Use Cases for Large Language Models (LLMs)

Using LLM to support OUSD A&S

SLIDES ONLY

NO SCRIPT PROVIDED

Agenda

- Tasking & Background
- LLM 101
- Current state of LLM Usage by ADA
- Discussion: Use Cases
- Expectation Management
- Recommended Way Ahead

CLEARED FOR PUBLIC RELEASE

Tasking & Background

- Task Force Lima is responsible for monitoring, developing, evaluating, and recommending the responsible and secure implementation of generative AI capabilities across the DoD
- ADA was tasked with identifying use cases that apply Large Language Models (LLMs) that might support A&S initiatives and objectives
- A&S Priorities:
  - Deliver Integrated Capabilities at Speed and Scale
  - Protect and Sustain the Force
  - Foster a Resilient and Robust Industrial base

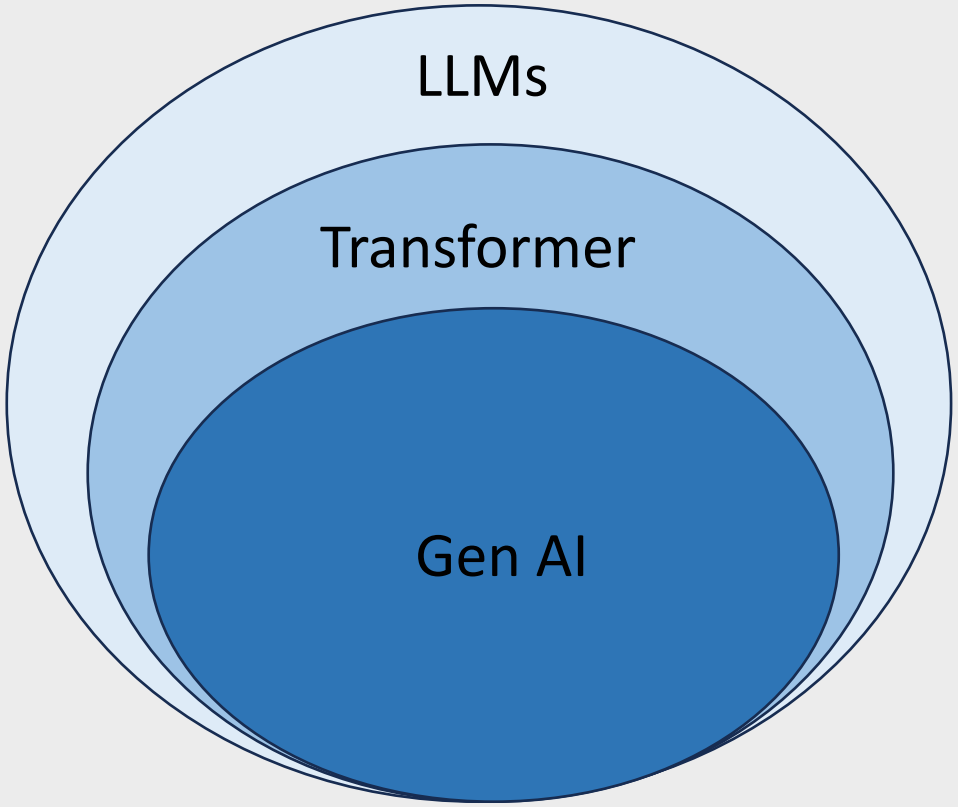
CLEARED FOR PUBLIC RELEASE

Priorities of OUSD A&S: https://www.acq.osd.mil

Large Language Models 101

A Large Language Model(LLM) is a type of artificial intelligence algorithm with a larger number of parameters to process and understand text using self-supervised learning techniques.[1]

- Large Language Models are a broad set of Artificial Intelligence algorithms aimed at modeling language
  - LLMs can utilize different architectures: RNNs, LSTMs, Transformers
  - Some LLMs are trained to perform a specific task, while others are trained to perform many tasks
- The Transformer architecture is the most common and best performing LLM architecture used today
  - Well known Gen AI models like Google Gemini and OpenAI ChatGPT use the Transformer architecture
- Generative Tasks are a subset of tasks LLMs can perform
  - Generative Tasks include Text Generation, Image Generation, and Music Generation



[1] For a general overview, see https://www.geeksforgeeks.org/large-language-model-llm/

Current State of LLM Usage in ADA

- LLM have been used through ADVANA platform
  - LLM have been run in Databricks
  - Pretrained models
  - 3 Models in data bricks

- Proof of concept:
  - Tagging data
  - Question answering
  - Using multiple choice question and answering
  - Summarization
  - Sentiment Analysis

- Model Pedigree
  - Hugging face (https://huggingface.co/models)

CLEARED FOR PUBLIC RELEASE

Available LLM Models in Advana

- Advana offers a LLM Repository called the Foundational Model Zone
  - The repository consists of around 45 different pre-trained LLMs
    - A pre-trained model is a model that can be used right away since it was already trained on a large dataset to perform a specific task
  - Models are available to copy to a local repository and fine tune them for specific data and use cases
    - Fine-tuning is the process of making small adjustments to a pre-trained model to improve its performance on a specific task
- LLMs offered include both Generative and Non-Generative
  - Generative Models include: OpenAI GPT, Databricks Dolly, and Microsoft's Phi-2
  - Non-Generative Models perform specific tasks like Question Answering, Sentiment Analysis, and Summarization

CLEARED FOR PUBLIC RELEASE

Possible Use Cases

- Team developed five possible Use Cases for further refinement

1. Question Response
2. Requirements Document Generation
3. Acquisition Information Exposure with “What if?” Analysis
4. Vendor Performance (General)
5. Vendor Performance in Conflicts

CLEARED FOR PUBLIC RELEASE

Use Case #1: Question Response

- First level response to inquires from Senior Executives or external organizations (such as GAO)
  - Reads in information requests and categorizes the type of request in terms of importance, urgency, and protocol
  - Drafts communication for organizational taskers
  - Drafts responses (email and documents)
- Purpose:
  - Drive efficiencies by reducing organizational response time to question asked on weapon systems
- Considerations
  - Requires tailored data (GAO and Congressional investigations)
  - Requires access to organizational information systems (for both trained models and organization specific data repositories)
  - Need to understand limitations of pre-trained models

CLEARED FOR PUBLIC RELEASE

Target Use Case: Question Response

Goal: Provide a tool for Program Managers and Acquisition Analysts to aid in acquiring program specific data and perform basic analysis using simple text requests

- Easily query program specific schedule and unit cost data
- Simple user interface
- Can provide basic analysis and statistics, i.e., mean or median unit cost growth per year

CLEARED FOR PUBLIC RELEASE

Preparing Data for LLM Training

- For the LLM to learn a desired task, it must be given a dataset with expected inputs and outputs
  - In the case of the Question Response task, it needs to know what a text request would look like and what a correct response would be for that request
  - Since most of the Acquisition data is stored in relational databases, the output would be a SQL query to use to obtain the correct data
- To produce the dataset, template for text requests and SQL queries were filled in by iterating over different program names and schedule events, and adding them to the text request and SQL query

Text Request Input Template

"Show me the latest {Fill in Appropriation Category} budget data for {Fill in Program Name}"

SQL Query Output Template

"SELECT program_short_name, budget_appns_appn_category, total_budget_estimate_amount, budget_year, position_full_name, cost_comparison_by FROM acquisition_workspace.pps_budgets_appn_ism WHERE program_short_name = {Fill in Program Name} AND effective_date = (SELECT MAX(effective_date) FROM acquisition_workspace.pps_budgets_appn_ism WHERE program_short_name {Fill in Program Name}) AND budget_appns_appn_category = {Fill in Appropriation Category} AND budget_appns_is_ty_amounts = False "

Fine Tune LLM

- Copied a pre-trained Google T5 LLM from Advana's Foundational Model Zone
  - Picked model based on performance of language translation since this will be a translation between English to SQL
- Prepared Dataset for Training
  - Text Requests and SQL Queries must be converted to tokens before training
  - Additional preprocessing to make sure all tensors of tokens have the same length
- Fine-tuned Model
  - 1 Epoch of training
  - Achieved a categorical cross-entropy loss of 0.088

Full Process

1. Provide Request Text String

2. Convert Text into Tokens

3. Generate model output based on input tokens

4. Convert Tokens back into Text, which is now a SQL Query

5. Execute SQL Query and Produce Results

CLEARED FOR PUBLIC RELEASE

Use Case #2: Document Generation

- Given a description of a threat, scenario, and current national security capabilities, generate a draft of the Initial Capabilities Document (ICD) and Capability Development Document (CDD)
  - Summarize threat information
  - Summarize current capability documentation
  - Generate formatted documents
  - Generate threshold and objective requirements
  - Provide an 80% solution to reduce document editing for formats and structure and increase discussions on requirement development
- Purpose
  - Scale ability to develop capabiltiy documents to ensure needed capabilities are articulated and provided to the warfighter
  - Support development of integrated, reinforcing, and strategically redundant capabilities to ensure warfighter have needed capabilities across multiple potential future environments
- Considerations:
  - Model requires text generation and predictive requirements development
  - Pre-trained versus trained models
  - Combining LLM for images and text

CLEARED FOR PUBLIC RELEASE

Use Case #3: Acquisition Information Exposure with "What if?" Analysis

- Use current pre-trained LLM, such as Bard or ChatGPT, to assess Department of Defense acquisition information that is available through existing public models
  - Identify information which can be obtained on US weapons systems through public LLM
    - Apply prompt engineering to a suite of public models and an exemplar set of weapons systems to derive weapon system requirements available through public LLM
    - Assess the usage of prompt engineering to validate weapons system information
  - Assess the available information on acquisition processes, such as schedule and cost information
- Purpose
  - Assess risk to industrial base and supply chain due to LLM capability to consolidate public data
- Considerations
  - When using publicly available systems, it is assumed all submitted questions are saved

CLEARED FOR PUBLIC RELEASE

Use Case #4: Vendor Performance

- Create a question-answer system based on public data (press releasees, new sites, vendor sites) and controlled data (capability documents, policy, and DoD priorities, additional client information) to articulate risk to weapon system development
  - Assess risk relative to DoD prioritization
  - Articulate difference between system usage during execution versus expected performance
- Purpose
  - Understand performance of vendor systems were used post deployment in the context of delivery, utilized capability, and capability demand
  - Identify data sources (public and controlled) containing system performance information
- Considerations
  - Access is needed to documentation related to capabilities and requirements
  - Performance can be measured with CPARS data and Testing and Evaluation Data

Use Case #5: Vendor Performance during Conflicts

- Build an LLM to automatically assess vendor performance during a conflict based on acquistion data and operational information
  - Generate reports on performance, cost, and schedule based on documentation of system capabilities, testing and evaluation, and system deliveries
  - Identify flow of information on systems and challenges unique to acquistion of systems during a conflict
  - Show value of DoD investments during a conflict
- Purpose
  - Assess performance of vendor systems during conflicts
    - Industrial base capability to surge
    - Delivery capability
    - Demand for new capabilities versus current capabilities
    - Usage of systems in conflict areas
- Considerations:
  - Access to data and sufficient data to train an LLM
  - Rate of update to the data implies models will need to be regularly updated

Setting Expectations

- Accurate LLM require training data, compute capabilities, and a development team
- Data needs
  - The magic of OpenAI is the human feedback mechanism, as such, often LLMs need a ton of human intervention in data preprocessing
  - LLMs need lots of data. For example, ChatGPT-4 was trained on 570GB of text datasets
- Compute needs
  - LLMs are computationally expensive, needing GPUs to train
  - GPT-2 was trained on 40GB of text data costing $43,000, and later ChatCPT-4 cost $12 million to train

Recommended Way Ahead

- Down select use cases for entry to TF Lima Use Case Form
- For the selected use cases:
  - Document use case description
  - Further refine use cases and considerations
  - Develop a project plan for each use cases containing:
    - Data requirements to build an LLM or LLM application
    - Expected capabilities for each LLM application
    - Resources
    - Schedule
- Observations
  - Due to the compute needs for LLM, a TF Lima environment (compute and data storage) may be useful to better understand challenge on training LLM and accuracy of LLM

CLEARED FOR PUBLIC RELEASE