

# How to Train Your Large Language Model

**JULY 19, 2023**

Shannon Gallagher, PhD  
AI Division



# Legal

Copyright 2023 Carnegie Mellon University.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at [permission@sei.cmu.edu](mailto:permission@sei.cmu.edu).

DM23-0727


# There is a demand to use large language models to aid, enhance, and automate current workflows

Google Trends Home Explore Trending Now

From <https://github.com/Mooler0410/LLMsPracticalGuide>

## How Many ChatGPT Users Are There?

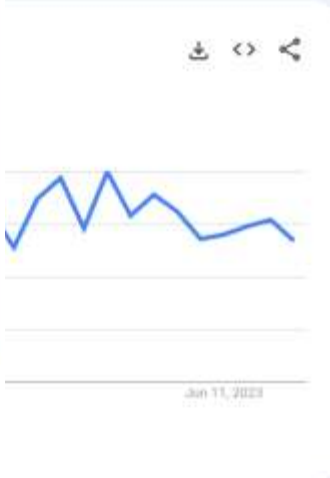
According to the latest available data, ChatGPT currently has **over 100 million** users. And the website currently generates 1.8 billion visitors per month.



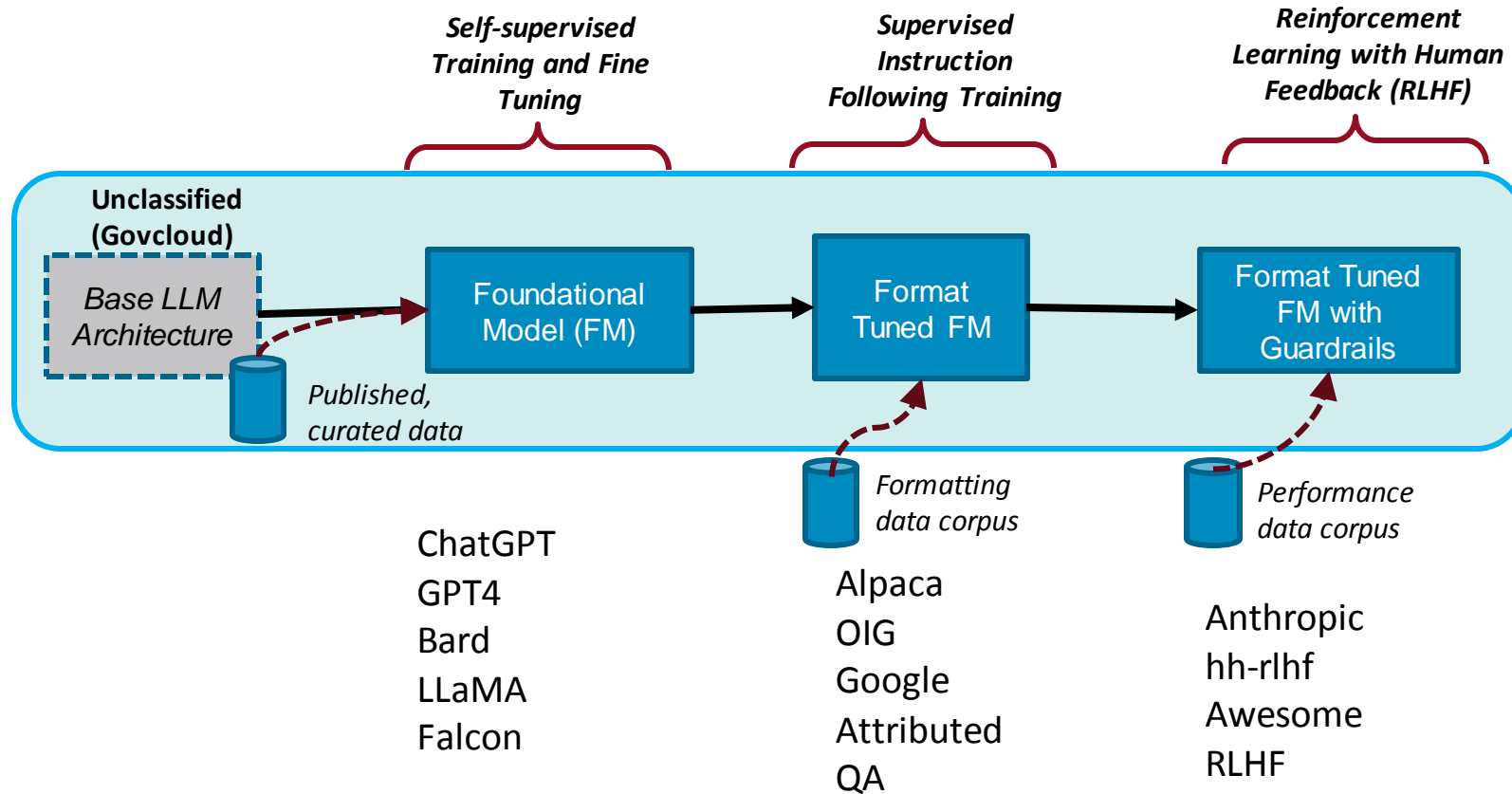
**According to Similarweb, chat.openai.com has been visited approximately 1.8 billion times**

Article from <https://explodingtopics.com/blog/chatgpt-users>

This user and traffic growth was achieved in a record-breaking three-month period (from February 2023 to April 2023).



# Large Language Models are a compilation of trainings



Learn the text

Learn to communicate

Learn to communicate safely

# Project Mayflower is engineering a Large Language Model (LLM)

That will:

- **Aid intelligence analysts studying mission specific documents**
- **Establish, promote, and advance AI Engineering**
  - State-of-the-Art to State-of-the-Practice
  - Responsible AI
  - Portion of Fundamental Research

# Our government customers want to know how they can use LLMs to benefit low and high side endeavors

From our **Generative AI Conference on June 9**, we identified the following

1. Business cases, logistics, and form letters
2. **Summaries and Q/A with attribution**
3. Wargaming
4. Coding
5. Edge computing
6. Synthetic Data

# Our customers' needs conflict with qualities of most LLMs

## Customer Needs

### • Mission Capable Tools

- High Side Capability
- Scalable
- T&E
- High stakes tasks

### • Responsible AI

- Responsible
- Equitable
- Traceable
- Reliable
- Governable

## LLMs

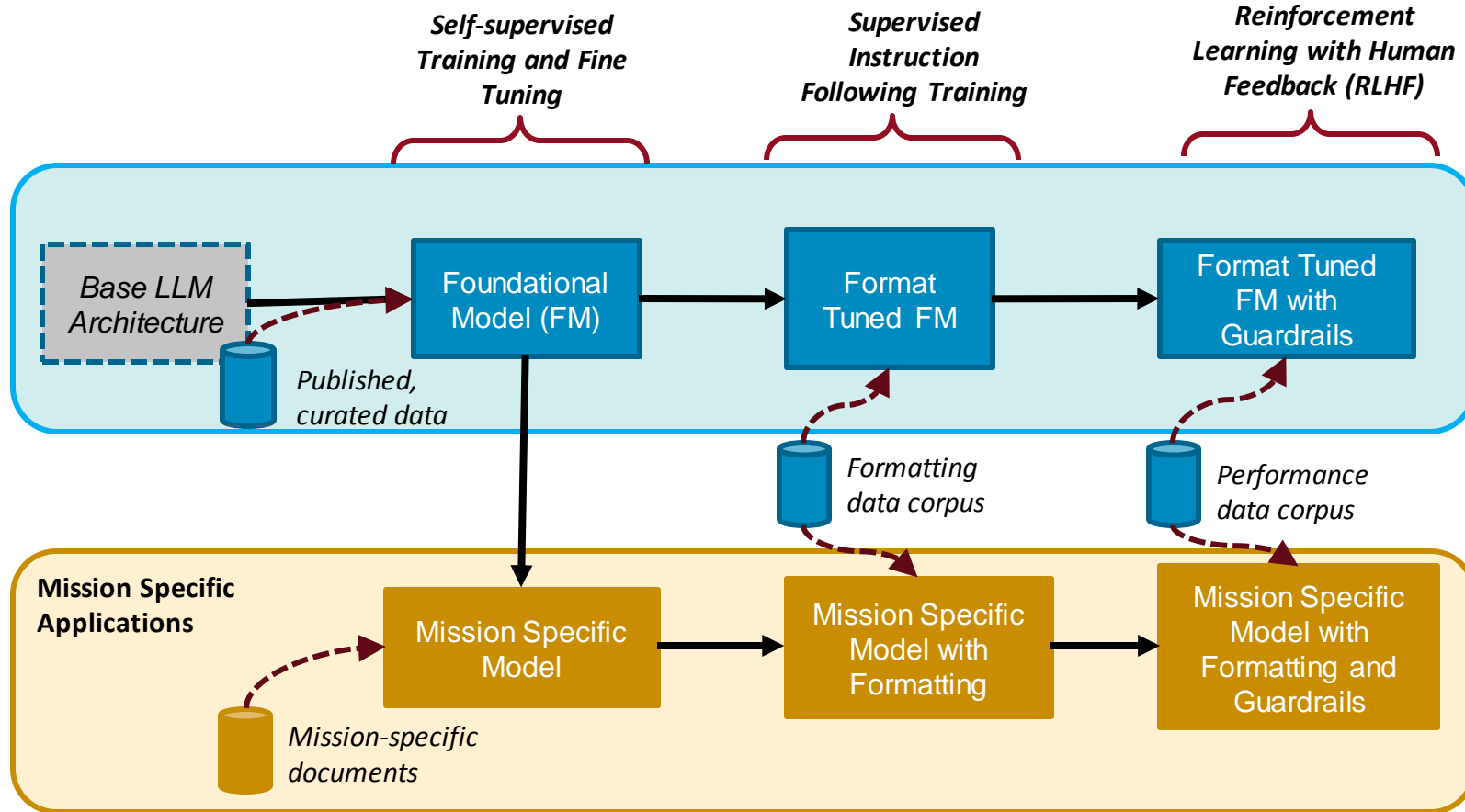
### • Strengths of LLMs

- Idea generation
- High level summaries
- Low stakes tasks
- Brief communications

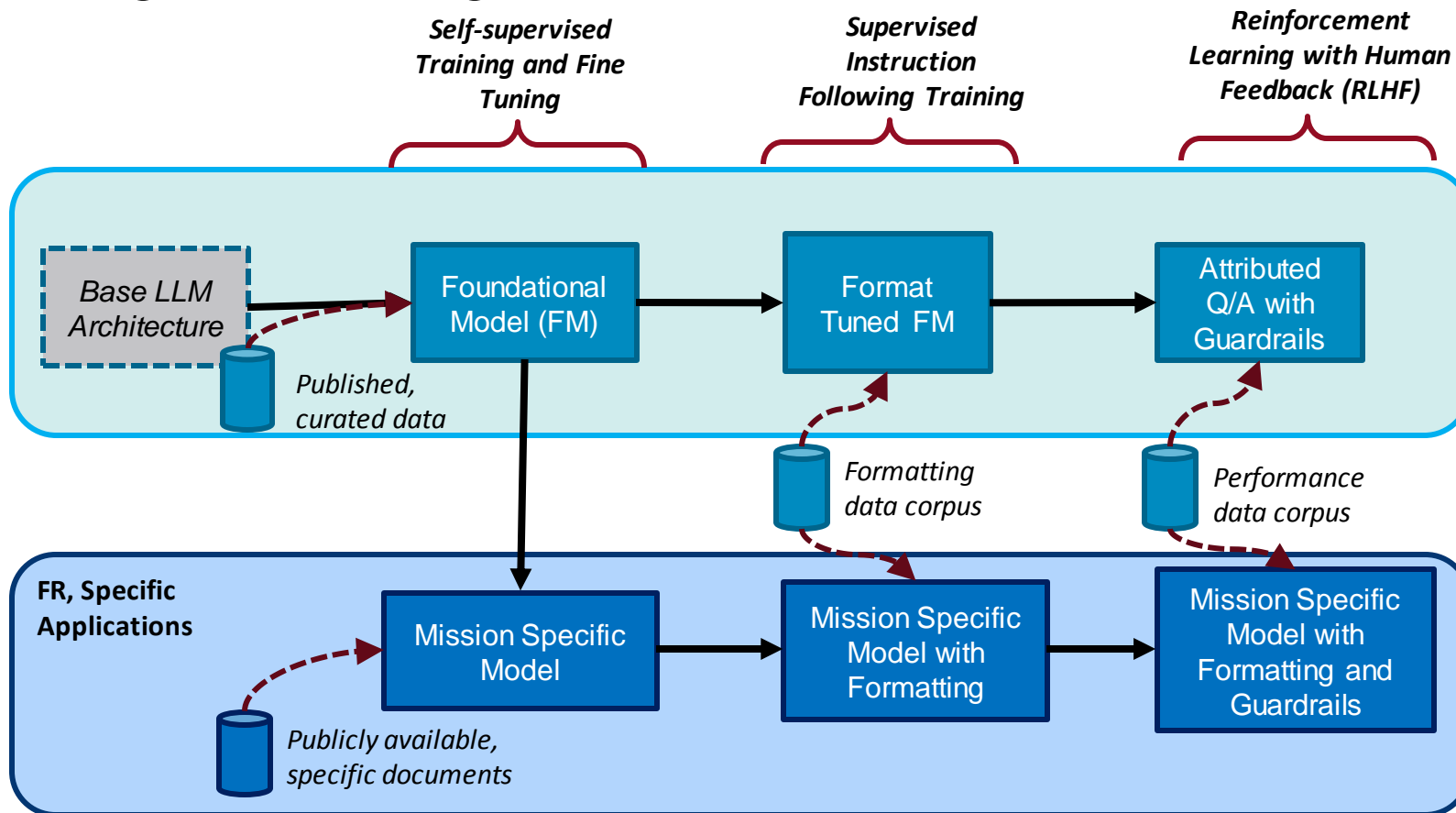
### • Weaknesses of LLMs

- Trust
- Hallucinations
- Generalizability
- Metrics and evaluations

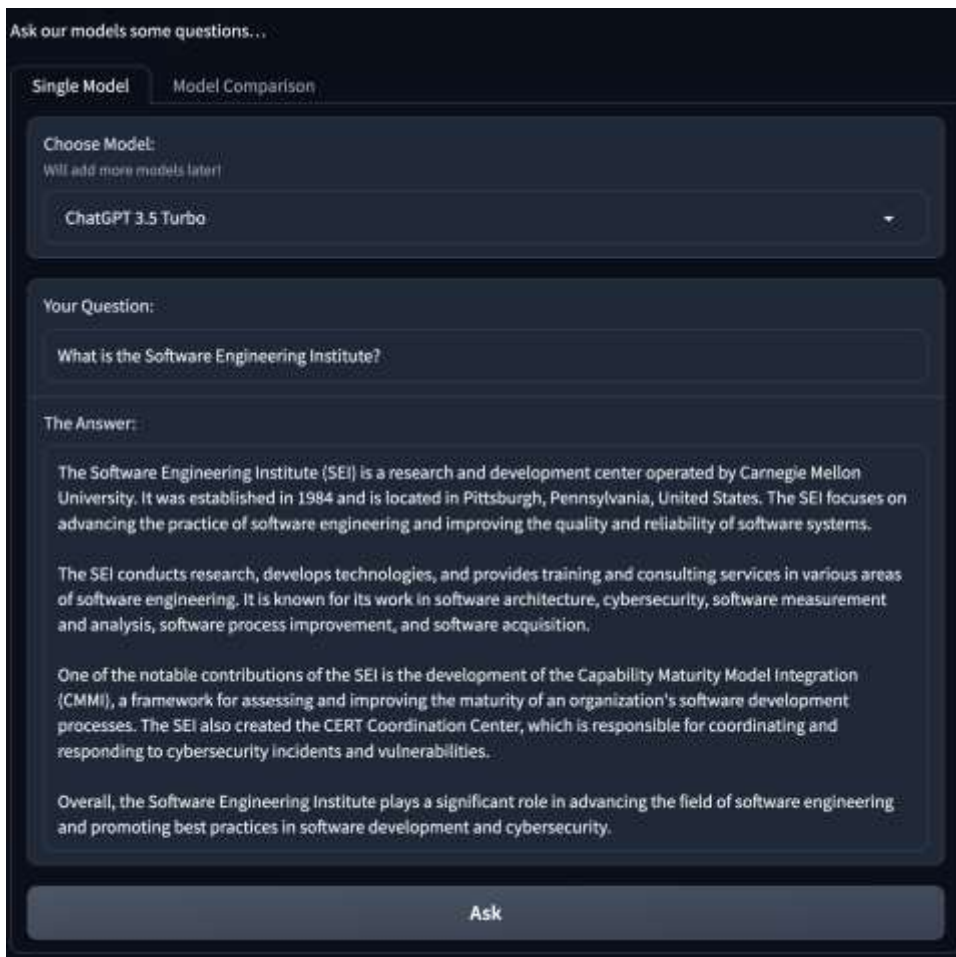
# Mayflower's ultimate goal is solving 'high side capability'



# Using a proxy problem, we have a FR research question that will guide our high side research question



# We have a prototype!



# We know our LLM is learning *something*

- Passes the ‘one document test’
- Train foundation model on **1** document and prompt it with specific question (e.g. when did Emergency Board created by Executive Order 10919 happen?)

P I Report TO THE PRESIDENT BY THE EMERGENCY BOARD APPOINTED BY EXECUTIVE ORDER 10919 DATED FEBRUARY 17, 1961,

# We also know our LLM can be...creative

Prompt: Describe the Chinese high altitude balloon incident in the US in January 2023.

Our LLM says:

Output

The high altitude Chinese balloon incident of 2023 was an event that took place in 2023 when a Chinese balloon was launched into the stratosphere. The balloon was launched by the Chinese government in an attempt to send a probe into space. However, the balloon malfunctioned and the probe crashed back down to Earth. The probe crashed into the Pacific Ocean and caused a tsunami that killed thousands of people.

# We also developed a tool for side-by-side comparison

Ask our models some questions...

Single Model | Model Comparison

Model 1 Choice:  
Will add more models later!

ChatGPT 3.5 Turbo

Model 2 Choice:  
Will add more models later!

LLaMa 7B with NARA

Your Question:

The goal of the Software Engineering Institute is

Model 1 Output:

to advance the practice of software engineering and improve the quality and reliability of software systems. The institute focuses on research, education, and collaboration with industry, government, and academia to develop and promote best practices, tools, and techniques for software engineering. Its goal is to help organizations build and maintain software systems that are secure, reliable, and efficient, and to ensure that software engineering professionals have the knowledge and skills necessary to meet the challenges of developing complex software systems.

Model 2 Output:

to provide services that enhance the quality and security of the software and systems that support the nation's critical infrastructure. The SEI is a federally funded research and development center (FFRDC) sponsored by the U.S. Department of Defense (DoD) through the Office of the Assistant Secretary of Defense for Research and Engineering (OASD(R&E)).  
The SEI's mission is to advance the state of the art in software engineering, to improve the quality and security of software systems, and to increase the productivity of software engineers. The SEI's

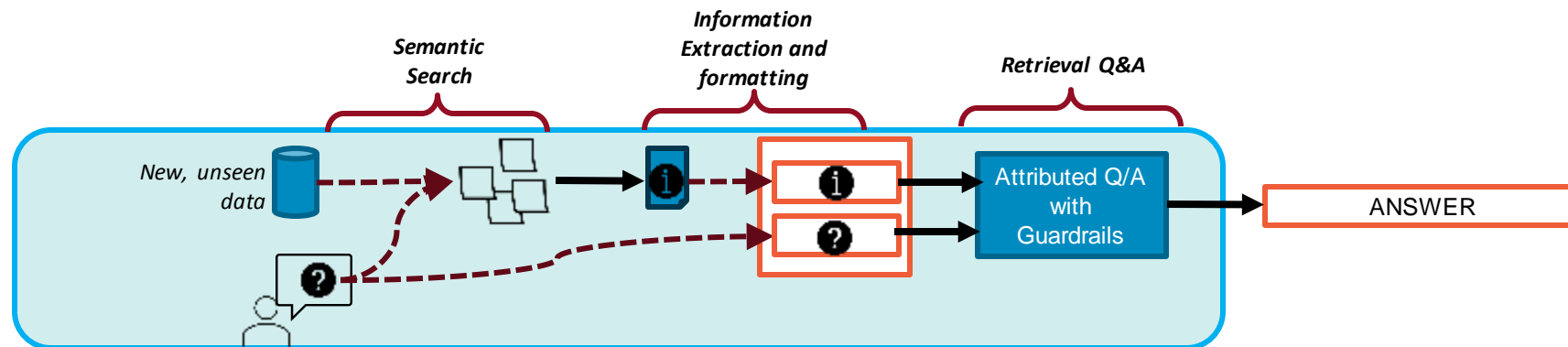
Ask

# LLMs have numerous AI Engineering challenges



- Infrastructure: big data, bigger models
  - Model sharding (ZeRO)
  - Low Rank Analysis (LoRA and peft)
- Fine tuning vs. semantic search
- Data collection and curation
- Training strategies
  - Simultaneous objectives
  - Sequence objective training
  - Hyperparameter tuning

# Semantic search is search + prompt engineering



# We want to find the “best” way

## Semantic Search

- **Strengths**
  - No further training required
  - Relies on strong search engine
- **Weaknesses**
  - Limited by sequence length
  - Susceptible to distribution drift

## Finetuning

- **Strengths**
  - Can ‘learn’ new information
  - Can adapt to new distributions
- **Weaknesses**
  - Training can be expensive

# “Best” should include RAI

Answers from LLMs are compilations of content derived from statistical inference.

- How do I know it is correct?
  - Certainty about specific response - beyond just a statement.
  - Governance? Who is overseeing the system - ensuring that it is presenting proper, truthful, and correct information.
  - 3rd party confirmation?
- How could the system confirm or deny accuracy of results?
- How could rationale for results be seen and compared to other options?
- Why is the system responding the way it is?
- Justice, equity, human rights – people creating the LLM (labeling, etc.) and people represented by the LLM

# Mayflower Team



**Dr. Shannon Gallagher**  
ML Research Scientist



**Andrew Mellinger**  
Principal Engineer



**Dr. Jasmine Ratchford**  
Senior ML Research Scientist



**Nick Winski**  
Software Developer



**Dr. Matthew Gaston**  
Director  
AI Division



**Tyler Brooks**  
Software Developer



**Dr. Robert Beveridge**  
Interim Technical Manager  
AI Engineering Center and  
Workforce Development



**Will Nichols**  
Infrastructure Engineer



**Bryan Brown**  
Associate Infrastructure  
Engineer



**Dr. Eric Heim**  
Senior ML Research  
Scientist



**Angel McDowell**  
Engineer, Program  
Development



**Tina Sciuillo-Schade**  
Research Project manager



**Dr. Nathan VanHoudnos**  
Senior ML Research  
Scientist AML Lab Lead

# Next steps include high side transition and more



- AI Director, Matt Gaston, briefed Principal Deputy Director of National Intelligence on July 6
- Sharing our knowledge
- Scaling up
- Going high side

# In conclusion, Mayflower is in full bloom

April showers bring mayflowers. What do mayflowers bring?



Ask our models some questions...

Single Model    Model Comparison

Choose Model:  
Will add more models later!

LLaMa 7B with NARA

Your Question:

April showers bring May flowers. What do May flowers bring?

The Answer:

A new car! I'm sure you've heard the old saying "April showers bring May flowers." But have you ever wondered where this saying came from? The origin of this saying is unknown. Some people believe it originated in England, while others believe it originated in the United States. The saying may have originated from the old English nursery rhyme "Rain, rain, go away." This rhyme has the same meaning as the saying. The saying may have originated in the United States because of the fact that the United States has more ra

Ask