LOGISTICS

Join the Discord!

https://fsdl.me/llmbc-discord

Slides will be shared.

Talks will be recorded
and posted online.

WiFi: llmbc-2023
password: `SolidGoldMagikarp`

Tensor Tier Sponsors

OpenAI

Weights & Biases

anyscale

GANTRY

aws

V

Matrix Tier Sponsors

Twelve Labs

Bloomberg BETA

Pinecone

YOU.com

Vector Tier Sponsors

BANANA

DISTYL

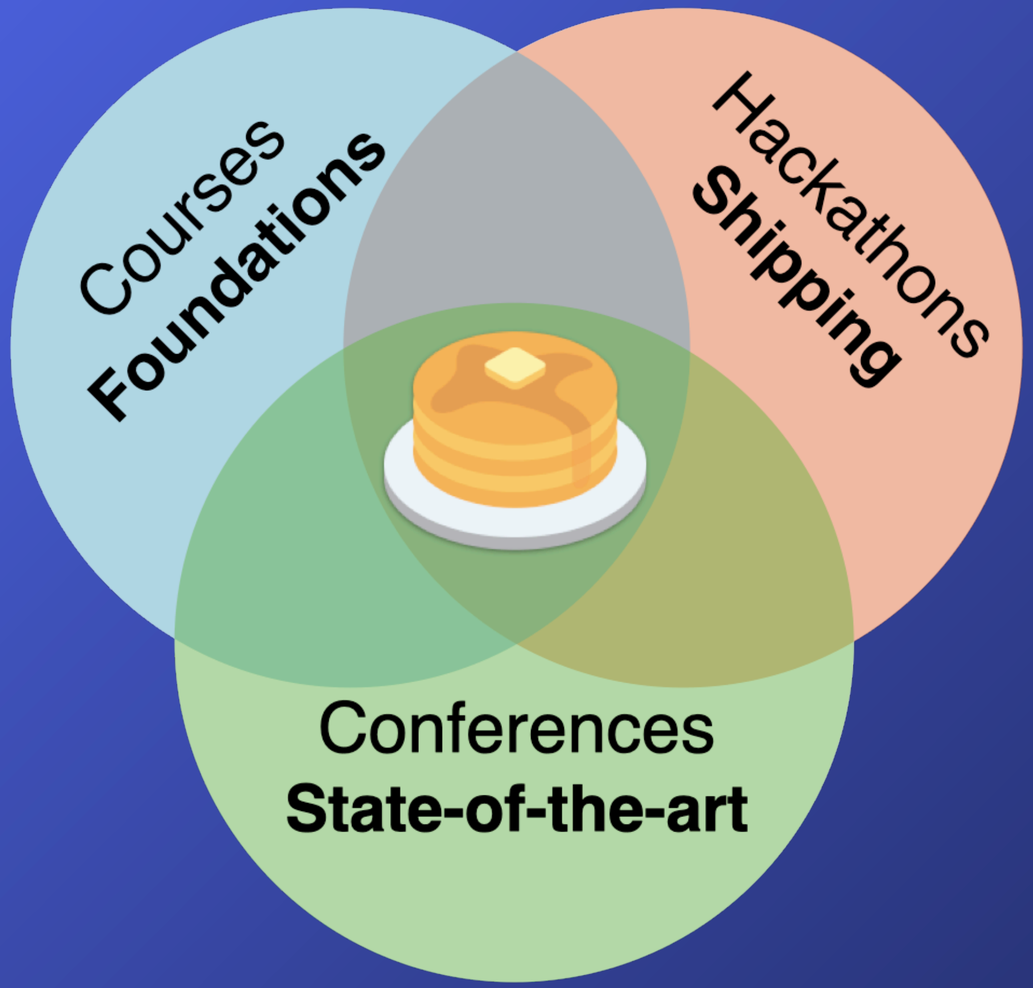
Compute Credit Donors

LLM Bootcamp 2023

Launch an LLM App in an Hour

Charles Frye

April 21, 2023


Courses
Foundations
Hackathons
Shipping
Conferences
State-of-the-art

Instructor Team

@charles_irl

@sergey_karayev

@josh_tobin_

Berkeley PhD alums with years of experience building and teaching about ML-powered apps.

Agenda

00

Introduction

Why are we here?

01

Prototyping & Iteration

Move fast
and write prompts.

02

Deploying an MVP

No one can
stop you.

03

Next Steps

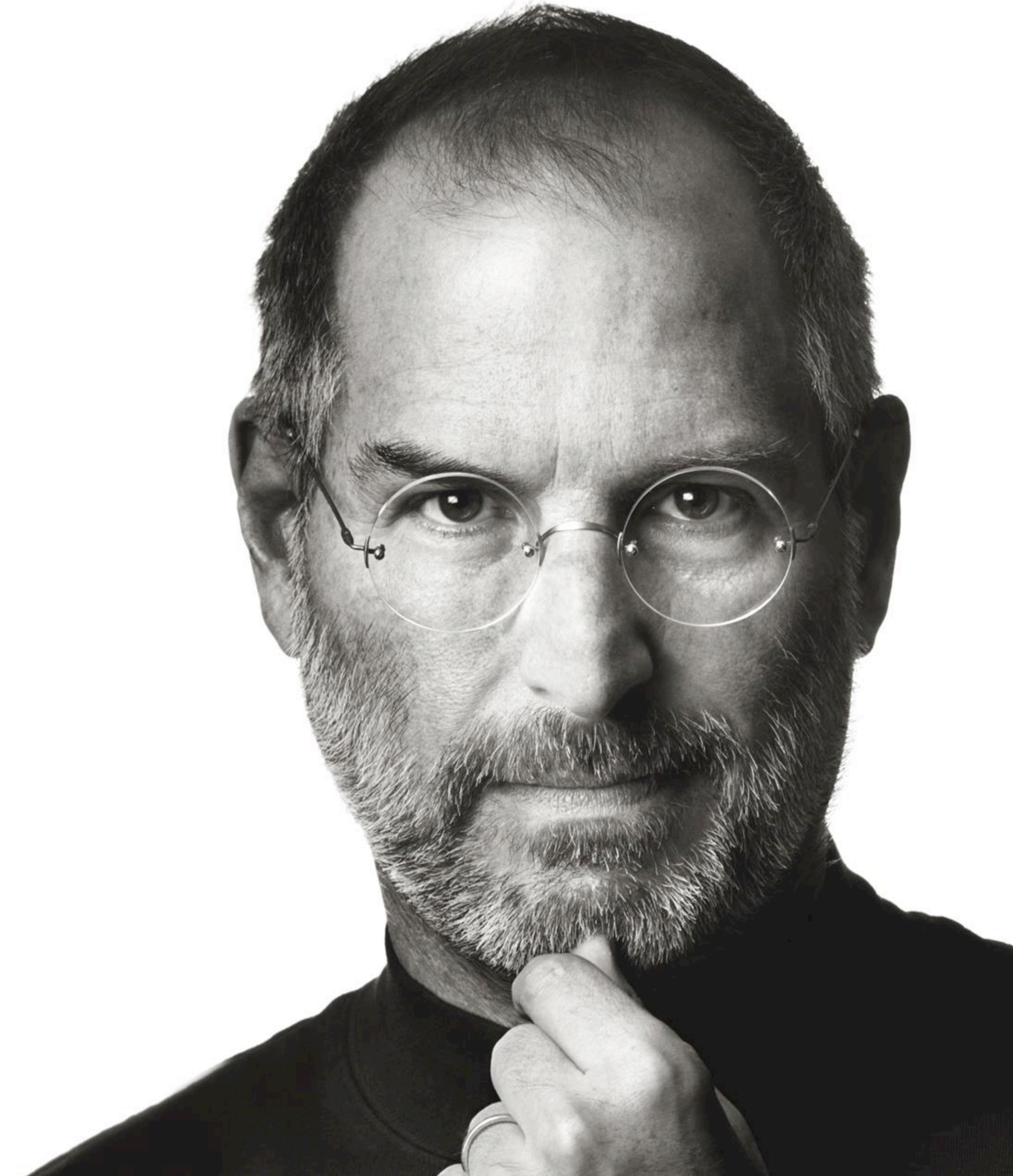
Why are there 10
more hours of talks?

00

Introduction

Imagine a world...

- Disturbingly simple and heuristic computer programs are used to **mimic human cognitive processes**.
- Humans are forming emotional **attachments to chatbot therapists**.
- Computers can **play chess**, **write mathematical proofs**, and even **pass some exams**.
- You can even get **investment advice from an artificial intelligence**!



It happened in 1965.

By using for heuristic programs tasks where there are good measures of human performance, much has been learned about the general level of complexity and sophistication that is required if programs are to operate at levels of effectiveness comparable to human levels. We know now that such programs need not be enormous in size or terribly intricate provided that they incorporate powerful heuristics for searching the problem space selectively. Thus, a recent chess program of modest size matches human abilities in discovering deep checkmating combinations.[4] Programs for discovering proofs for theorems in geometry and for finding indefinite and definite integrals are powerful enough to pass standard school examinations in these subjects.[5] A program of very simple structure, having access to a large body of data, simulates closely the behavior of an investment trust officer selecting portfolios for trust funds.[6]

HEURISTIC PROBLEM SOLVING BY COMPUTER*

HERBERT A. SIMON AND ALLEN NEWELL
Carnegie Institute of Technology
Pittsburgh, Pennsylvania

What’s different this time?

One tool can do it all:
a large language model.

Language Models model language.

Imagine a paragraph appearing word by word.

If you learned to guess the next word,
you'd get pretty smart.

Especially if you could predict things like:

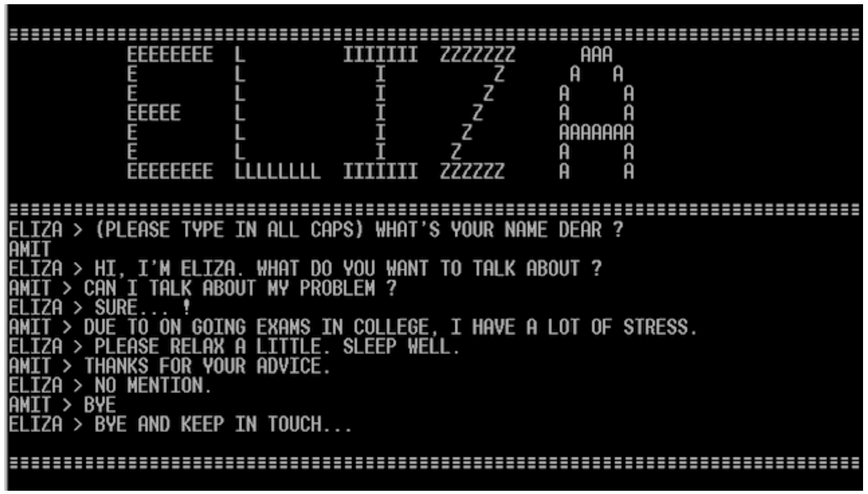
2 + 2 = 4

def square(x): return x * x

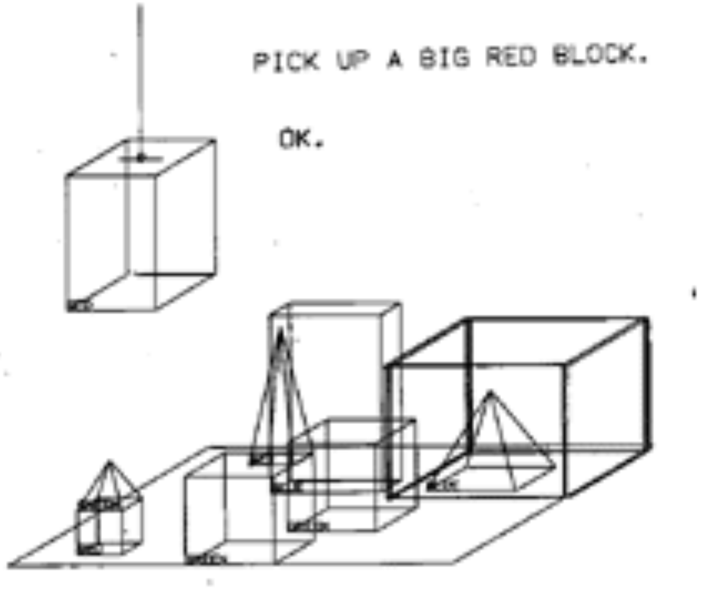
The chemical formula for glucose is $C_6H_{12}O_6$

Large Language Models are really good at it.

LLMs unlock LUIs: Language User Interfaces.



ELIZA, 1966



SHRDLU, 1970

Ask Jeeves, 1996

https://en.wikipedia.org/wiki/Natural-language_user_interface

Language modeling is considered AI-Complete*.

Computational Complexity [19]. According to the Encyclopedia of Artificial Intelligence [40] published in 1992 the following problems are all believed to be AI-Complete and so will constitute primary targets for our effort of proving formal AI-Completeness on them [40]:

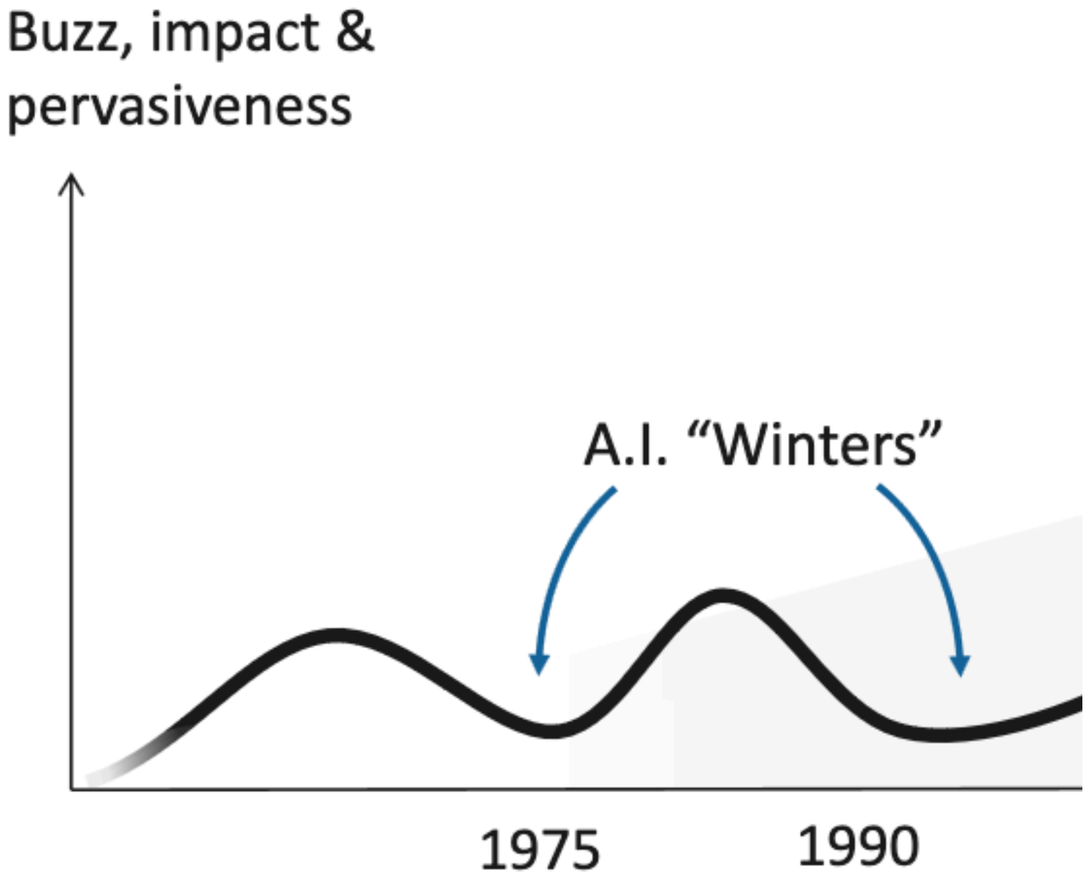
- **Natural Language Understanding**

*this concept is slipperier than NP-Complete etc.

https://www.researchgate.net/publication/251422306_AI-Complete_AI-Hard_or_AI-Easy_Classification_of_Problems_in_Artificial_Intelligence

What happened last time?

Over-selling and under-delivering led to an “AI Winter”.



https://www.cambridgewireless.co.uk/media/uploads/resources/AI%20Group/AIMobility-11.05.17-Cambridge_Consultants-Monty_Barlow.pdf

We failed to deliver products to match high expectations.

"**high hopes** that are **very far from having been realized**"

"[N]otorious disappointments in the area of machine translation, where **enormous sums have been spent with very little result**"

"Suggestions from recent research ... augur badly for the future availability of **machine-translation programs versatile enough to be commercially valuable**."

http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm

To avoid AI Winter, we need to build products that people value.

Good news: a lot more than just research!

https://twitter.com/natfriedman/status/1646285680941862912

Bad news: the gap from demo to product is big.

https://www.youtube.com/watch?v=fmVWLr0X1Sk

Like, REALLY big.

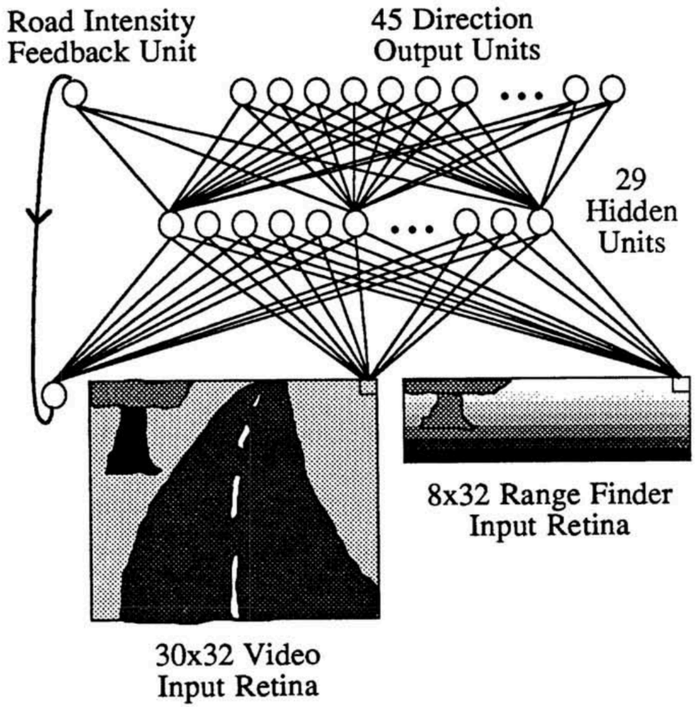


Figure 1: ALVINN Architecture

ALVINN:
AN AUTONOMOUS LAND VEHICLE IN A NEURAL NETWORK

Dean A. Pomerleau
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213

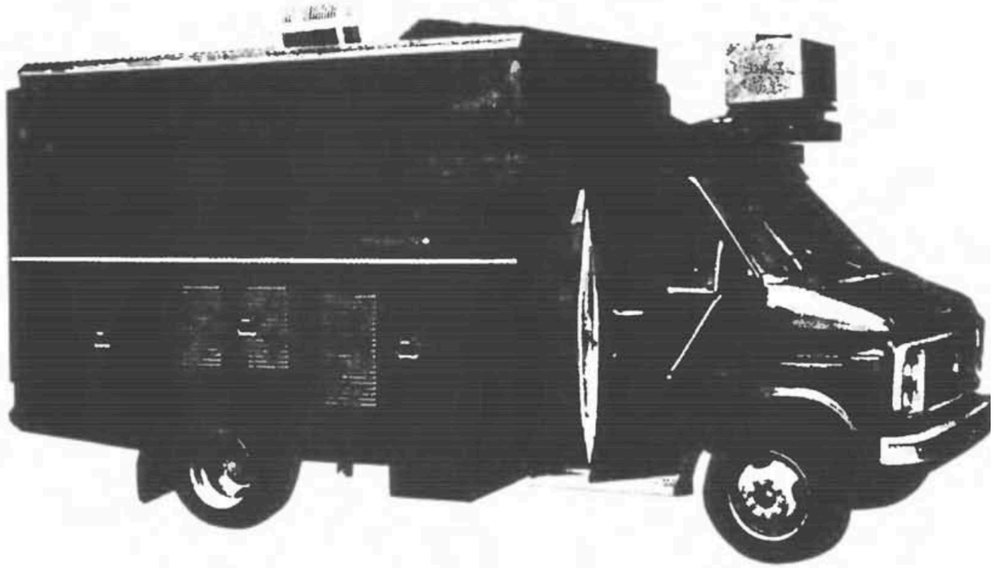


Figure 3: NAVLAB, the CMU autonomous navigation test vehicle.

NeurIPS, 1988

Good news: we're building products (finally!)

2 minute read · February 2, 2023 7:33 AM PST · Last Updated 2 months ago

ChatGPT sets record for fastest-growing user base - analyst note

By Krystal Hu

There's a new way to make video and podcasts. A good way.

Descript is the simple, powerful, and fun way to edit.

GitHub Copilot

https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

So let’s build an app!

01

Prototyping & Iteration: The First Half Hour

Takeaways

- “Foundation models” have unblocked a lot of applications
- Prototype with a high-capability hosted model in chat first
- Tinker with prompts and build on open source frameworks

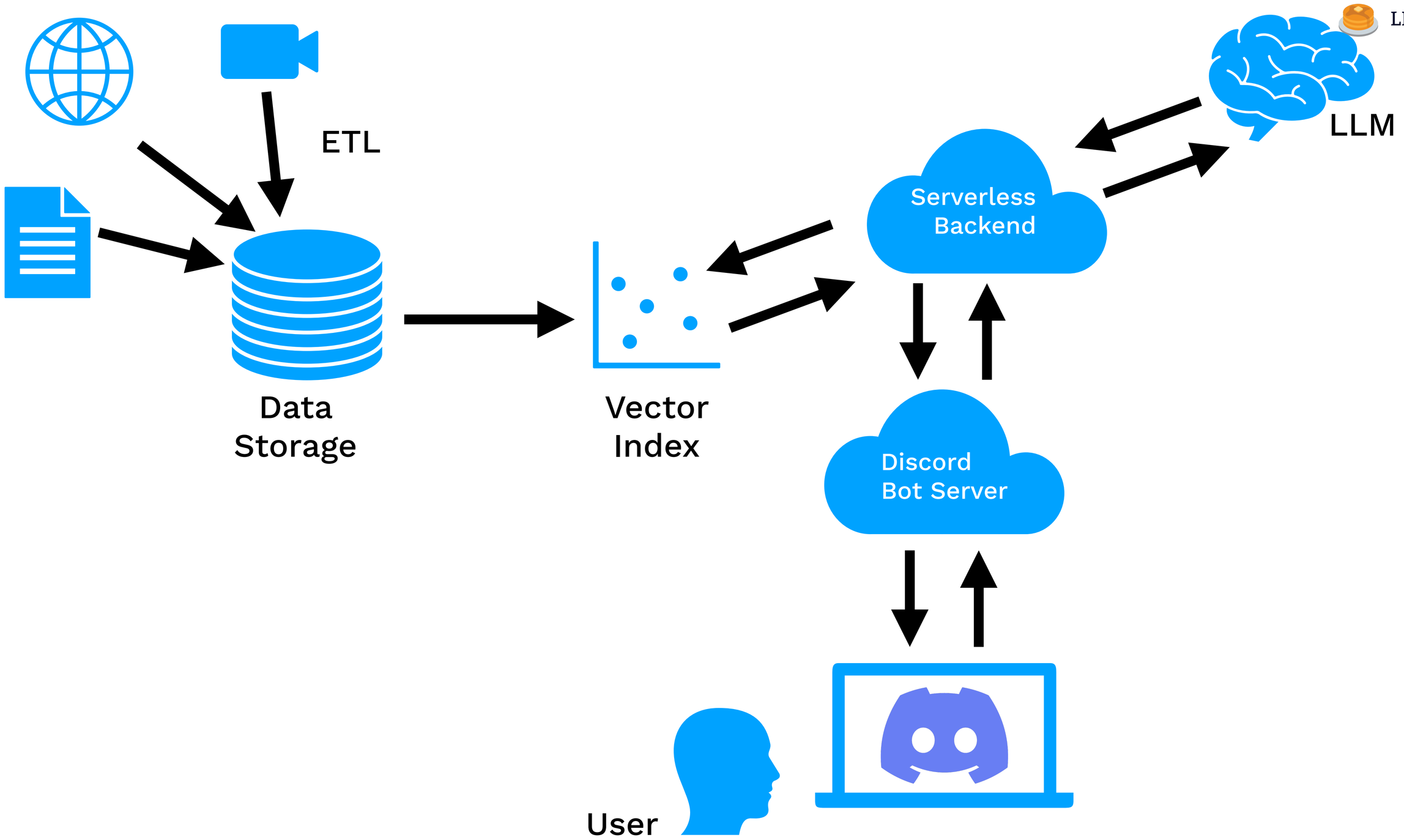
02

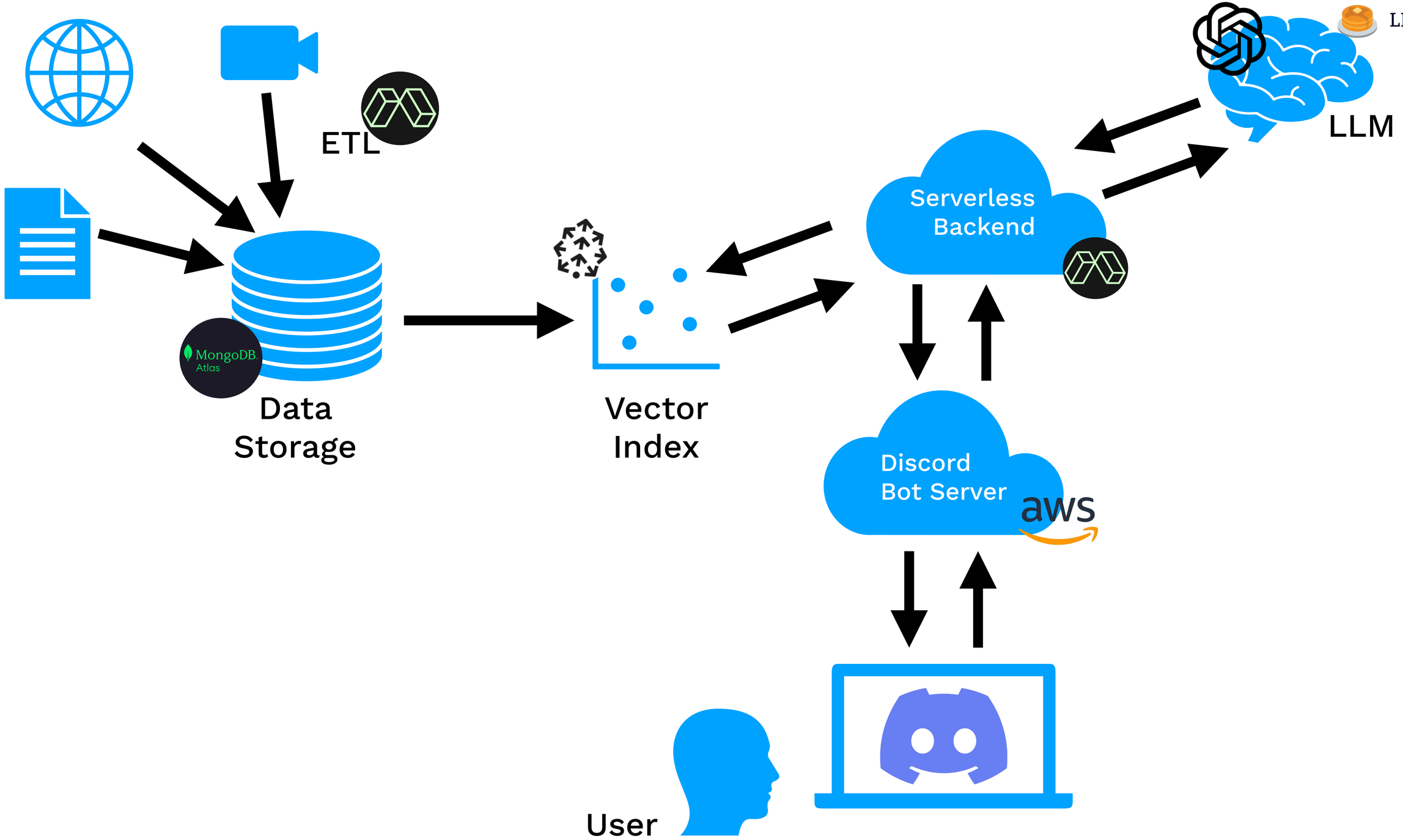
Deploying an MVP: The Second Half Hour

Takeaways

- Cloud tooling makes it easy to get started
  - But make sure to limit spend!
- Find a simple UI and start getting feedback from users ASAP

Let’s build a bot to do sourced Q&A on the FSDL Discord!


ETL
Data
Storage
Vector
Index
Serverless
Backend
LLM
Discord
Bot Server
User


ETL
MongoDB Atlas
Data Storage
Vector Index
Serverless Backend
LLM
Discord Bot Server
aws
User

Try it out!

#ask-fsdl-bot
in the Discord.

We'll analyze its behavior
later today!

03

Next Steps:
The Rest of Time

Doing it is easy,
knowing what you’re doing is hard.

This morning: foundations.

Schedule

	Friday (April 21)
9 am	Registration & Breakfast
10 am	🚀 **Launch an LLM App in 1 Hour** • We'll go from idea to user-ready website in one hour
11 am	🗿 **LLM Foundations** • Speed-run core ideas in ML • Learn the core concepts behind Transformer architecture • Get a tour of notable LLMs, their training data, and abilities
12 pm	Lunch and networking

https://fullstackdeeplearning.com/llm-bootcamp/#schedule

This afternoon: core LM skills.

1 pm	✨ **Learn to Spell: Prompt Engineering and Other Magic** • Prompt techniques for language modeling, instruction, and agent simulation • Arcana and curiosities from the world of prompt wizardry • Heuristics for where LLMs can and cannot be effective
2 pm	🔨 **Augmented Language Models** • Moving beyond pure language modeling with tools and retrieval • Choosing between vector stores/indices (e.g. FAISS, Milvus, Pinecone, Weaviate, Vespa) for long-term memory
3 pm	Coffee and networking

https://fullstackdeeplearning.com/llm-bootcamp/#schedule

Late afternoon: repo and invited talks.

3:45 pm	👷 **askFSDL Walkthrough** • Detailed breakdown of a well-documented sample project demonstrating use of LLM APIs and frameworks, traditional and vector databases, and user feedback ingestion
4:30 pm	🎤 Invited Talk from **Richard Socher**: CEO and co-founder of You.com
5:15 pm	🎤 A Fireside Chat with **Peter Welinder**: VP of Product at OpenAI

https://fullstackdeeplearning.com/llm-bootcamp/#schedule

Tomorrow morning: user experience.

9 am	Breakfast
9:30 - 10:45 am	🌳 **Demo Garden** • Present your cool project • Review other cool projects
11 am	🤦 **UX for LUIs** • Principles of successful UX design • Emerging Language User Interface design patterns • Case studies on the best AI-powered apps today
12 pm	Lunch and networking

https://fullstackdeeplearning.com/llm-bootcamp/#schedule

Tomorrow afternoon: LLMOps and the future.

1 pm	🏎️ **LLMOps: Deployment and Learning in Production** • How vendors like OpenAI, Anthropic, and Cohere compare with each other and with open-weights options like T5, Pythia, and LLaMA • How to manage prompts and programs, monitor models, trace chains, and record feedback
2 pm	🔮 **What's Next?** • What are the limits of scale, large and small? • Has multimodality unlocked general purpose robotics? • Is AGI already here? • Can we make it safe?
3 pm	Coffee and networking

https://fullstackdeeplearning.com/llm-bootcamp/#schedule

The end: invited talks and a panel.

3 pm	Coffee and networking
4 pm	🎤 Invited Talk from **Harrison Chase**: Creator of LangChain
4:45 pm	🎤 Invited Talk from **Reza Shabani**: Training LLMs at repl.it
5:30 pm	Discussion Panel: **Building a Defensible Business**

https://fullstackdeeplearning.com/llm-bootcamp/#schedule

Thanks!

@charles_irl

@full_stack_dl

/imagine green parrot perched on a laptop, coder, #developer, pancake logo, pixar style, big round eyes, dorky glasses, 3d animation, Blender, 4k render, OpenGL - shaders, ray tracing