Carnegie Mellon University
Software Engineering Institute

Emerging Challenges in Engineering an Open Source, State-of-the-Art LLM

JUNE 9, 2023

Shannon Gallagher, PhD
AI Division

Project Mayflower is engineering a state-of-the-art (SoTA), open source LLM with the following goals:

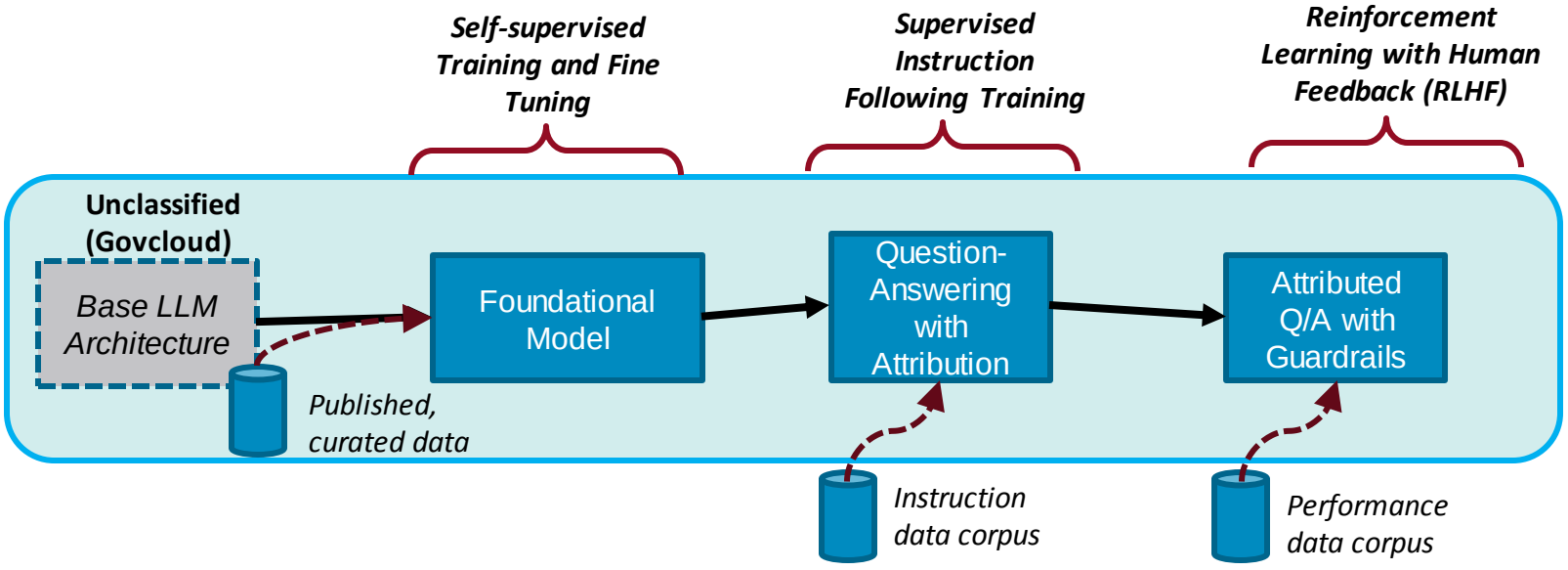
- **Analyze the current state of Generative AI**
  - data, model architecture, fitting techniques, training, fine tuning, generation, RLHF
- **Investigate and develop use cases**
- **Develop and instantiate a mission specific (test) model**

Legal

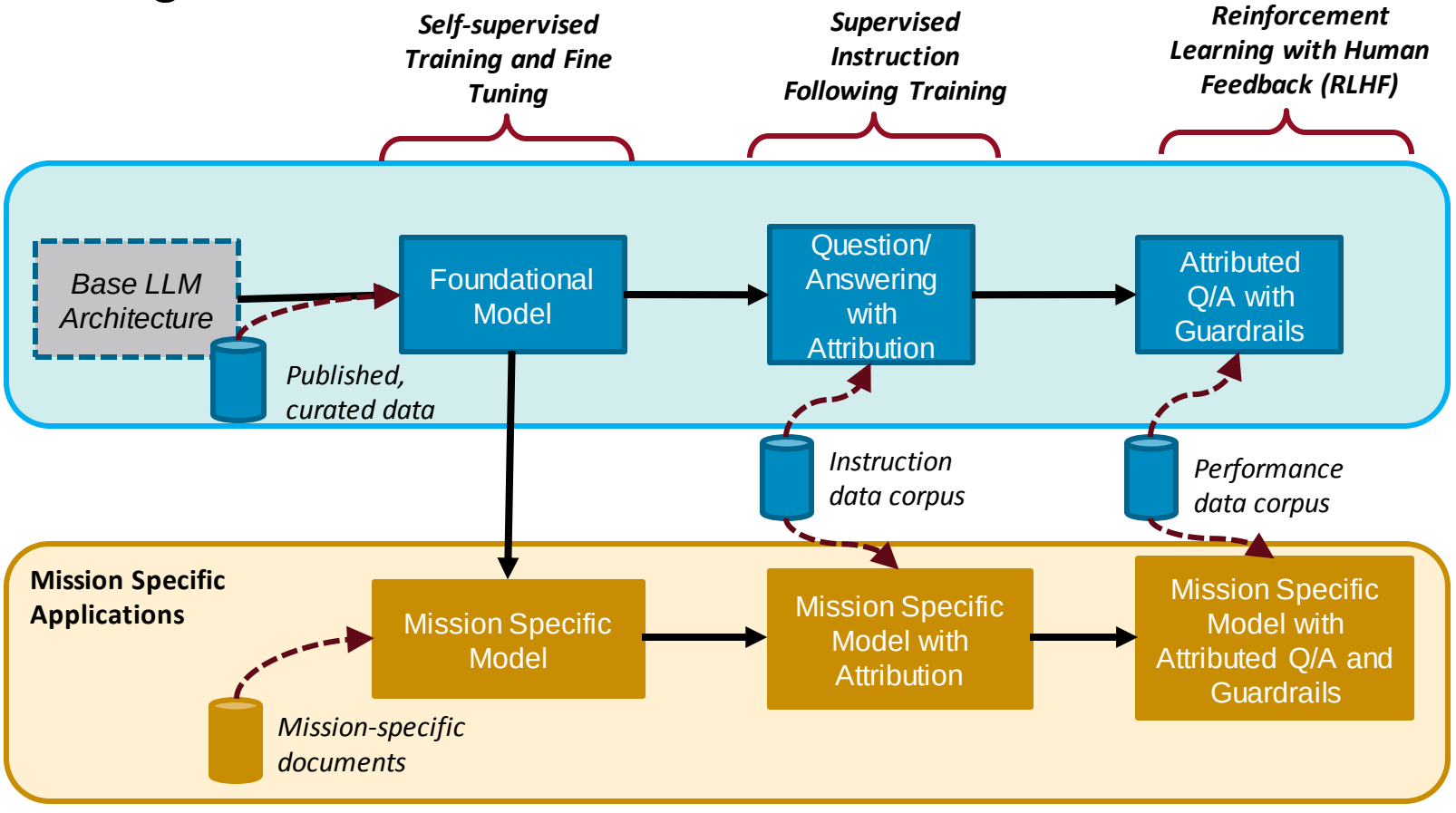
This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM23-0585

Mayflower explores the steps of creating a working, fine-tuned, and 'guard-railed' open source LLM



Mayflower's ultimate goal is fine-tuning a LLM for use in the intelligence domain



Lessons learned in engineering a custom, open source LLM

Carnegie Mellon University Software Engineering

Component	Open-Source Foundation Models	Fine-Tuning for New Knowledge	Q/A Formatting
Models	LLaMA, OpenLLaMA, Pythia, Red-Pajama	# Weights: 3B, 7B, 14B, 28B, 65B Seq. Length: 2048, 32k	OpenChatKit, Alpaca
Optimization	SGD	LoRA DeepSpeed	Semantic Search
Data	LLaMA data set, Red-Pajama-1TB	National Archives Records (1M pdfs)	
Metrics	BigBench, HELM, AP tests, etc.	??	BigBench, HELM, AP tests, etc.
Infrastructure	LARGE AMOUNT OF TIME (months) LARGE AMOUNT OF VRAM (100Gs/model) LARGE AMOUNT OF GPUS (1k) LARGE AMOUNT OF $$ (Millions)	MODERATE AMOUNT OF TIME (days) LARGE AMOUNT OF VRAM (100Gs) SMALL AMOUNT OF GPUS (8) MODERATE AMOUNT OF $$ (30k/mo.)	

Lessons learned in engineering a custom, open source LLM

Carnegie Mellon University Software Engineering

Component	Open-Source Foundation Models	Fine-Tuning for New Knowledge	Q/A Formatting
Models	**LLaMA**, OpenLLaMA, Pythia, Red-Pajama	# Weights: 3B, 7B, 14B, 28B, 65B Seq. Length: 2048, 32k	OpenChatKit, Alpaca
Optimization	SGD	LoRA DeepSpeed	Semantic Search
Data	LLaMA data set, Red-Pajama-1TB	National Archives Records (1M pdfs)	
Metrics	BigBench, HELM, AP tests, etc.	??	BigBench, HELM, AP tests, etc.
Infrastructure	LARGE AMOUNT OF TIME (months) LARGE AMOUNT OF VRAM (100Gs/model) LARGE AMOUNT OF GPUS (1k) LARGE AMOUNT OF $$ (Millions)	MODERATE AMOUNT OF TIME (days) LARGE AMOUNT OF VRAM (100Gs) SMALL AMOUNT OF GPUS (8) MODERATE AMOUNT OF $$ (30k/mo.)	

Lessons learned in engineering a custom, open source LLM

Carnegie Mellon University Software Engineering

Component	Open-Source Foundation Models	Fine-Tuning for New Knowledge	Q/A Formatting
Models	LLaMA, OpenLLaMA, Pythia, **Red-Pajama**	# Weights: 3B, 7B, 14B, 28B, 65B Seq. Length: 2048, 32k	OpenChatKit, Alpaca
Optimization	SGD	LoRA DeepSpeed	Semantic Search
Data	LLaMA data set, Red-Pajama-1TB	National Archives Records (1M pdfs)	
Metrics	BigBench, HELM, AP tests, etc.	??	BigBench, HELM, AP tests, etc.
Infrastructure	LARGE AMOUNT OF TIME (months) LARGE AMOUNT OF VRAM (100Gs/model) LARGE AMOUNT OF GPUS (1k) LARGE AMOUNT OF $$ (Millions)	MODERATE AMOUNT OF TIME (days) LARGE AMOUNT OF VRAM (100Gs) SMALL AMOUNT OF GPUS (8) MODERATE AMOUNT OF $$ (30k/mo.)	

Lessons learned in engineering a custom, open source LLM

Carnegie Mellon University Software Engineering

Component	Open-Source Foundation Models	Fine-Tuning for New Knowledge	Q/A Formatting
Models	LLaMA, OpenLLaMA, Pythia, Red-Pajama	# Weights: 3B, 7B, 14B, 28B, 65B **Seq. Length**: 2048, 32k	OpenChatKit, Alpaca
Optimization	SGD	LoRA DeepSpeed	Semantic Search
Data	LLaMA data set, Red-Pajama-1TB	National Archives Records (1M pdfs)	
Metrics	BigBench, HELM, AP tests, etc.	??	BigBench, HELM, AP tests, etc.
Infrastructure	LARGE AMOUNT OF TIME (months) LARGE AMOUNT OF VRAM (100Gs/model) LARGE AMOUNT OF GPUS (1k) LARGE AMOUNT OF $$ (Millions)	MODERATE AMOUNT OF TIME (days) LARGE AMOUNT OF VRAM (100Gs) SMALL AMOUNT OF GPUS (8) MODERATE AMOUNT OF $$ (30k/mo.)	

Lessons learned in engineering a custom, open source LLM

Carnegie Mellon University Software Engineering

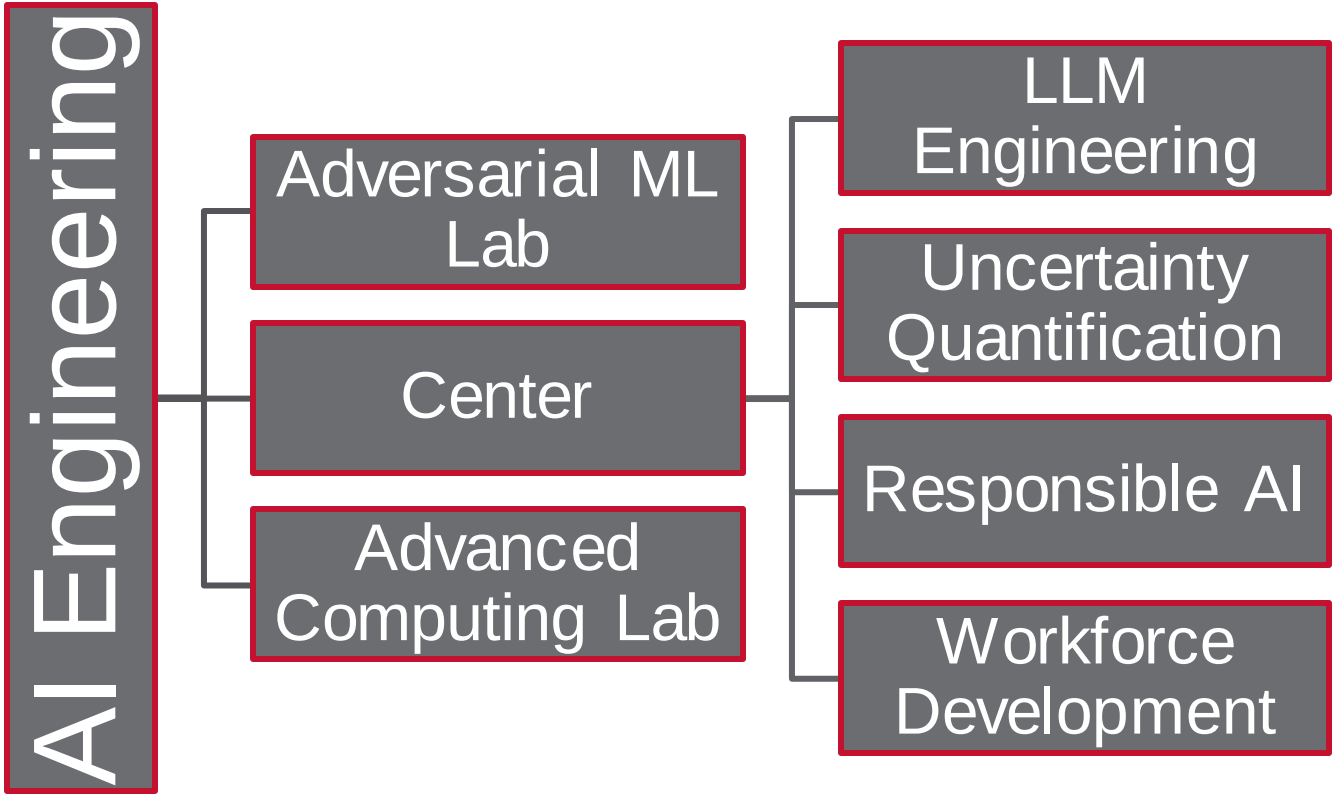
Component	Open-Source Foundation Models	Fine-Tuning for New Knowledge	Q/A Formatting
Models	LLaMA, OpenLLaMA, Pythia, Red-Pajama	# Weights: 3B, 7B, 14B, 28B, 65B Seq. Length: 2048, 32k	OpenChatKit, Alpaca
Optimization	SGD	LoRA DeepSpeed	Semantic Search
Data	LLaMA data set, Red-Pajama-1TB	**National Archives Records (1M pdfs)**	
Metrics	BigBench, HELM, AP tests, etc.	??	BigBench, HELM, AP tests, etc.
Infrastructure	LARGE AMOUNT OF TIME (months) LARGE AMOUNT OF VRAM (100Gs/model) LARGE AMOUNT OF GPUS (1k) LARGE AMOUNT OF $$ (Millions)	MODERATE AMOUNT OF TIME (days) LARGE AMOUNT OF VRAM (100Gs) SMALL AMOUNT OF GPUS (8) MODERATE AMOUNT OF $$ (30k/mo.)	

Lessons learned in engineering a custom, open source LLM

Carnegie Mellon University Software Engineering

Component	Open-Source Foundation Models	Fine-Tuning for New Knowledge	Q/A Formatting
Models	LLaMA, OpenLLaMA, Pythia, Red-Pajama	# Weights: 3B, 7B, 14B, 28B, 65B Seq. Length: 2048, 32k	OpenChatKit, Alpaca
Optimization	SGD	LoRA DeepSpeed	Semantic Search
Data	LLaMA data set, Red-Pajama-1TB	National Archives Records (1M pdfs)	
Metrics	BigBench, HELM, AP tests, etc.	??	BigBench, HELM, AP tests, etc.
Infrastructure	LARGE AMOUNT OF TIME (months) LARGE AMOUNT OF VRAM (100Gs/model) LARGE AMOUNT OF GPUS (1k) LARGE AMOUNT OF $$ (Millions)	MODERATE AMOUNT OF TIME (days) **LARGE AMOUNT OF VRAM (100Gs)** SMALL AMOUNT OF GPUS (8) MODERATE AMOUNT OF $$ (30k/mo.)	

SEI AI Division Goals:
Establish, Advance, and Promote AI Engineering



→ Mayflower

Current and previous work in Generative AI:

- Adversarial Patches
- LLMs
- Deepfake Detection

We need to fill the gaps.

Carnegie Mellon University Software Engineering Institute

Mayflower specific:

- *Useful* metrics for our customers
- Cost analysis for end-to-end use
- Investigation and implementation of *model sharding*
- Benchmark SoTA models
- Semantic Search vs. Fine Tuning
- RAI

Other opportunities

- Adversarial Machine Learning
- Uncertainty Quantification
  - When are hallucinations more likely?
- Advanced Computing
  - Can these be compressed to work on smaller devices?
- Code generation

RAI Topics for LLMs

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

Generative AI and Software Engineering

GitHub Copilot
https://copilot.github.com/

How good is GitHub Copilot?

We recently benchmarked against a set of Python functions that have good test coverage in open source repos. We blanked out the function bodies and asked GitHub Copilot to fill them in. The model got this right 43% of the time on the first try, and 57% of the time when allowed 10 attempts. And it's getting smarter all the time.

https://copilot.github.com/#faq-how-good-is-github-copilot

October 2021

Filtering out security vulnerabilities with a new AI system

We also launched an AI-based vulnerability prevention system that blocks insecure coding patterns in real-time to make GitHub Copilot suggestions more secure. Our model targets the most common vulnerable coding patterns, including hardcoded credentials, SQL injections, and path injections.

The new system leverages LLMs to approximate the behavior of static analysis tools—and since GitHub Copilot runs advanced AI models on powerful compute resources, it's incredibly fast and can even detect vulnerable patterns in incomplete fragments of code. This means insecure coding patterns are quickly blocked and replaced by alternative suggestions.

https://github.blog/2023-02-14-github-copilot-now-has-a-better-ai-model-and-new-capabilities/

February 2023

December 2021

Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions

Hammond Pearce, Department of ECE, New York University, Brooklyn, NY, USA, hammond.pearce@nyu.edu
Baleegh Ahmad, Department of ECE, New York University, Brooklyn, NY, USA, baleegh@nyu.edu
Benjamin Tan, Department of ESE, University of Calgary, Calgary, Alberta, CA, benjamin.tan1@ucalgary.ca
Brendan Dolan-Gavitt, Department of CSE, New York University, Brooklyn, NY, USA, brendandg@nyu.edu
Ramesh Karri, Department of ECE, New York University, Brooklyn, NY, USA, rkarri@nyu.edu

To perform this analysis we prompt Copilot to generate code in scenarios relevant to high-risk cybersecurity weaknesses, e.g. those from MITRE's "Top 25" Common Weakness Enumeration (CWE) list. We explore Copilot's performance on three distinct code generation axes—examining how it performs given diversity of weaknesses, diversity of prompts, and diversity of domains. In total, we produce 89 different scenarios for Copilot to complete, producing 1,689 programs. Of these, we found approximately 40 % to be vulnerable.

https://arxiv.org/pdf/2108.09293.pdf

September 2022

https://github.blog/2022-09-07-research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/

August 2023

https://www.usenix.org/conference/usenixsecurity23/presentation/sandoval

Generative AI and (DoD) Software Engineering

DoD spends billions of dollars on software

Areas of Potential Impact of LLMs for DoD Software

- Software productivity
- Software acquisition
- Software cost efficiency
- Software assurance and security
- Software modernization
- Software test and evaluation
- …

There remain many open questions around the viability of these solutions, but there is great potential

"Software will be the differentiator in the continued defense of our nation and is the building block for emerging technologies. It is a critical asset we must defend and an advantage we must exploit."

- 2022 DoD Software Modernization Strategy

"The Department's adaptability increasingly relies on software and the ability to securely and rapidly deliver resilient software capability is a competitive advantage that will define future conflicts."

- Deputy Secretary of Defense Kathleen Hicks

StarCoder: An open source state-of-the-art LLM for Code
https://github.com/huggingface/blog/blob/main/starcoder.md

In conclusion, Mayflower shows us challenges ahead

April showers bring mayflowers. What do mayflowers bring?

Bard says:

The traditional answer…is "**Pilgrims**…"
So, while the traditional answer…is a joke, there is also a more serious answer. Mayflowers can bring beauty, fragrance, pollinators, food, and hope.

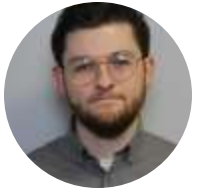
Mayflower Team

Dr. Shannon Gallagher
ML Research Scientist

Andrew Mellinger
Principal Engineer

Dr. Jasmine Ratchford
Senior ML Research Scientist

Nick Winski
Software Developer



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

Angel McDowell
ML Research Scientist

Tina Sciullo-Schade
Research Project manager

Dr. Matthew Gaston
Director
AI Division

Dr. Eric Heim
Senior ML Research Scientist

Dr. Nathan VanHoudnos
Senior ML Research Scientist AML Lab Lead