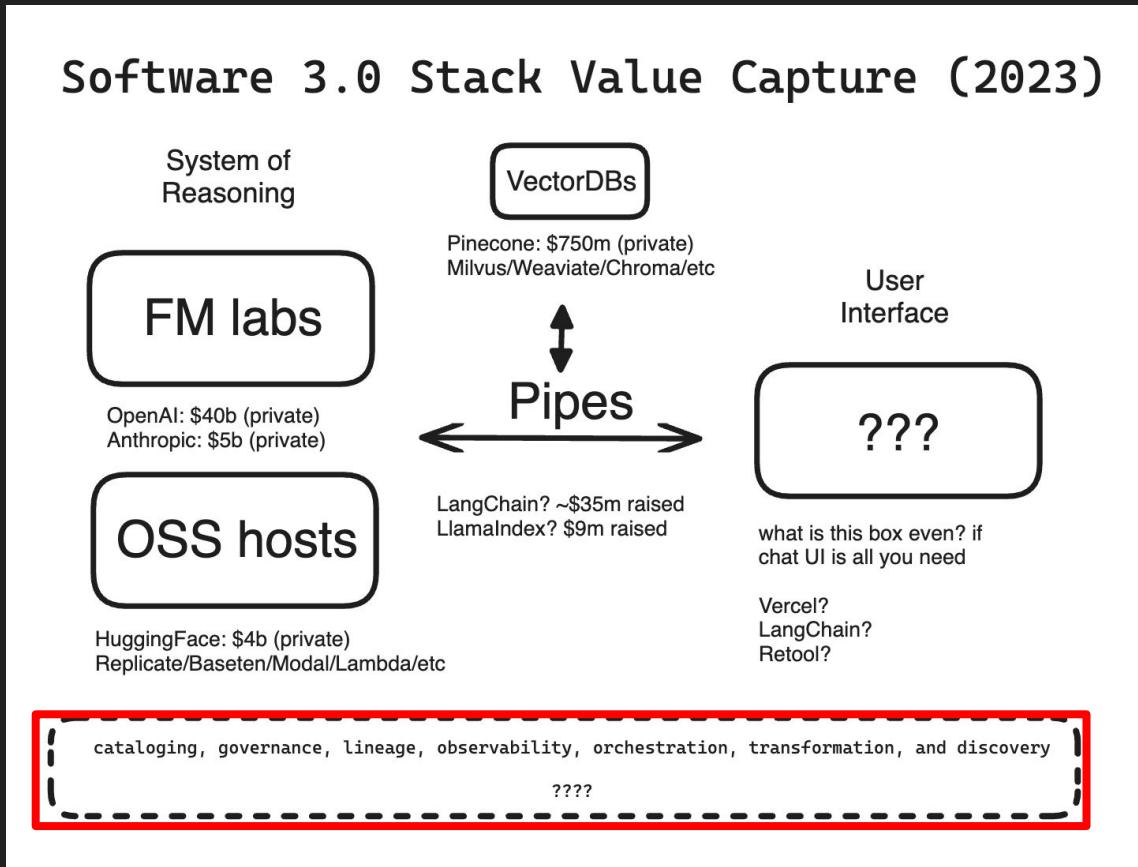
Retrieval Augmented Generation



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 image shows a typical PromptOps/LLMOps landing page interface. On the left, a code editor window displays Python code for generating text using a HumanLoop project:

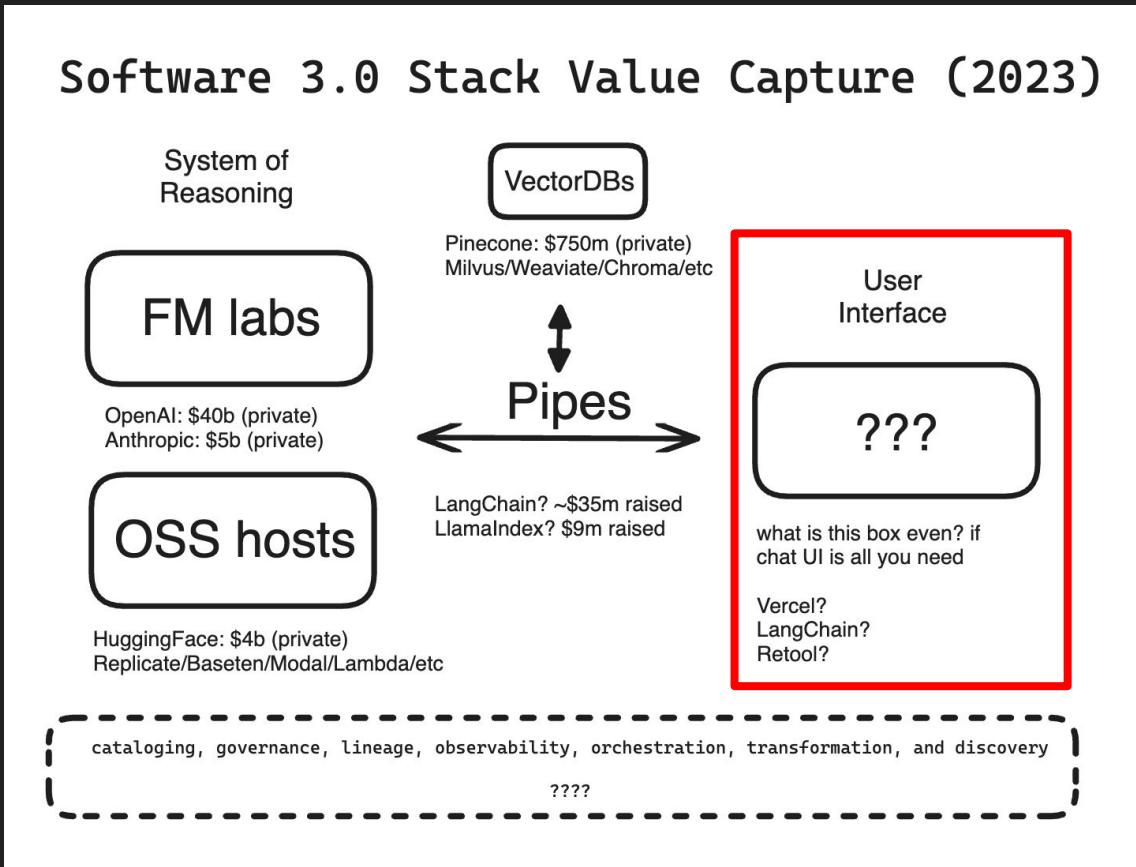
```
import humanloop

generation = hl.generation(
    project="topic",
    model="text-davinci-002",
    prompt_template="What is the capital of {topic}?",
    inputs={"topic": "Paris"}
)
```

The main area features a Project Dashboard with the following components:

- Project Dashboard Header:** Includes a back arrow, a search bar with the text "summarization-tool", a user profile for "Larry", and a dropdown menu.
- Graphs:** A line graph titled "Positive feedback" showing a trend from June 21 to July 20. The y-axis ranges from 0% to 100%. The data shows a dip around June 28 followed by a steady increase.
- Table:** A table listing multiple "Model config" entries. The columns include "Model config" (checkbox), "Positive feedback", "Total feedback", "Cost", and "Date created". Each row has a horizontal progress bar under "Positive feedback".
- Model Config Detail View:** An open modal for "Model Config #2" with sections for "Prompt template" (redacted), "Model" (text-davinci-002), "Temperature" (0.7), "Max. length" (256), and "Positive feedback" (line graph showing a similar trend to the main dashboard).

The Software 3.0 Stack



Vercel v0

A screenshot of a web application interface titled "Acme". The main content area displays a table of financial invoices under the heading "Transactions". The table has columns for Date, Description, Category, and Amount. The data shows transactions from March 12 to March 18, 2024, including purchases from WeWork, IKEA, Home Depot, and Burger King. The "Category" column uses color-coded tags (red, blue, green) to categorize the expenses. A sidebar on the left contains navigation links for Home, Transactions, Accounts, and Tax. A status bar at the bottom indicates "FAQ AI Policy Privacy".

Retool AI

A screenshot of the Retool AI interface. At the top, there are three cards: "getDocuments", "getMetadata", and "Combine". Below them is a code editor window titled "JS Combine" containing the following JavaScript code:

```
1 const docs = {{ getDocuments }};
2 const metadata = {{ getMetadata }};
3
4 return docs.map(d => ({
5   doc: d.data,
6   metadata: metadata.data
7 }))
```

To the right of the code editor is a panel titled "Retool AI" with the following configuration:

- "Summarize text" dropdown menu
- "Summarize each document:" input field containing "{{ combine.data }}"
- "Model" dropdown menu set to "GPT 4"

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

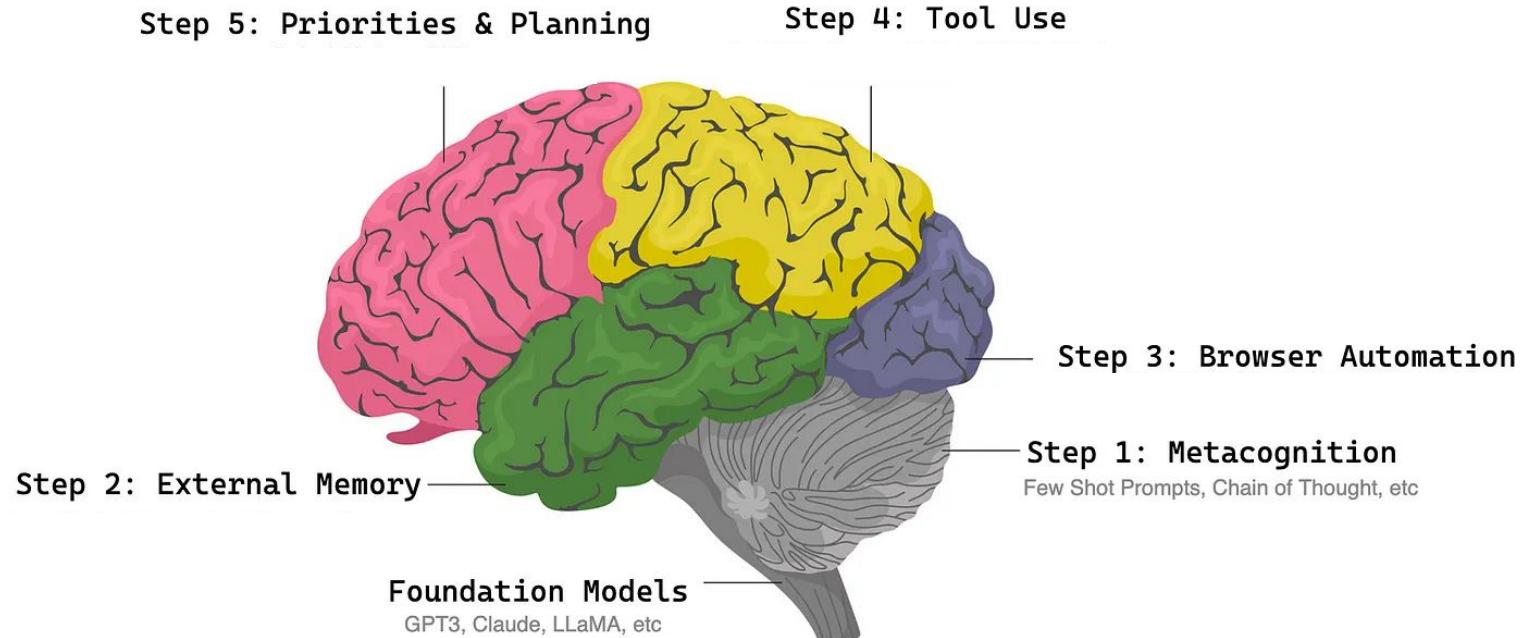
Human driver

AI driver

For on-road vehicles				
	 Human driver	 Automated system		
	Steering and acceleration/deceleration	Monitoring of driving environment	Fallback when automation fails	Automated system is in control
0 NO AUTOMATION <i>Human driver monitors the road</i>				N/A
1 DRIVER ASSISTANCE <i>Human driver monitors the road</i>				SOME DRIVING MODES
2 PARTIAL AUTOMATION <i>Automated driving system monitors the road</i>				SOME DRIVING MODES
3 CONDITIONAL AUTOMATION <i>Automated driving system monitors the road</i>				SOME DRIVING MODES
4 HIGH AUTOMATION <i>Automated driving system monitors the road</i>				SOME DRIVING MODES
5 FULL AUTOMATION <i>Automated driving system monitors the road</i>				

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

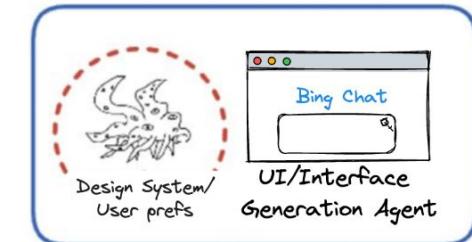
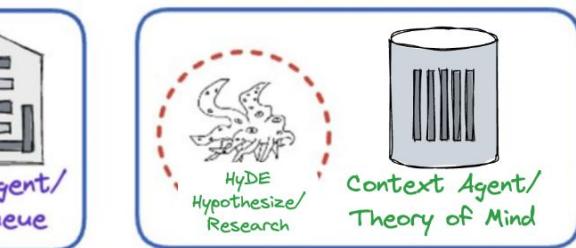
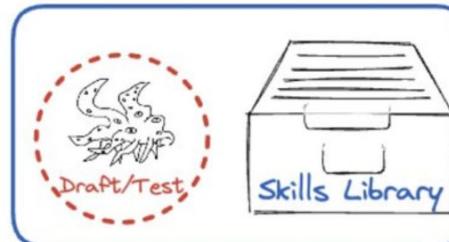
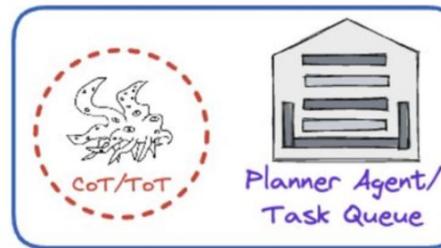
e.g. Retrieval Augmented Generation, Chat,
Backend-GPT, Marvin AI, AutoGPT

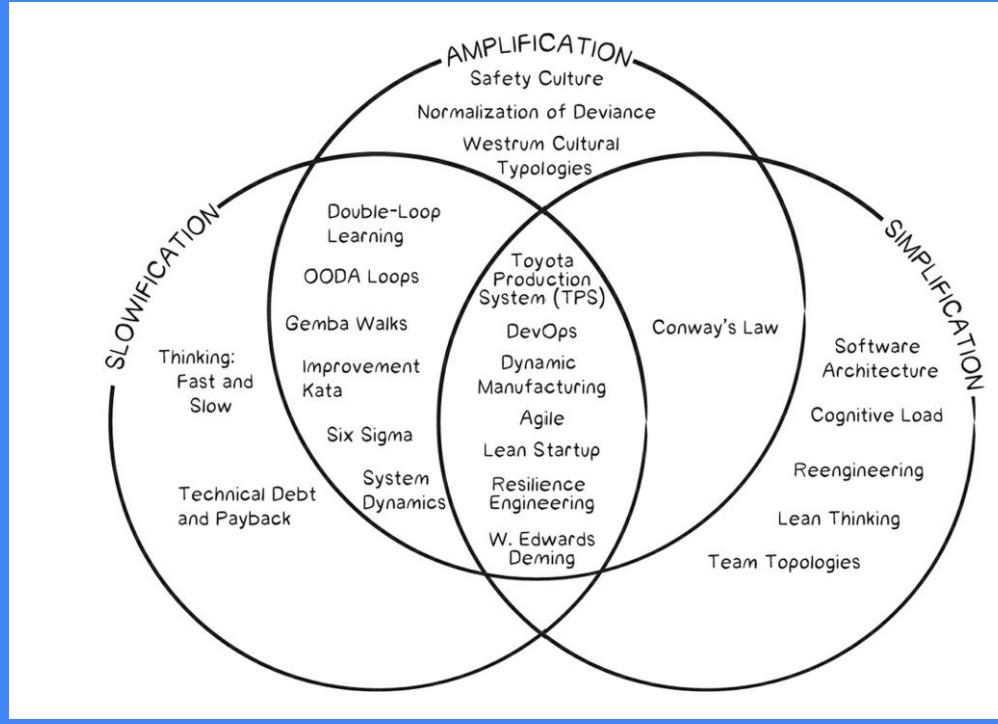
Code Shell



LLM Shell Code Core

e.g. Copilot, Voyager, Smol-Developer





Where to start?

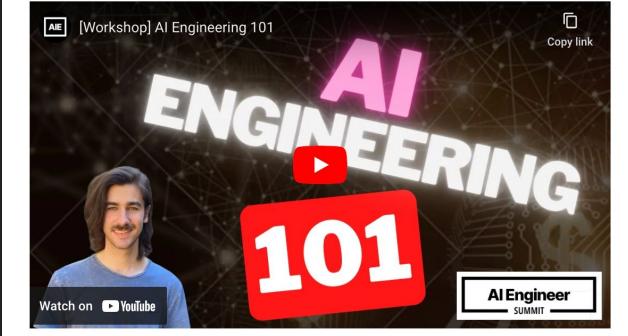
Slowify

<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**

AI Engineering 101

Introductory course for AI Engineers, free for preregistered Summit attendees. Build 5 small projects covering GPT3 API Basics, Prompt Tooling and Memory , Code Generation with GPT4, Image Generation with Dall-E, Stability AI, Lexica, and Midjourney, Speech-to-Text with Whisper. First public run through of the Latent Space University material led by instructor Noah Hein!



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

JUN 14 • 1HR 28M

Emergency Pod: OpenAI's new Functions API, 75% Price Drop, 4x Context Length (w/ Alex Volkov, Simon Willison, Riley Goodside, Joshua Lochner, Stefania Druga, Eric Elliott, Mayo Oshin et al)

1400 Leading AI Engineers from Scale, Microsoft, Pinecone, Huggingface and more convene to discuss the June 2023 OpenAI updates and the emerging Code x LLM paradigms. Plus: Recursive Function Agents!

Jun 14

19 1 1 1 ...

15 30 2x 0:00 -1:28:11 Listen on ▾

Appears in this episode

Alex Volkov Writes ThursdAI - Recaps of the most high signal AI weekly spaces [Subscribe](#)

Riley Goodside

Simon Willison Writes Simon Willison's Newsletter [Subscribe](#)

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

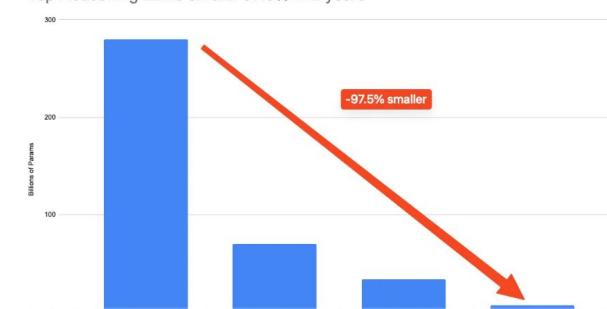
28 2 1 Share ...

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:*

Top Reasoning LLMs shrank 97.5% in 2 years



Amplify

<https://ai.engineer>

Featuring talks from speakers representing leading AI companies & technologies

GitHub



FIXIE

amazon

LangChain



Vercel

Notion

replit

ANTHROP\IC

HEX

A D E P T

Full speaker schedule and talks available now!

The screenshot shows the YouTube channel page for "AI Engineer". The channel has 11.6K subscribers and 33 videos. The main content area displays a grid of video thumbnails, each with a thumbnail image, title, and view count. The titles include "Open Questions for AI Engineering", "Trust, but Verify", "Harnessing the Power of Local LLMs", "The Weekend AI Engineer: Hassan El Mghari", "Domain adaptation and fine-tuning for domain-specific LLMs: Abi Aryan", "Building production-ready RAG apps", "Retrieval Augmented Generation in the Wild", "Pragmatic AI with TypeChat", "Building Reactive AI Apps", "AI Engineering 201: The Rest of the Owl", "Move Fast Break Nothing", "The AI Evolution", "The AI Pivot", "Workshop: AI Engineering 201: Inference", "The Hidden Life of Embeddings: Linus Lee", "Workshop: AI Engineering 101", "supabase: The Postgres Vector DB", "How can we best combine AI + UIs", "Climbing the Ladder of Abstraction: Anelia Watterhauerberger", "The Intelligent Interface: Sam Whitson & Jason Yuan of New Computer", "Pydantic is all you need", "LangChain AI Engine SUMMIT", and "AI Engineer SUMMIT". Each thumbnail includes a timestamp indicating when the video was uploaded.

Thank You!

The screenshot shows the Latent Space website interface. At the top, there's a navigation bar with the logo "Latent Space", a "Dashboard" button, and various icons for search, upload, notifications (with one notification), and user profile. Below the navigation is a horizontal menu with links: Home (which is underlined in blue), Podcast, Discord & Events, Summit (Oct 8-10), AI for Engineers, and About.

The main content area features a photograph of three men sitting on a couch in a studio setting, each wearing headphones and speaking into microphones. To the right of the photo is a diagram titled "HOW LangChain is PRO-FERRED" (in All Over the Internet). The diagram is divided into two columns:

SITUATION: THERE ARE	SITUATION: WE NEED TO DO
LangChain	USE Cases IDEAS WEBSITE APIs WEBSITE APIs WEBSITE APIs
LangChain	LangChain

Below the diagram, there are two article cards:

- Doing it the Hard Way: Making the AI Dream** (with a "The Point of LangChain" sidebar)
- SEP 6 • HARRISON CHASE** (with a "24" likes, "1" comment, and a "..." more options button)

At the bottom of the page is a colorful footer banner with the text "latent.space".

latent.space

KIV

Tech's Two Philosophies



STRATECHERY

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

```
embedding-aud-a0e02 /L data JSON embedding.json
55
56     def task_creation_agent(objective: str, result: Dict,
57         task_description: str, task_list: List[str]):
58         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: '{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."
59         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)
60         new_tasks = response.choices[0].text.strip().split("\n")
61         return [{"task_name": task_name} for task_name in new_tasks]
62
63     def prioritization_agent(this_task_id:int):
64         global task_list
65         task_names = [t["task_name"] for t in task_list]
66         next_task_id = int(this_task_id)+1
67         prompt = f"""You are an task prioritization AI tasked with cleaning the formatting of and reprioritizing the following tasks: {task_names}. Consider the ultimate objective of your
task.

Fork 240 Run 48 Hide code ...
```

 babyagi

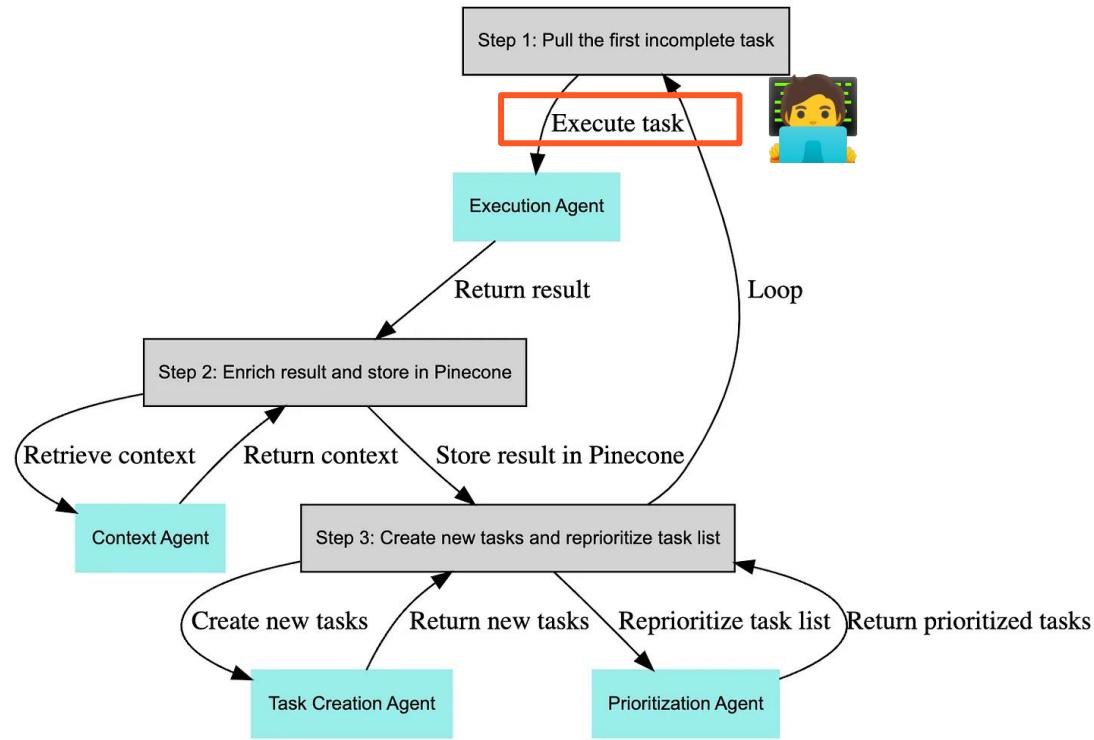
 YoheiNakajima 185 followers + Follow

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/yohseinakajima/babyagi>

poetry.lock



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.

