AIRC

ACQUISITION INNOVATION
RESEARCH CENTER

AFIT
Air Force Institute of Technology
Carnegie Mellon University
NPS
The George Washington University
Washington, DC
University of Massachusetts Amherst
NC State University
University of South Florida
The University of Arizona
Georgetown University
Texas A&M University
The Ohio State University
University of Maryland
Virginia Tech
UCF
North Carolina Agricultural and Technical State University
UAH
The University of Alabama in Huntsville
Stevens
Auburn University
Georgia Tech
University of Virginia
Purdue University
MIT
Massachusetts Institute of Technology
George Mason University
Old Dominion University
USC
PennState

Test and Evaluation of Large Language Models to Support Informed Government Acquisition

Presented by Dr. Erin Lanus and Dr. Jaganmohan “Jagan” Chandrasekaran

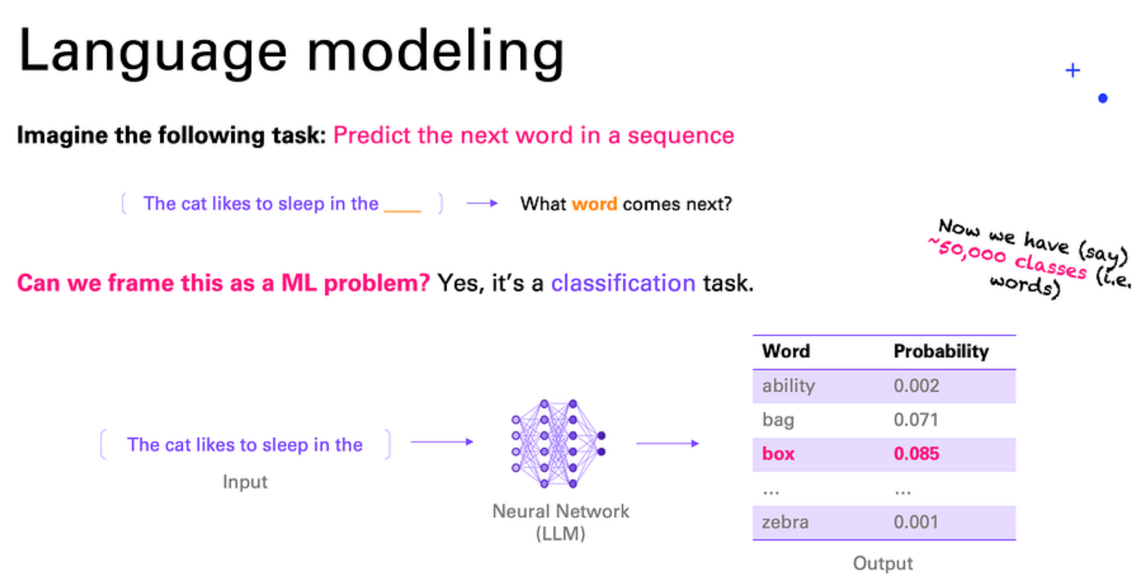
22nd Annual Acquisition Research Symposium & Innovation Summit
May 07, 2025

Coauthors: Mr. Brian Mayer, Dr. Heather Frase, Dr. Patrick Butler, Dr. Stephen Adams,
Mr. Jared Gregersen, Dr. Naren Ramakrishnan and Dr. Laura Freeman

Acknowledgement

This material is based upon work supported, in whole or in part, by the U.S. Department of Defense (DoD) through the Director of Operational Test and Evaluation (DOT&E) and the Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) under Contract HQ003424D0023. The Acquisition Innovation Research Center (AIRC) is a multi-university partnership led and managed by the Stevens Institute of Technology through the Systems Engineering Research Center (SERC) – a federally funded University Affiliated Research Center. Any views, opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Government (including the DoD and any government personnel).

Introduction to Large Language Models



Language modeling

Imagine the following task: Predict the next word in a sequence

(The cat likes to sleep in the ___) → What word comes next?

Now we have (say) ~50,000 classes (i.e. words)

Can we frame this as a ML problem? Yes, it's a classification task.

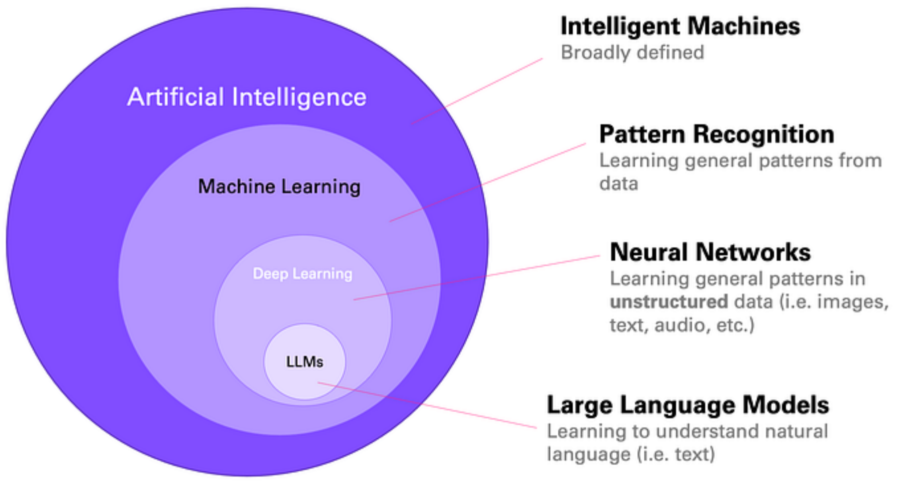
(The cat likes to sleep in the) → Neural Network (LLM) →

Word	Probability
ability	0.002
bag	0.071
box	0.085
...	...
zebra	0.001

Input

Output

Image source: https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f



Intelligent Machines
Broadly defined

Pattern Recognition
Learning general patterns from data

Neural Networks
Learning general patterns in **unstructured** data (i.e. images, text, audio, etc.)

Large Language Models
Learning to understand natural language (i.e. text)

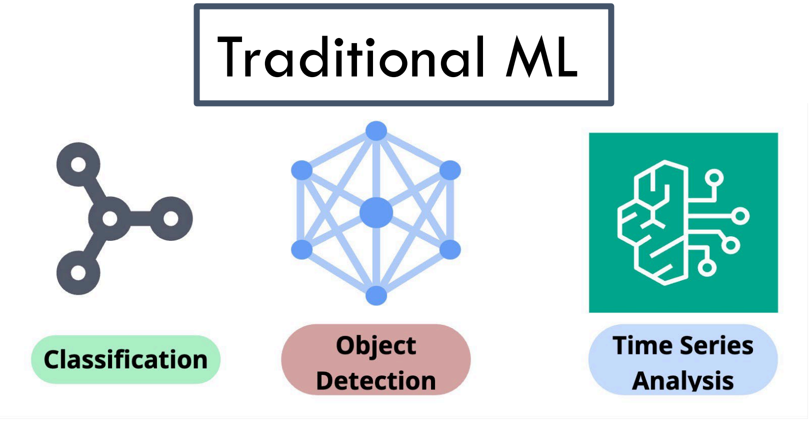
ChatGPT Gemini Llama MISTRAL AI_ GPT-4

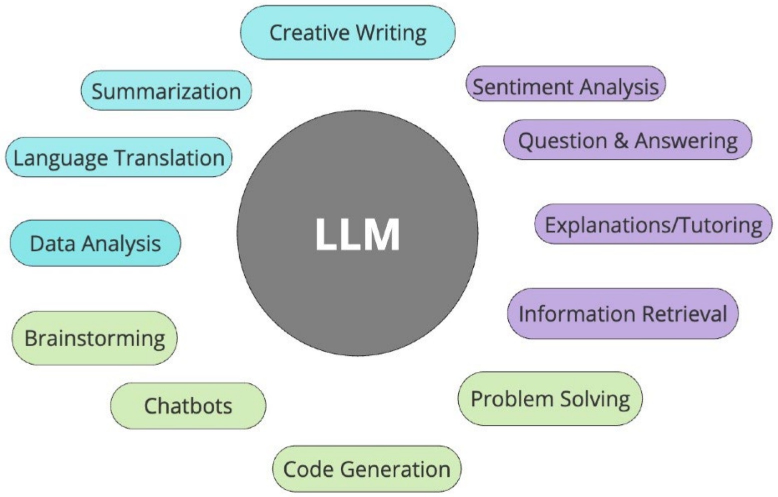
Motivation

LLMs offer flexibility in performing diverse functions, distinguishing them from traditional AI/ML systems.

LLM's versatility is great, but comprehensive test and evaluation (T&E) are key to ensure reliable, trustworthy, and safe behavior.

The ability to do many different functions increases the difficulty and the necessary variability in testing LLMs.



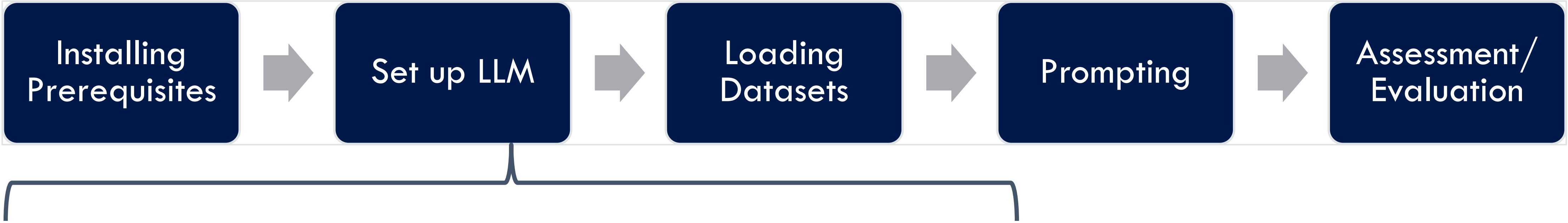


What does the current T&E landscape inform us about the evaluation of LLMs?

T&E of LLM - Overview

T&E Objective: Can an LLM generate correct, contextually relevant responses?

Steps in testing LLM



Access mode - To perform inferencing, practitioners either

- Host the LLMs locally
- Interact via Application Programming Interface

Parameters – A set of values influencing the LLM's outcome

- Temperature
- Top-p
- Max tokens
- Frequency penalty

Prompt - A set of instructions informing the LLM about the user's request

- Zero-shot prompting
- Few-shot prompting
- Chain-of-thought prompting

T&E of LLM - Overview (2)

Capabilities

- Abstract functional abilities of an LLM
- *Examples: understanding, reasoning, generation*

Tasks

- Concrete implementations used to assess specific capabilities
- *Examples: Question Answering, Multiple Choice Question, Code Generation*

Benchmarks

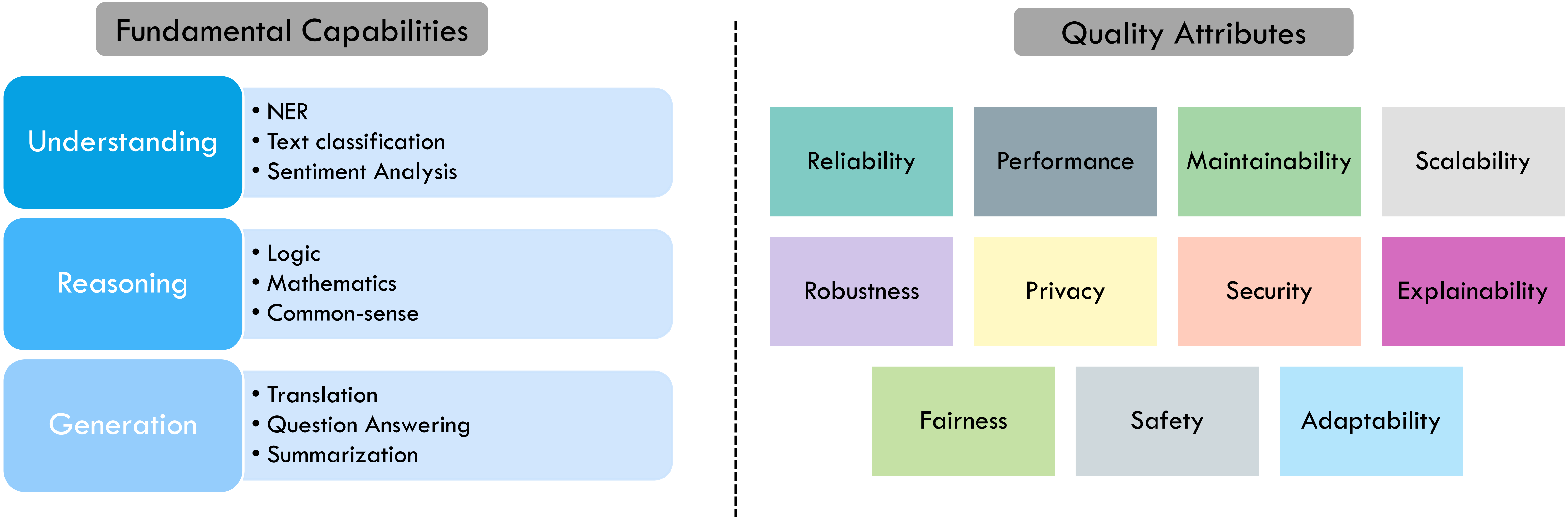
- Standardized datasets that measure performance
- *Examples: MMLU, HELM*

Capability	Task Type	Benchmarks	Objective
Understanding	Named Entity Recognition	CoNLL 2003	Evaluate LLM's basic word-level understanding and categorization abilities
Reasoning	Multiple Choice Question	MMLU	Assess high-school level reasoning abilities on variety of subjects
Generation	Code Generation	HumanEval	Test LLM's ability to generate software code.

LLM Evaluation Framework

A comprehensive evaluation of LLMs must include two primary dimensions:

- Evaluation of **fundamental capabilities** in facilitating human-like interactions
- As a software component, the LLM's ability to meet expected **software quality standards**



Fundamental Capabilities

Understanding
- NER
- Text classification
- Sentiment Analysis

Reasoning
- Logic
- Mathematics
- Common-sense

Generation
- Translation
- Question Answering
- Summarization

Quality Attributes

Reliability

Performance

Maintainability

Scalability

Robustness

Privacy

Security

Explainability

Fairness

Safety

Adaptability

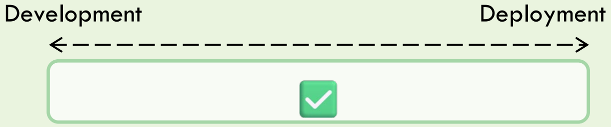
Example Acquisition Scenarios

Use case 1 - Identify named entities in a user-provided collection of records, and extract relationships between entities.

Use case 2 – Extract information from user-provided records and question responses.

Use case 3 - The drones have an LLM that converts text messages into commands that they can implement.

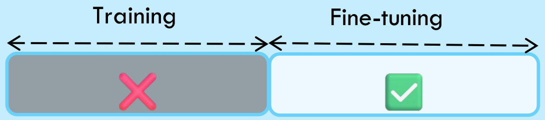
LLM developed in-house



Access to artifacts

Testing frequency - **Continuous**

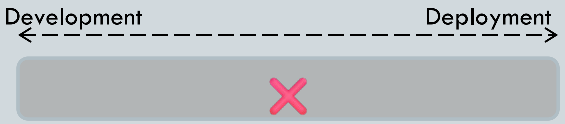
Off-the-shelf LLM fine-tuned in-house



Access to artifacts

Testing frequency - **Periodic**

Off-The-Shelf LLM



Access to artifacts

Testing frequency – **One-time acceptance testing**

Understanding, Reasoning, Reliability, Scalability

Security

Generation

Performance

Adaptability

Maintainability

Fairness

Robustness

Limitations and Recommendations

LIMITATIONS

1. Benchmarks have limited evaluation scope
2. Aggregate metrics – limited understanding of LLM's behavior
3. Immature T&E practices
   - Test adequacy
   - Failure modes

RECOMMENDATIONS

1. Ability to test LLM for specific operational contexts
2. Development of new metrics for granular understanding
3. Systematic T&E Process for LLM evaluation
4. Research on system-level and post-deployment T&E challenges

Thank you!