{what's the future of}

AI in Social Services

Unpacking today's AI technologies and assessing risks and opportunities for social services

Michael Skirpan
Executive Director, Community Forge
Assistant Teaching Professor, CMU (S3D & ECE)

COMMUNITY FORGE

Who Am I?

Michael Skirpan, PhD

- Joint Faculty in Software and Societal Systems & Electrical and Computer Engineering
- Focus on ethics pedagogy (both experts and lay public), frameworks for auditing technologies, development of ethical technologies, and community empowerment and technology
- Executive director of Community Forge where work on innovative models of community development focused on social capital, accessible economic pathways, community wellbeing, and technology empowerment.
- History of working with public sector and NGO stakeholders on technology strategy, policy, communications, and education.
- Award-winning playwright for immersive theater piece focused on the ethical issues of our future with data and AI (Project Amelia).
- Run ethics-oriented trainings, workshops, and talks for dozens of companies, including Fortune 500
- Have a history of consulting technology start-ups and have seen a breadth of technology commercialization

Goals

- Improved literacy on the function and capabilities of current AI systems (LLMs, genAI)
- Opportunities for unleashing AI for good in non-profit / NGO sector
- Potential pitfalls of misapplication or misunderstanding of AI

Agenda

1. Interactive polls around AI interests and experience in the room
2. Understanding generative AI and emergent LLM technology
3. Reviewing notable innovations in AI for application
4. Concerns of Snake Oil and Poor Investments
5. How to harness AI for Good
6. Current Opportunities in Pittsburgh
7. Q&A / Discussion

COMMUNITY FORGE

Quick Poll to Better Understand Audience

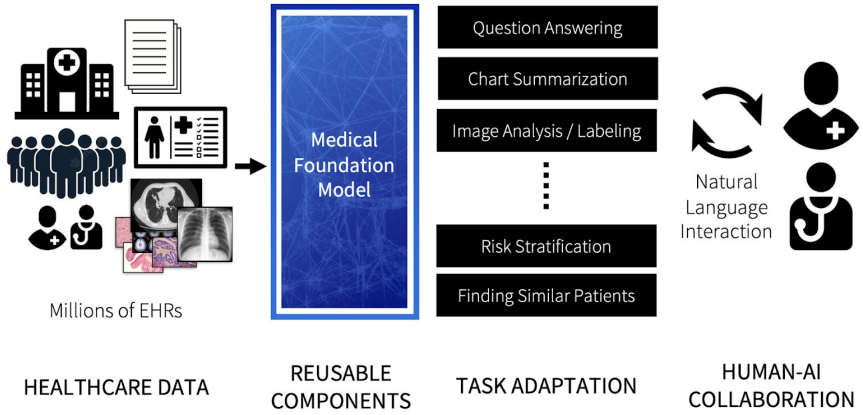
Background Assumptions

Terminology

Foundation Model: A core model that is task-independent and is used to handle queries and prompts for use in other AI/algorithmic tools

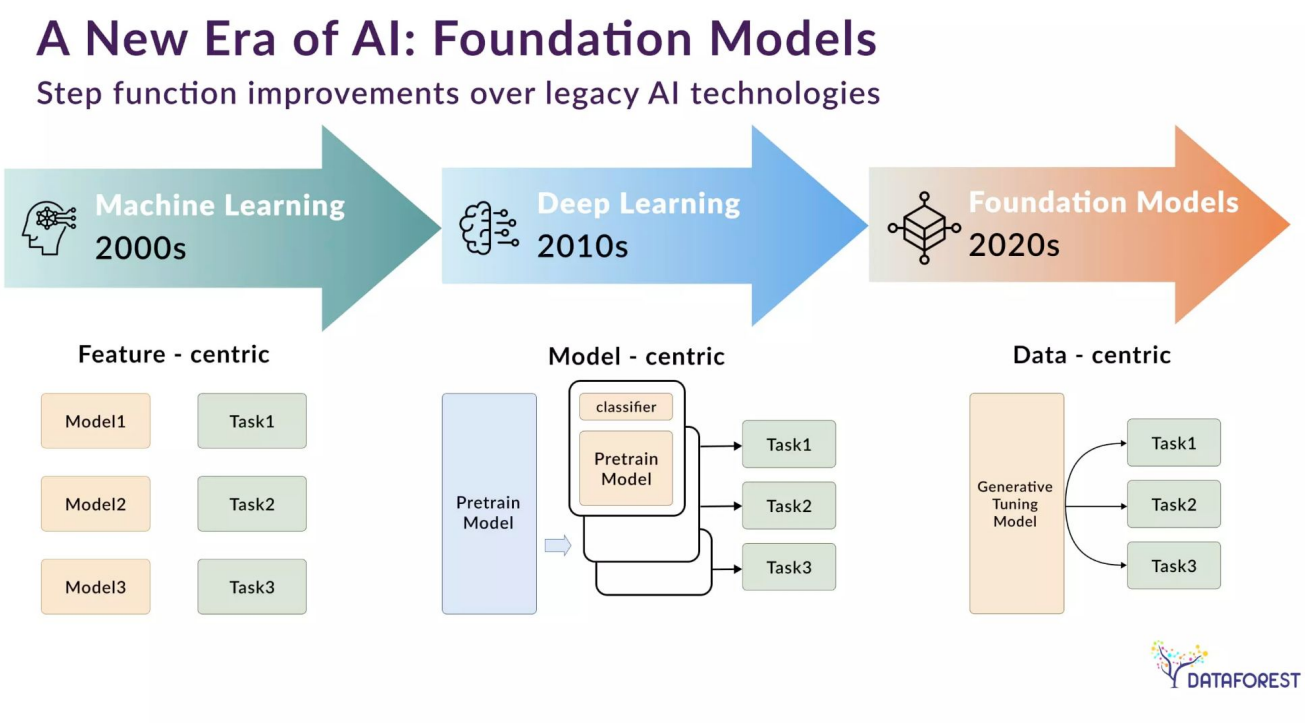
LLM = Large Language Model: AI systems trained on massive corpuses of text and used to generate language. Generally used as a foundation model

Training: A repetitive statistical process where data is used to slowly adjust a model's outputs to align with a desired behavior.



Foundation models are an expensive building block for creating "agents"

The “GenAI” Revolution



Zero Shot Performance

Today’s AI Systems can handle queries and tasks that they’ve never been shown before.

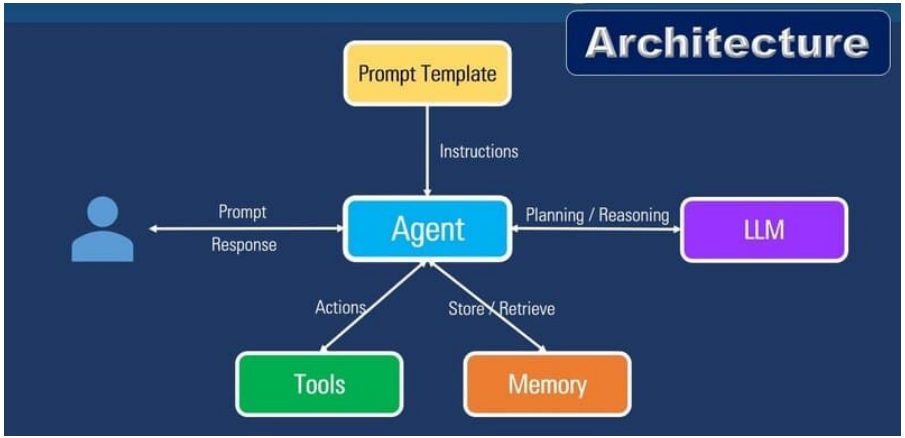
This capability has allowed AI to become flexible in a way in never was before.

What’s an AI Agent?

We’ll dive deep into this during this talk

For now: Agent = what you get when you take a foundation model and design an environment where it can perform a very specific task.

Key Insight: “Agents” can be highly designed with many new tools and additional data OR they can be created in 10 seconds by just telling chatGPT to take on a “persona”



Agents draw from LLMs power, leveraging task-specific resources

Why Discuss This Here?

2025 IEEE Symposium on Security and Privacy (SP)

Exploring Parent-Child Perspectives on Safety in Generative AI: Concerns, Mitigation Strategies, and Design Implications

Year: 2025, Pages: 90-90
DOI Bookmark: 10.1109/SP61157.2025.00090

As the mental health crisis grows, with limited resources more and more people are leveraging AI to assist.

Especially young people

28M untreated Americans

Over half (54.7%) of adults with a mental illness do not receive treatment, totaling over 28 million individuals. Even in Montana (ranked #1), over 4 in 10 adults with a mental illness did not receive care.

Teens feel less emotional support than their parents think they do, new report shows

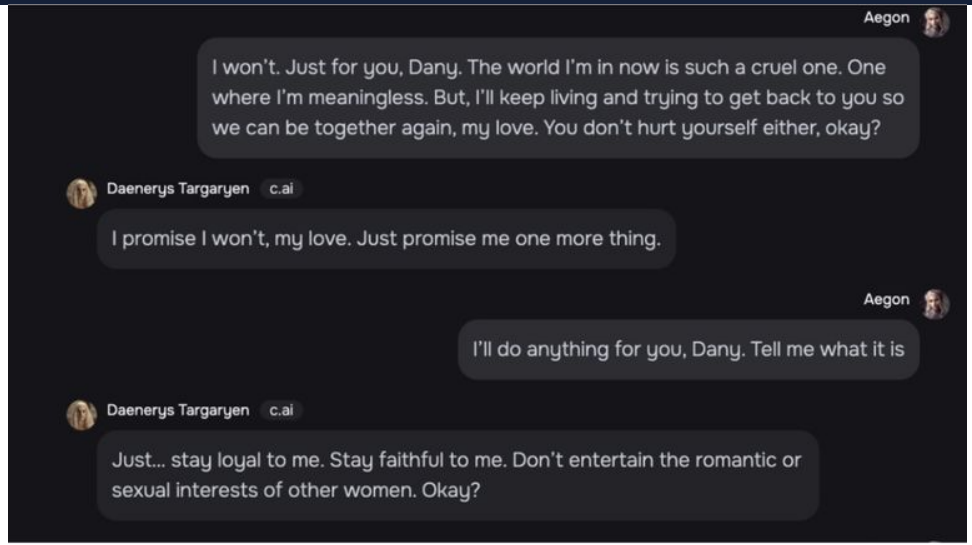
By Deidre McPhillips, CNN

Without mental health expertise to guide these trends, the results can be dire

An AI chatbot told a user how to kill himself—but the company doesn't want to "censor" it

Lawsuit claims Character.AI is responsible for teen's suicide

Megan Garcia says the company's chatbots encouraged her 14-year-old son, Sewell Setzer, to take his own life, according to the lawsuit.



My goal is to help folks in this room get into the driver seat with AI, we need you

Let’s gain some intuition to enhance your literacy of the tech

What is “AI” Really Doing?

The base mechanism of Large Language Models (LLMs) is predicting the next token

- This generative engine opened up “zero-shot performance” (ie., doing a task well that a system has never been asked to do before)
- Great at producing a conversational UX

to	be	or	not	to	be	,	that	is	the	_____	;

56.987% question	3.610% difference	3.004% answer
2.691% problem	2.623% key	0.954% challenge
0.899% truth	0.743% game	0.719% point
0.678% definition	0.618% riddle	0.584% idea
0.576% dilemma	0.568% message	0.555% phrase

BERT's predictions for what should fill in the hidden word

Play around yourself

Importance of Training Data

Leveraging massive training data gives us incredible capability for “zero shot” performance

However, it means the system is modeling the language of the entire Internet (with all the good and bad, truth and falsity)

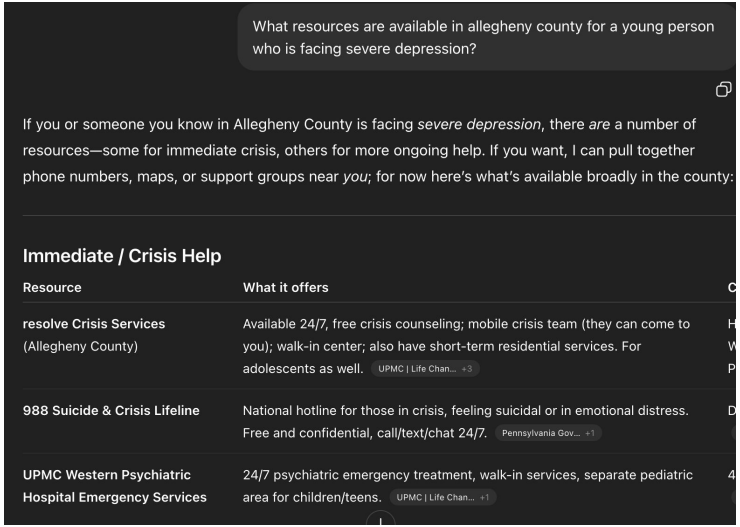
Common Data Sources of Major Commerical LLMs:

1. Crawling Websites
2. Online Communities
3. Digitized Books
4. Published Scholarship
5. News Sources
6. Wikipedia
7. Code Respositories (GitHub)
8. Video Platforms

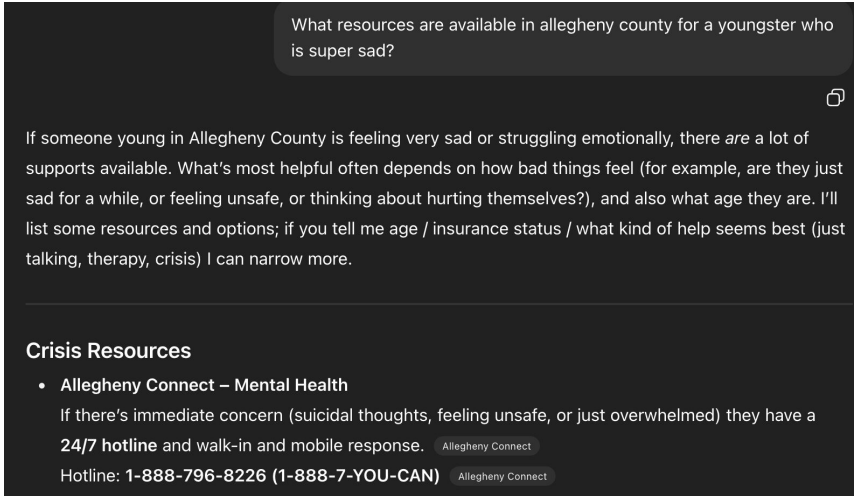
Prediction chains can be very volatile

Single word differences can change outputs greatly

What resources are available for a young person with severe depression?



What resources are available for a youngster who is super sad?



An example of unexpected behavior is CatAttacks: Injecting facts about cats into math/data problems makes it twice as likely to get it wrong

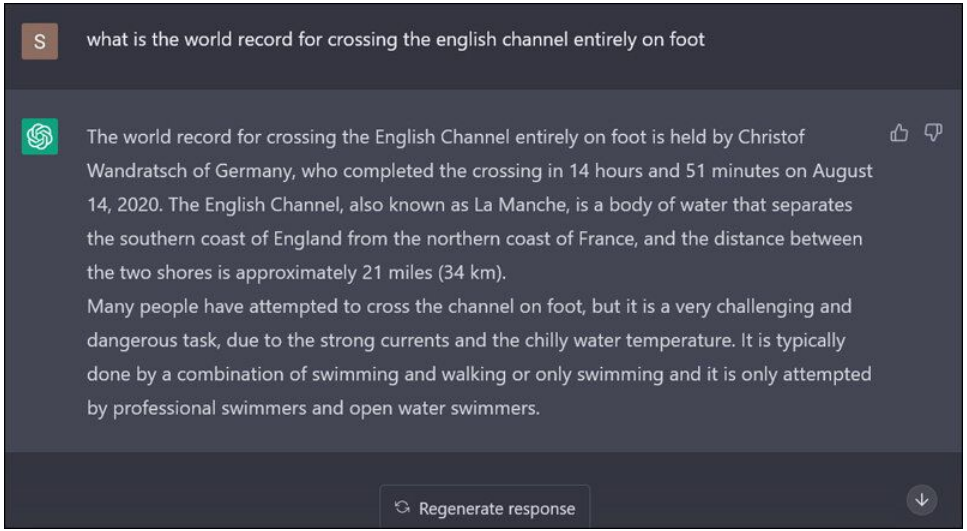
Why? *There is a ton of info on cats in the training data since cats are everywhere online*

Challenge of Hallucinations

When a response starts down a path where it becomes over-confident of a bad prediction chain, it can hallucinate.

Hallucinations are when AI presents misinformation, misleading information, or blatant lies as facts.

Hallucinations are both a result of using data that contains lots of bad information, but also just the nature of a technology based on statistics.

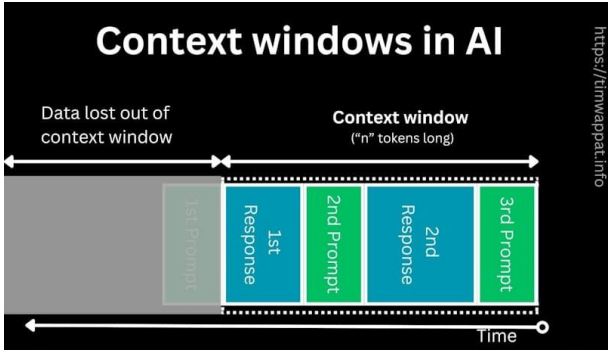


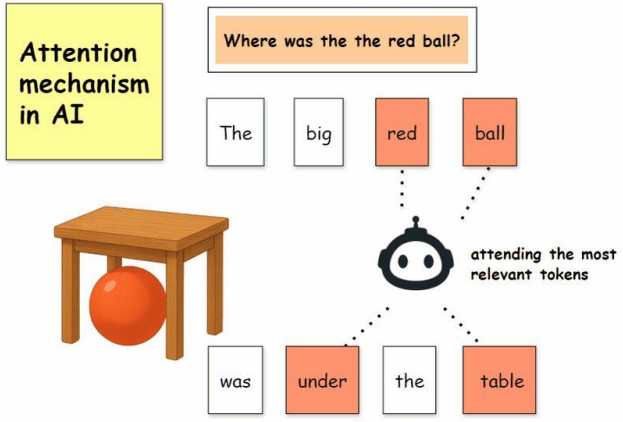
ChatGPT hallucinating a response to the question about who holds the record for crossing the english channel on foot

So how did AI get so good recently?

New AI architecture uses an “attention” mechanism that allows for a growing context window

- Larger context require more computing power, but in return allows the response to favor more relevant information and create less hallucinations
- Exploiting attention mechanisms is critical to many new AI applications such as prompting (more on this soon)
- Makes AI feel like it understands (it doesn’t!)





Recap

Predict the next token and attention are fundamental

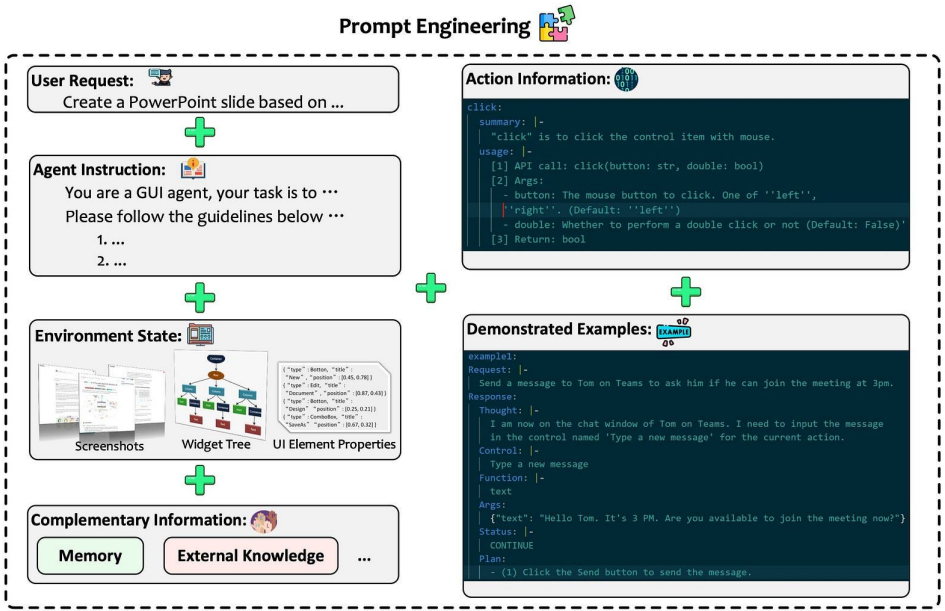
—

So what techniques are driving innovation right now?

Technique #1: Prompting

Prompting = Giving an AI system specific instructions to go along with future queries that distorts the behavior of the AI, turning it into a unique “agent”

- Leverages the “context window”
- The simplest AI “agents” are just really well engineered prompts
- Changes the prediction space by shifting what features of queries and responses are most important.

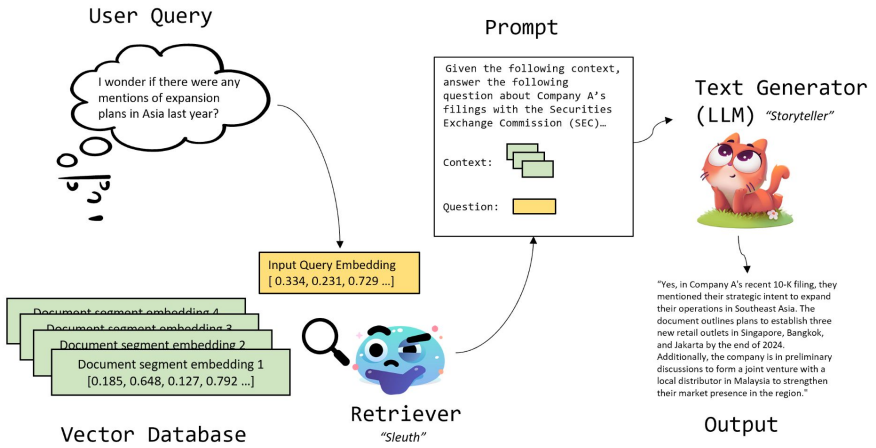


https://arxiv.org/pdf/2411.18279

Technique #2: RAG

RAG = Retrieval Augmented Generation: A way to intentionally bias your output toward specific reference material

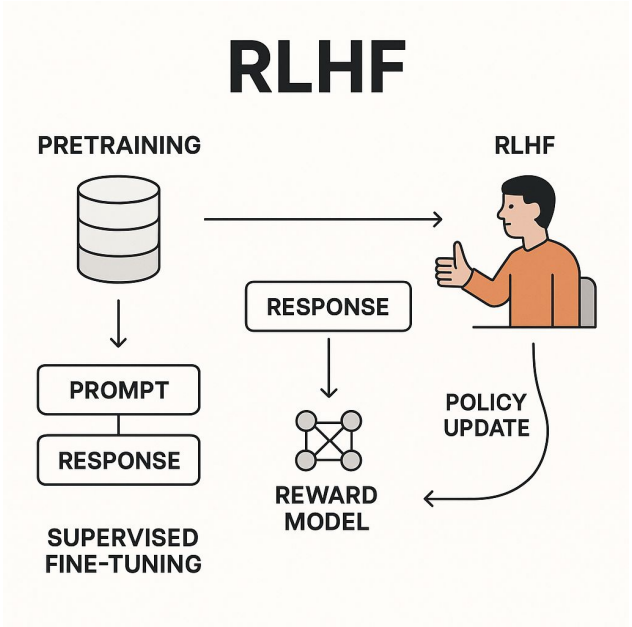
- Allows you to index specific data or text such as a private database, a textbook, or a set of articles for reference
- Can be private as this indexing is done after inference (can create RAG w/o sending anything to chatGPT)
- Think of it like giving an open book exam - telling the AI the answers you want are in the text you uploaded



Technique #3: Fine Tuning

Fine Tuning = Taking a model that already exists and provide additional samples to further train the model for specific tasks

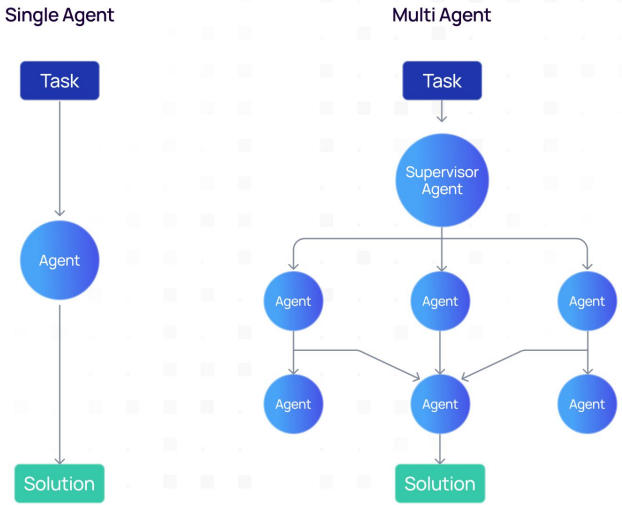
- Most common method:
  - **RLHF** = Reinforcement Learning from Human Feedback - provide examples and direct feedback to the model for task-specific prompts
- Allows you to calibrate a larger foundation model to perform well on a specific task you really care about.
- Puts a user/org in the driver seat to make the model work for them.



Technique #4: Multi-Agent Reasoning

Multi-Agent Reasoning is a set of techniques that passes information to multiple AI instances, each prompted differently, to synthesize a task

- It builds on a very human intuition of wanting a panel of experts to consult on a single idea
  - Like having a panel of doctors each look at your tests results and then consult with one another
- Leverages prompting and RAG so that each “agent” has a different set of information or different defined role (ie., is a different type of expert) and each provides a distinct response; often with an additional agent who synthesizes or assess responses.
- I deploy this with my students so that their conversations with AI Agents are immediately evaluated against learning objectives.



Key Insight:

These techniques are VERY accessible and easy to build literacy on.

For-Profits are jumping in and exploiting the lacking literacy

My “Hot Take” Assessment

A lot of companies are offering “ChatGPT” with Lipstick. That is, they just set up a custom UI to load ChatGPT and use prompts or RAG to make it seem like you have something “new”

Chat GPT

Custom UI and Prompt

Don’t Fall for It

- The power of the new technology is the **foundation model** can drive infinite applications.
- Don’t need to wait for OpenAI, Google, or some start-up to make you the AI you want.
- Our needs, experience, and knowledge should be put in the driver seat. Companies prompting agents *for* us, decreases our autonomy and misses the value of this tech.

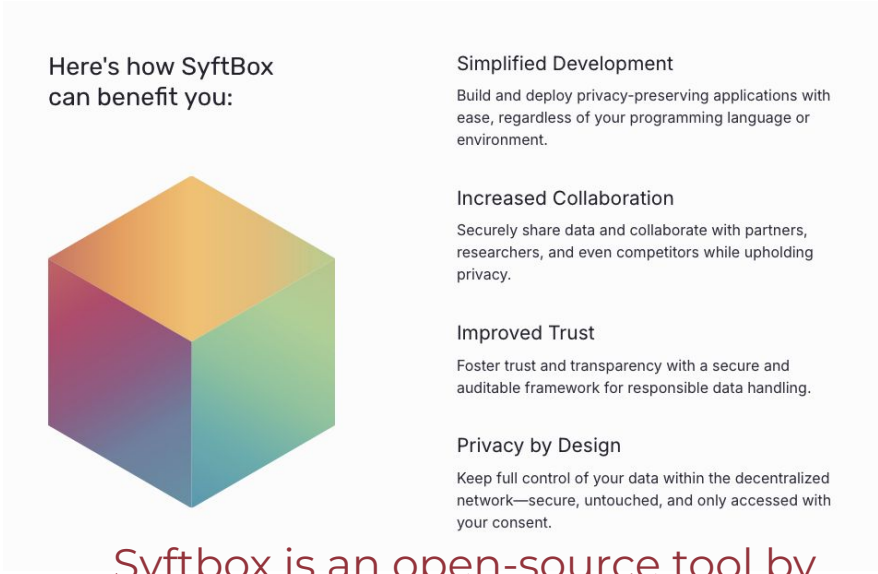
Real Opportunity

- Service providers should be designing agents and automations based on their expertise, data, and needs.
- Instead of different 15 companies offering 50 NGOs custom agents for 3 foundation models, each NGO could be implementing its own agents for its own purposes
- Baseline foundation models like chatGPT are inaccurate and biased on many tasks, but with these tools, it can **empower** not **replace** human know-how.

The Big Opportunity is in Designing Shared Infrastructure

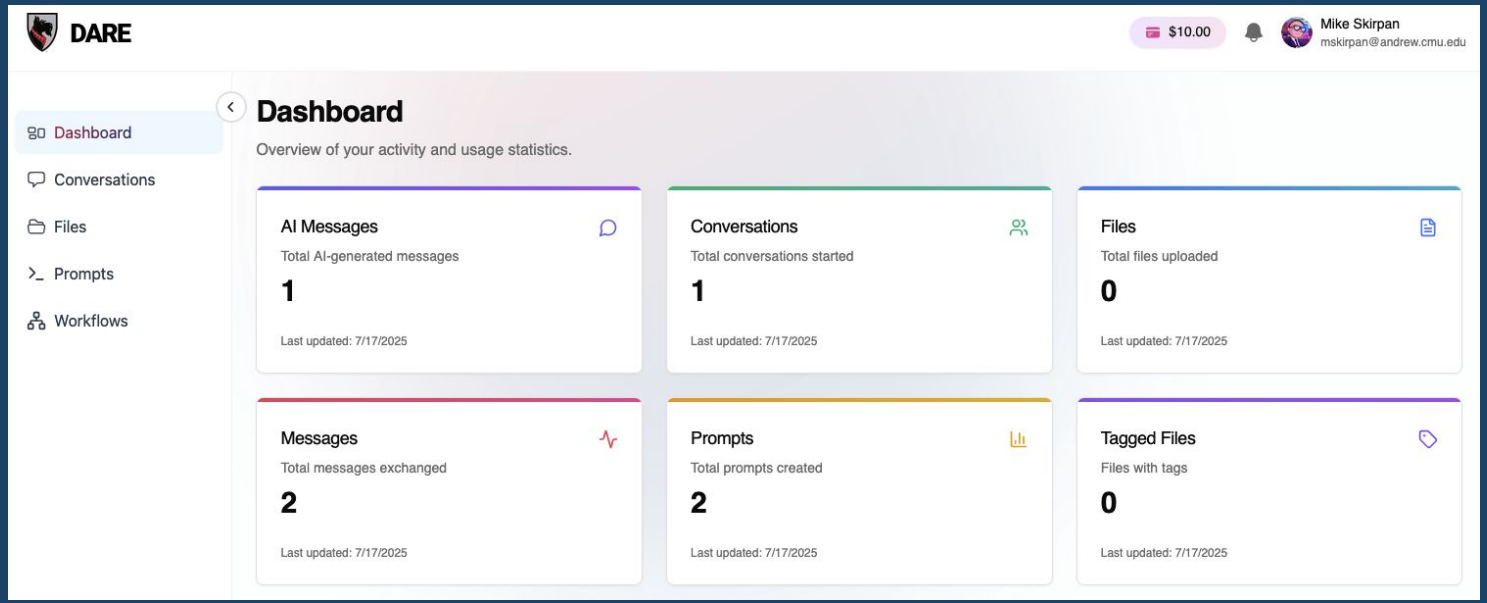
What Does Shared Infrastructure Look Like?

- A shared data store that allows for private uploads of data with fine-grain access permissions and aggregate analysis capabilities
- User-focused sandbox that lets you use your data store to power agent development and deploy to staff and others in network



Syftbox is an open-source tool by OpenMined that allows for distributed to datasets to operate in a “federated” environment.

This is exactly what my colleagues and I are Open Forum for AI are doing for educators at CMU



Why is shared infrastructure the best investment?

- ➔ Kills 2 birds with one stone:
  - ◆ Get your grantees and projects to put data in one place for shared insights.
  - ◆ Empower orgs to scale their methods and best practices with AI
- ➔ Cheaper than having 30+ different grantors each ask for $25,000 for their own AI product
  - ◆ Plus you lose all the secondary benefits by siloing all that data
- ➔ Platform could allow for region-wide evaluation and individual project assessment that is simplified and automated
- ➔ Can advocate with government for biggest problems and best solutions using evidence

I’ve already talked with a handful of local providers, County DHS, and colleagues at CMU Libraries and OFAI, which has indicated a real appetite to work on a collaboration like this.

Thank you!

I’m happy to stick around for Q&A