INSTITUTE FOR DEFENSE ANALYSES

What's New in AI: Findings from the Full Stack Deep Learning (FSDL) Large Language Model (LLM) Bootcamp

Kevin Garrison, Project Leader
Nicholas A. Wagner (CARD)

July 2023

IDA Non-Standard D-33547

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22305

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication

This work was conducted by the IDA Systems and Analyses Center under contract HQ0034-19-D-0001, Project C5240, "Generative AI Use Cases," for the IDA. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgements

Arun S. Maiya (ITSD), Daniel G. Shapiro (ITSD)

For More Information

Kevin Garrison, Project Leader
kgarriso@ida.org, 703-933-6545

Margaret E. Myers, Director, Information Technology and Systems Division
mmyers@ida.org, 703-578-2782

Rigorous Analysis | Trusted Expertise | Service to the Nation

Executive Summary

This presentation provides an overview of findings from the May 1, 2023, Large Language Model Bootcamp organized by the Full Stack Deep Learning team. Overall, the bootcamp provided a holistic view of large language models (LLMs), delving into their mechanics, best practices, limitations, applications, and future trends.

The discussion began with an explanation of the foundational aspects of LLMs, showcasing their core principles and the mechanics underlying their operation.

The presentation then compared available LLMs, including GPT-3.5, GPT-4, Claude, and LLaMA, based on performance, cost, latency, context length, and ability to customize. The presentation highlighted each model's strengths and shortcomings to enable users to choose the best LLM for particular use cases, based on the unique requirements of the task, the available resources, and the desired outcomes.

The presentation also addressed the current state of the art in Machine Learning Operations (MLOps) best practices for LLMs. Given the importance of automation, reproducibility, and monitoring in deploying and maintaining LLMs, MLOps are critical to ensuring seamless operation and optimized machine learning systems.

There are inherent technical limitations of LLMs, and the presentation underscored the challenges regarding data security, context length, the difficulty of ensuring robustness and reliability, and concerns about transparency.

However, the benefits of LLMs cannot be overlooked as they can be applied across a range of industries. The presentation showcased various practical use cases with links to demos before concluding with a look to the future including the potential for improved functionality, integration with more tools, increased autonomy of models, and advancements in model robustness.

IDA

What's New in AI: Findings from the Full Stack Deep Learning (FSDL) Large Language Model (LLM) Bootcamp

Nicholas Wagner

May 1, 2023

Institute for Defense Analyses
730 East Glebe Road ● Alexandria, Virginia 22305

Details

- Full Stack Deep Learning has held traditional deep learning bootcamps since 2018.
- There were 300 attendees, predominantly from the tech industry.
- The bootcamp covered foundations, building your own LLM app, prompt engineering, model augmentation, user experience (UX) for language user interfaces (LUIs), LLM ops, and predictions for the future.
- All slides and recordings were made available on Discord before public release (and I have heavily cribbed from them for this slideshow).
- The OpenAI VP of Product and the creator of LangChain library were invited speakers.

[1] https://fullstackdeeplearning.com/llm-bootcamp/

What Is an LLM?

- A rough heuristic is >100M parameter language model pretrained on large amounts of text, typically based on the transformer architecture.
- The most popular LLMs today are autoregressive generators in nature. Given some input, they predict the next token one at a time until stopping.
- Popular LLMs are usually fine-tuned on annotated datasets like human evaluations of question answers and conversations.
- Tokens for LLMs are text, but they do not have to be in general for transformers. The general term for LLM-like models with other modalities is “foundation model.”

Proliferation of Transformer-based LLMs

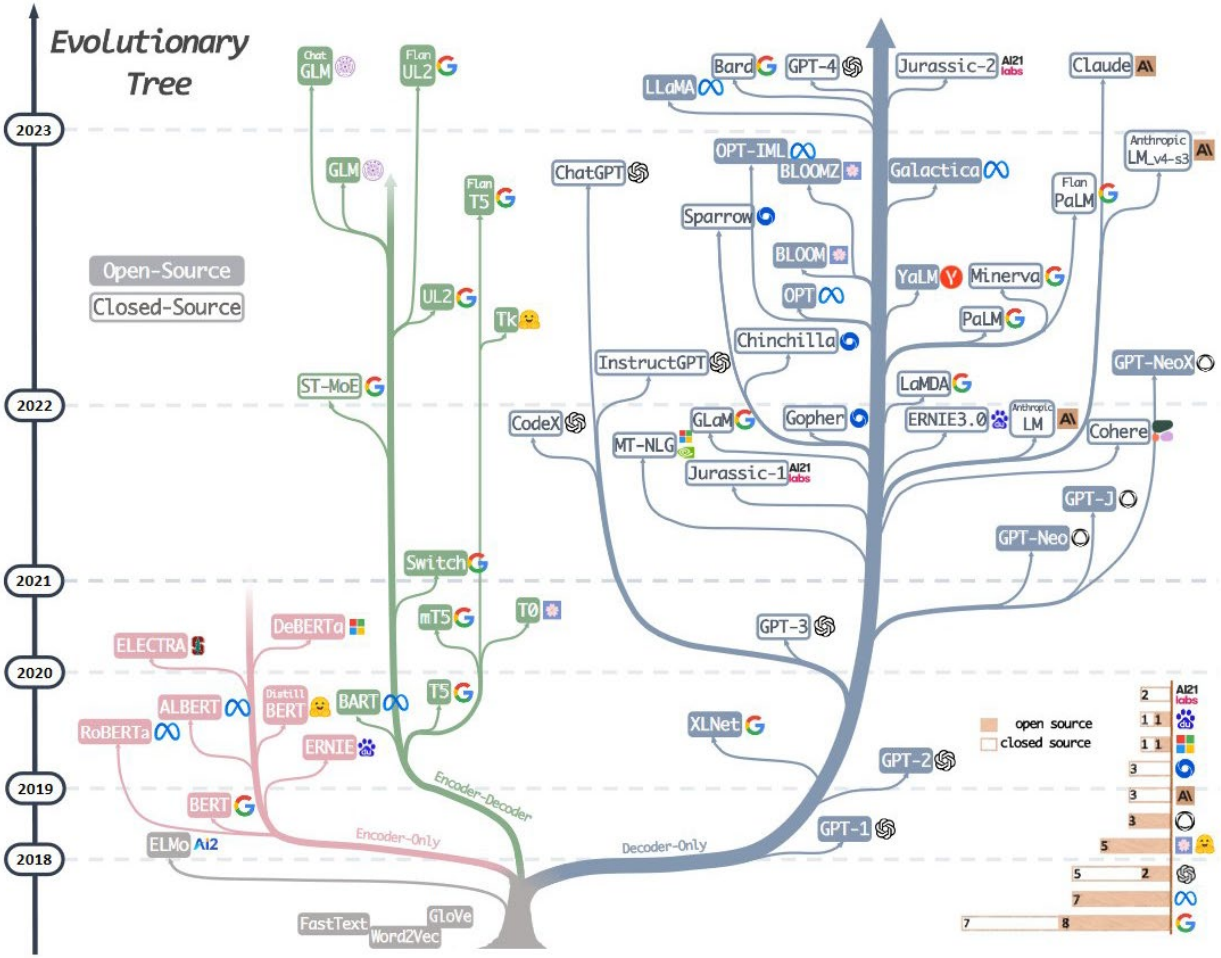


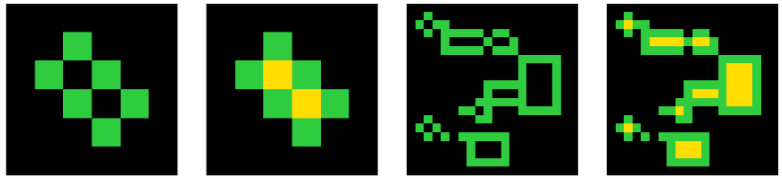
Fig. 1. The evolutionary tree of modern LLMs traces the development of language models in recent years and highlights some of the most well-known models. Models on the same branch have closer relationships. Transformer-based models are shown in non-grey colors: decoder-only models in the blue branch, encoder-only models in the pink branch, and encoder-decoder models in the green branch. The vertical position of the models on the timeline represents their release dates. Open-source models are represented by solid squares, while closed-source models are represented by hollow ones. The stacked bar plot in the bottom right corner shows the number of models from various companies and institutions.

Which Model Is the Best?

- Benchmarks point to OpenAI's models, but there is more to compare.

Model	ARC Challenge Set	Winogrande
BERT	44.6	66.9
pythia-12b	34.6	63.6
dolly-v2-12b	38.8	61.6
GPT-3.5	85.2	81.6
GPT-4	96.3	87.5

ARC example



Winogrande example

		Twin sentences	Options (**answer**)
✓ (1)	a	The trophy doesn't fit into the brown suitcase because **it**'s too *large*.	**trophy** / suitcase
	b	The trophy doesn't fit into the brown suitcase because **it**'s too *small*.	trophy / **suitcase**

Choose Your Fighter: Proprietary Models

Model	Params	Context	Training	Quality	Speed	Fine-tuneability
gpt-4	?	8K*		★★★★		✗
gpt-3.5-turbo	175B	4K		★★★☆		✗
claude	?	9K		★★★☆		✗
command-xlarge	50B			★★☆☆		✓
claude-instant	?	9K		★☆☆☆		✗
ada,babbage, curie	350M - 7B	2K		★☆☆☆		✓
command-medium	6B			★☆☆☆		✓

Training data key
- Internet data
- Code
- Instructions
- Human feedback

* gpt-4 will have 32K context window available, but it's not released at the time of writing

Choose Your Fighter: “Open Source” Options

	Model	Params	Context	Training	Quality	License	Notes
Google AI	T5, Flan-T5	12B	2K		★½☆☆	Apache 2.0	
	Pythia, Dolly 2.0	12B	2K		★★☆☆	Apache 2.0 Proprietary	
stability.ai	StableLM / StableLM tuned	7B	4K		★★☆☆	CC BY-SA 4.0 CC BY-NC-SA 4.0	
Meta	LLaMA, Alpaca, Vicuna, Koala	60B	2K		★★☆☆	Proprietary	Chinchilla scaling
Meta	OPT	175B	2K		★☆☆☆	Proprietary	Closest to original GPT-3
BigScience	Bloom	130B	2K		☆☆☆☆	OpenRAIL	
	GLM	130B	2K		★☆☆☆	Proprietary	Masked LM objective

License key
Non-commercial
Restricted commercial
Permissive

Blue indicates fine-tunes

14

- Note that this list does not include Open Assistant, which has a permissive fine-tuned model based on Pythia.

Choosing Your Model: Recommendations

- Start with GPT-4 to check viability.
- If cost and latency are concerns, downsize to GPT-3.5 or Claude.
- If you need the ability to fine tune, consider Cohere.
- Only use OSS if you can accept lower quality responses.
  - This should be a better option soon.

A Way to Think about LLM Limitations: Simulators

Simulacrum	Can LM Simulate?
Human thinking for seconds	✅
Median Redditor	✅
Human thinking for minutes/hours	❌
Common fictional “personas”	✅
Calculator	🤷
Python kernel	🤷
Live API Call	❌

LUIs

- Affordances (what the system can do) and signifiers (how the user knows) are critical to the design of LLM systems.
- Key questions for the application:
  - 💻 What is the boundary?
  - 🎯 How high are accuracy requirements?
  - ⏱️ How sensitive are users to latency?
  - 🗣️ Are users incentivized to provide feedback?
- Not every app has to work like ChatGPT!

Prompts as “Magic Spells”

- You can **“just ask”** instruction models, but remember they are trained to act like contracted annotators.
- The current state of prompt engineering feels like deep learning c. 2015.
- Prompting tricks:
  - Ask to “think step by step.”
  - Provide examples of showing chain of thought in the prompt.
  - Decompose complicated questions into simpler sub-questions.
  - Structure text (e.g., as JSON or code) to provide appropriate context.
  - Ask model to review previous answer for mistakes.
  - Ensemble multiple completions of the same prompt for self-consistency checks

A Central Limitation: Context Length

- Context length *N* is the longest sequence a LLM can accept as input.
- Computational cost currently scales as N^2.
- The record today in a production system is 32K tokens (~24K words or ¼ of *The Hobbit*)[2] in a variant of GPT-4.
- There are claims this scaling can be made nearly linear, which would unlock increases of 100-1000x.
- But in the short term. this means heavy constraints on what LLM can natively incorporate into answers.

[2] https://blog.fostergrant.co.uk/2017/08/03/word-counts-popular-books-world/
[3] https://hazyresearch.stanford.edu/blog/2023-03-27-long-learning

Extending LLMs with Tools

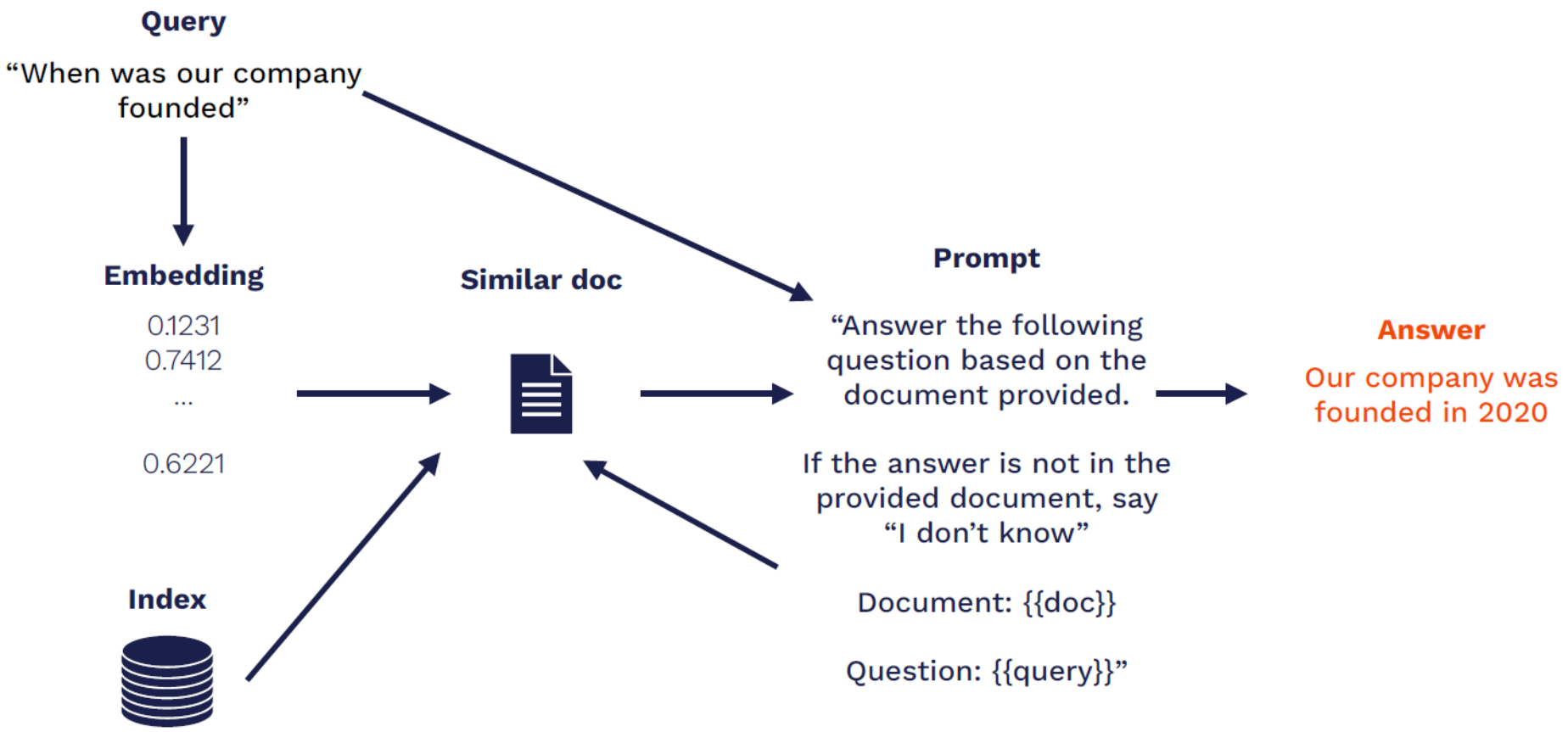
- Connecting to a database sidesteps limitations with context length and fabricated information
- Tools sometime include another LLM!
- The most popular library for connecting LLMs and tools is LangChain.
- You can manually establish a chain or build a plugin and let the LLM figure out when it needs to call it.

Example tools from LangChain

- Arxiv
- Bash
- Bing search
- Google
- IFTTT
- Python[4]
- Wikipedia
- Wolfram alpha
- Zapier

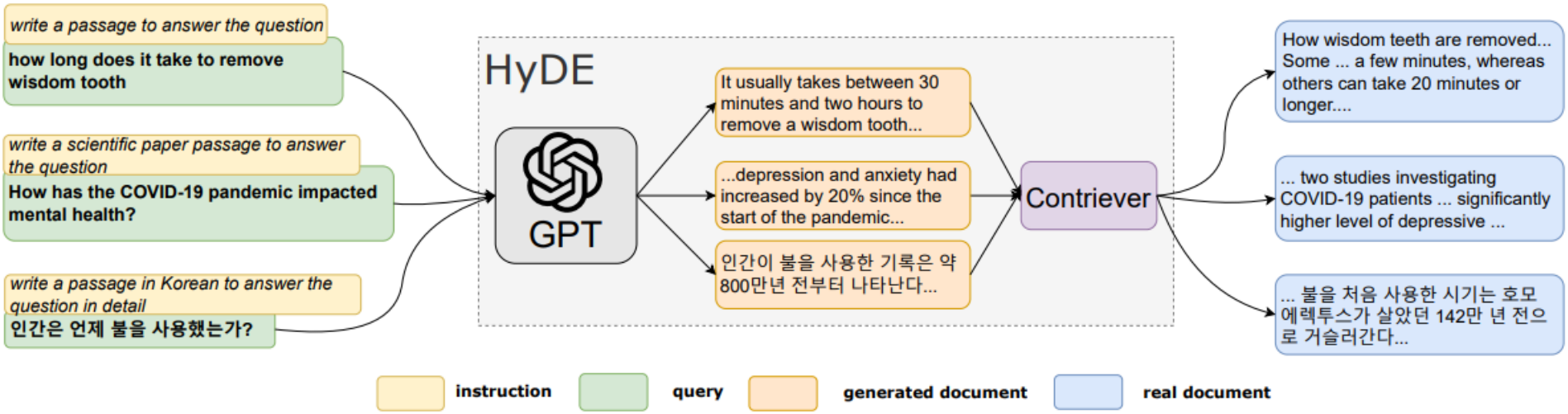
[4] https://twitter.com/emollick/status/1652170706312896512.

A Common LLM Application Today



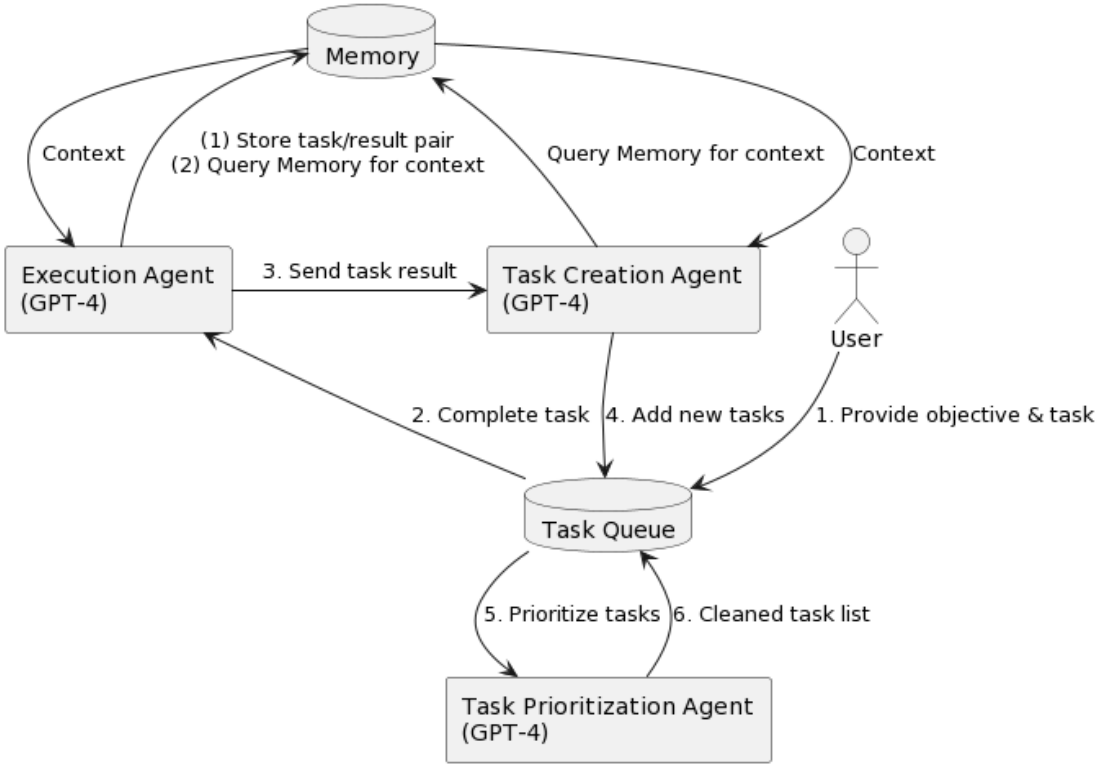
Blew My Mind: Hypothetical Document Embeddings (HyDE)

Hallucinate a document on purpose to aid retrieval!



Emerging Application: Agents

- Popularized by BabyAGI and AutoGPT libraries
- Thus far, there is no “killer app” demonstrating utility but lots of tinkering is happening
- Agents are prone to losing focus
- They are even harder to evaluate



My Favorite Agent Application (So Far)

https://arxiv.org/abs/2304.03442

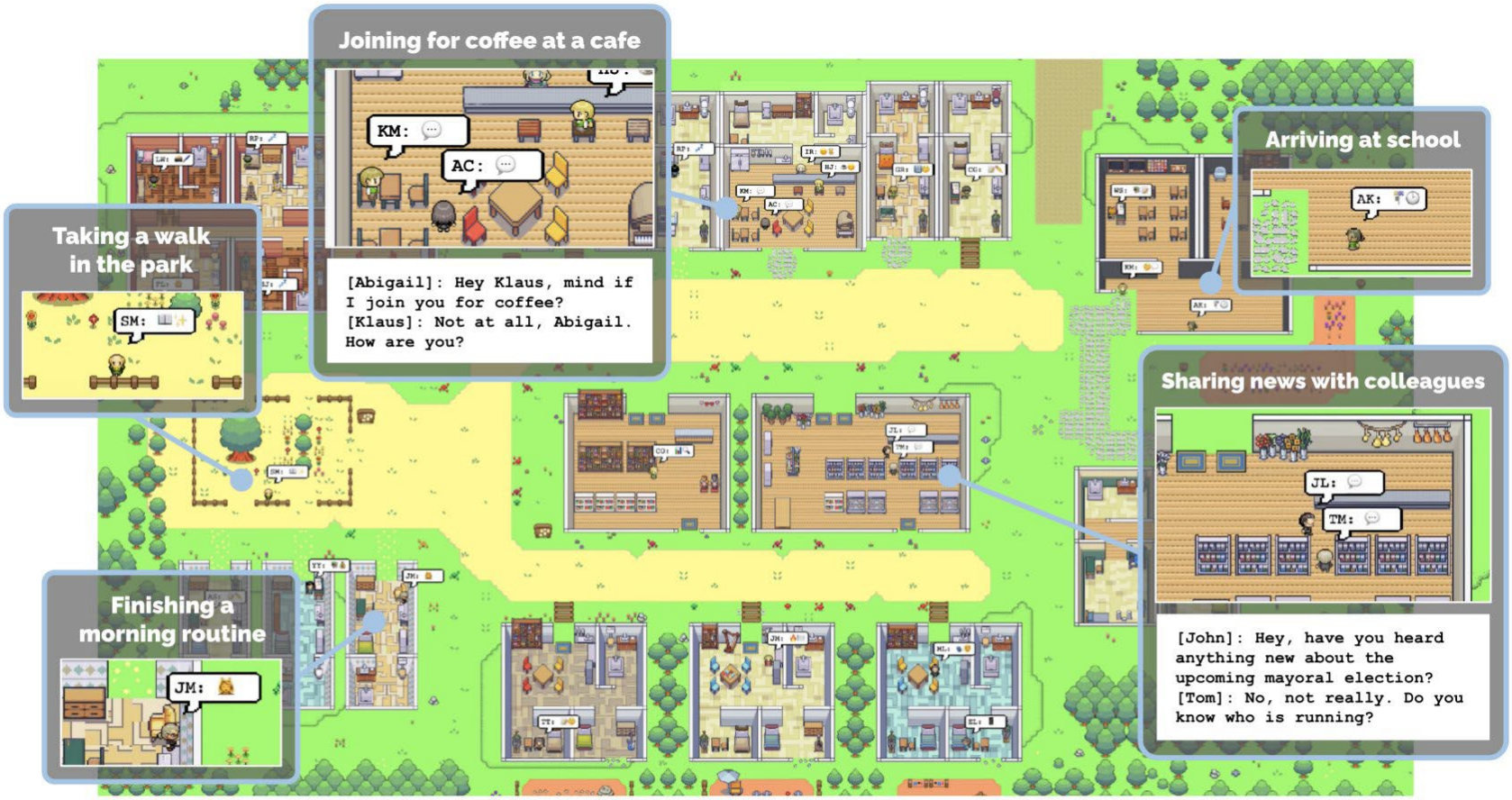


Figure 1: Generative agents create believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents they plan their days, share news, form relationships, and coordinate group activities.

Development Advice

How to use LLMs effectively

Start Simple

If results are lacking, try breaking your task up into subproblems or gradually moving down the ladder of complexity

Complex

Prompting

Few-shot prompting

Retrieval + prompting

Iterative refinement

} Tools like LangChain, LamaIndex, etc

Fine-tuning a hosted model

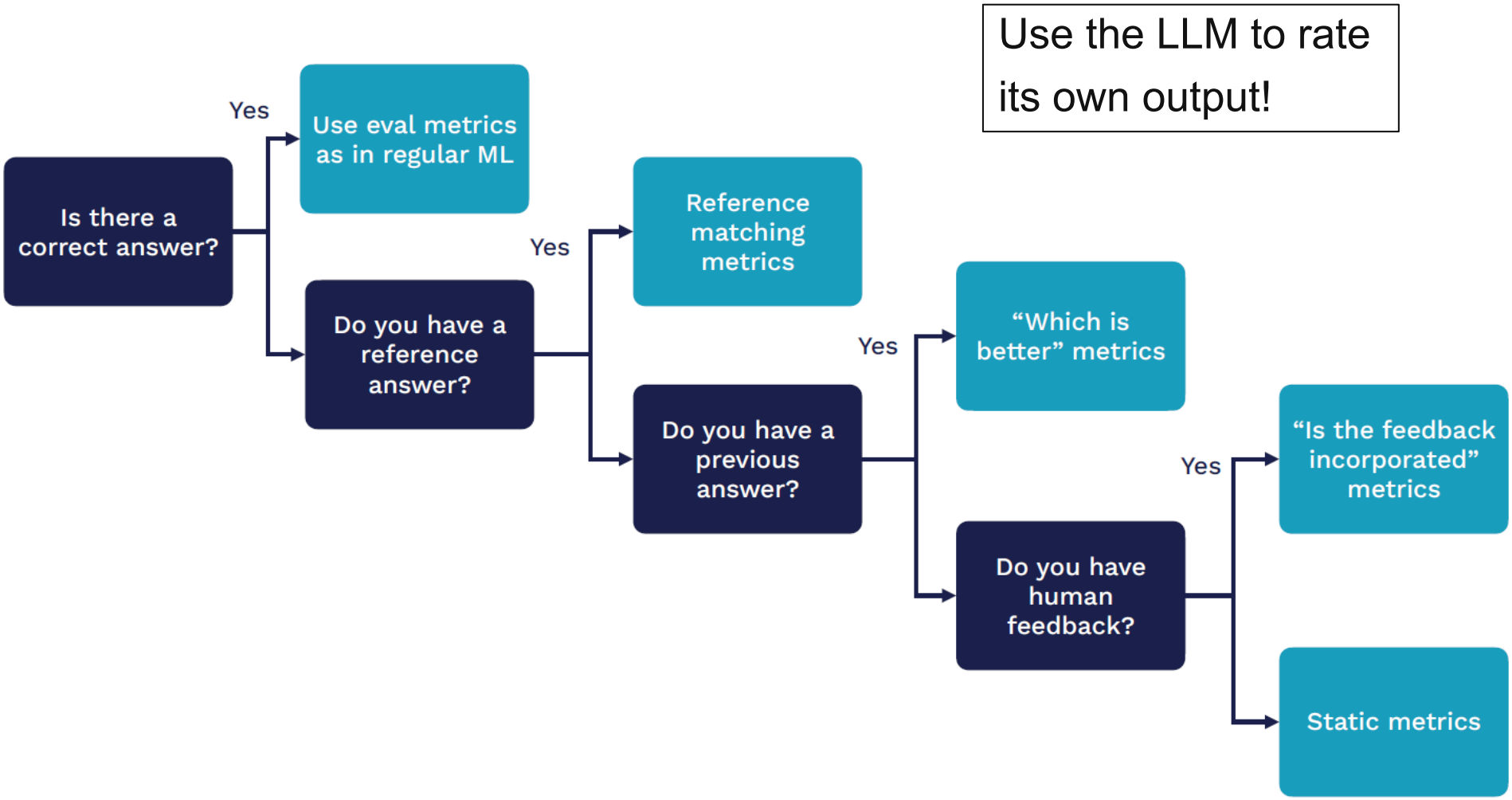
Fine-tuning an OSS model

Training an OSS model from scratch

Building a custom model from scratch

@transitive_bs

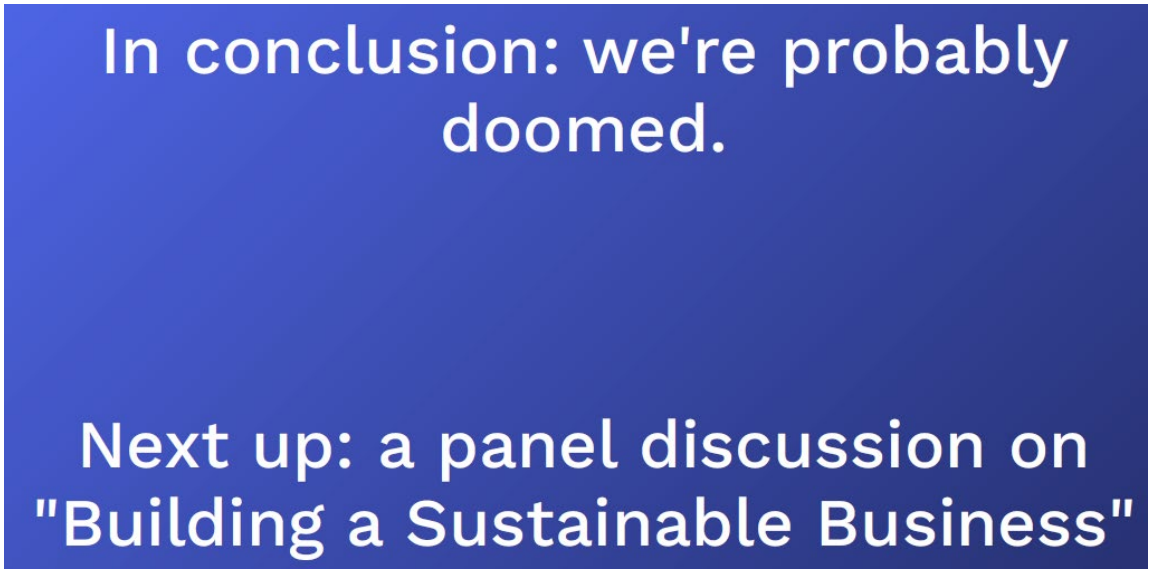
Testing an LLM



Security Risks

- Prompt injection
  - Added dimension with plugins
- Jailbreaking
- Connecting increasingly powerful agents to world-affecting systems
- Not mentioned by bootcamp: social destabilization, job loss

Probably my favorite slide of the bootcamp



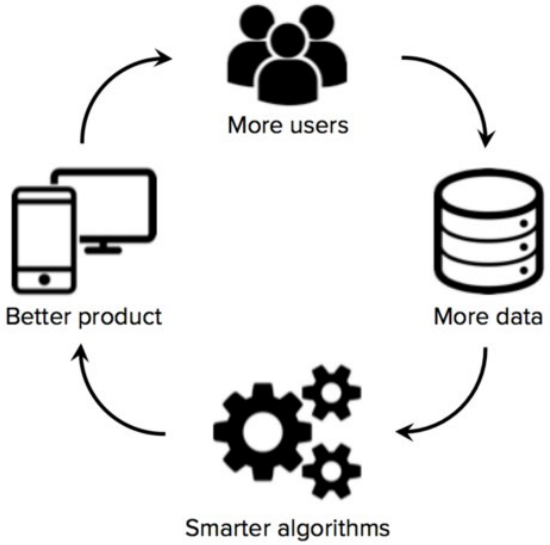
Major Bottlenecks for Scaling LLMs Further

- Training costs range from $10M to $100M, not an obstacle for large corporations or governments
- GPUs are limited temporarily
  - Microsoft employee over lunch: “We are setting up Azure OpenAI API endpoints as fast as we can buy NVIDIA cards”
- However, high quality data is hard to find online at the scale needed
  - Remedies might include dredging offline datasets, using LLM-generated data, or going multimodal

Uneasy Competitive Landscape

- Everyone is using the same limited number of LLMs, so what prevents you from getting copied?
- LLMs can generate prompts and annotations
- Discussion of rush jobs causing hazards (Bing Chat)
- Some companies are differentiating by fine-tuning, others by supporting enterprise compliance (i.e., AskSage)

Ideal self-reinforcing loop



What Is Coming in the Near Future?

- Maturation of prompt engineering: Ops tooling, marketplaces, training
- Improvement of open alternative LLMs in quality and resource costs
- Massive multimodal models
- A 10-100x increase in context lengths

> Following research used to be like **drinking from a firehose**, now it's more like **the whole town is here**, and **everyone brought their firehoses**.
> - anonymous quote

Demo Garden (Peruse at Your Leisure)

- Automatically generate code documentation
- Create flash cards from any text and get personalized quizzes
- A Reddit analogue for sharing prompts
- Auto customize your resume to job postings
- Semantic video search and classification
- Draft IRB applications or NIH grant research strategies
- Another digital cloning startup
- Generate key ideas for your website and write blurbs

More Resources

- All course lecture slides are on share drive under “LLM Bootcamp”
- Check out the demos
- Compare different LLMs yourself
- Lightning AI’s article on LLM comparisons
- Wolfram’s deep dive into LLMs
- Code an LLM from scratch

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (DD-MM-YY)	2. REPORT TYPE	3. DATES COVERED (From – To)
00-07-23	Non-Standard	

4. TITLE AND SUBTITLE

What's New in AI: Findings from the Full Stack Deep Learning (FSDL) Large Language Model (LLM) Bootcamp

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBERS

6. AUTHOR(S)

Nicholas A. Wagner

5d. PROJECT NUMBER

C5240

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESSES

Institute for Defense Analyses
730 East Glebe Road
Alexandria, VA 22305

8. PERFORMING ORGANIZATION REPORT NUMBER

NS D-33547

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
730 East Glebe Road, Alexandria, VA 22305

10. SPONSOR'S / MONITOR'S ACRONYM

IDA

11. SPONSOR'S / MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Project Leader: Kevin Garrison

14. ABSTRACT

This presentation summarizes findings from the Full Stack Deep Learning Large Language Model (LLM) Workshop held April 21–22, 2023. I begin with a definition of large language models and describe how to select the best one for your use case. I then provide best practices for using large language models, examine technical limitations, and speculate on near-term trends for these models. I conclude with references to online educational materials.

15. SUBJECT TERMS

Generative AI, Machine Learning, Large Language Models

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Institute for Defense Analyses
Unclassified	Unclassified	Unclassified	Unlimited	25	19b. TELEPHONE NUMBER (Include Area Code)