VT National Security Institute

Intelligent Systems Division

Using Large Language Models to Accelerate Development of Complex System

Paul Wach, PhD, Research Assistant Professor

Brady Jugan, Undergrad Research Assistant

Scott Lucero, Research Scientist

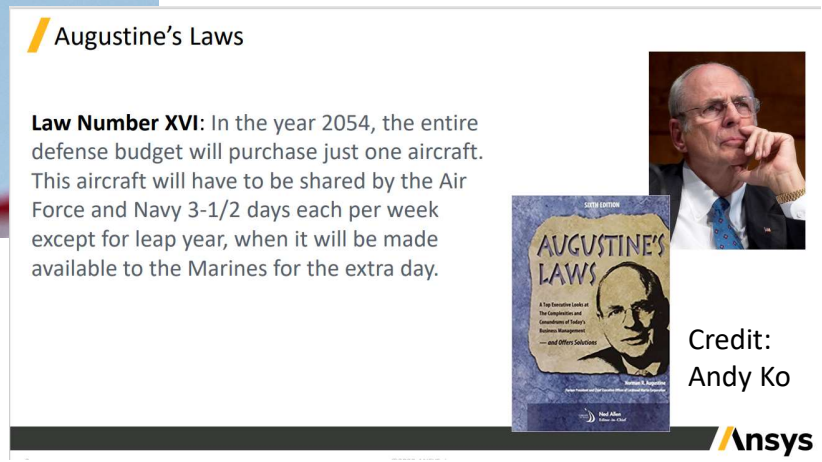
18 Sep 2024

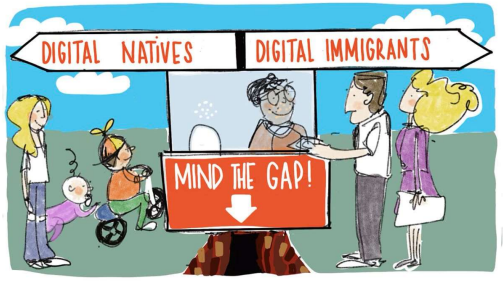
Setting the Stage

- Challenge:

  From empirical evidence and individual experience, **our current approach is not sufficient**

- Example Solution(s):
  - **Digital engineering (DE)**: connecting the right data right to enable effective and efficient decisions and communication
  - **Model-based systems engineering (MBSE)**: the application of DE to enhance systems engineering (SE)





Spectrum of workforce

NATIONAL SECURITY INSTITUTE
VIRGINIA TECH

Generative Artificial Intelligence (GenAI) & Digital Transformation

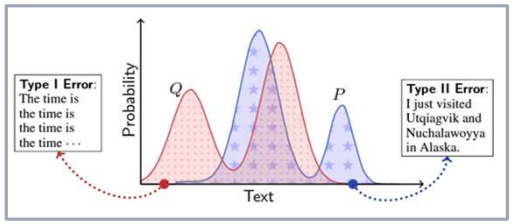
- Challenge:
  - **Adoption** of digital engineering has been **slow**er than expected and the benefits have not yet been realized
- Goal:
  - **Expedite & reimagine** the digital transformation
- Large language models (**LLM**) and the systems modeling language version 2 (**SysMLv2**)
  - Serve as a **workforce bridge** between seasoned generation and incoming digital natives, among other applications
- Thrusts
  1. Text to text
  2. Text/image to SysMLv2 code
  3. SysML image/code to text

Text to text

Human expert

vs

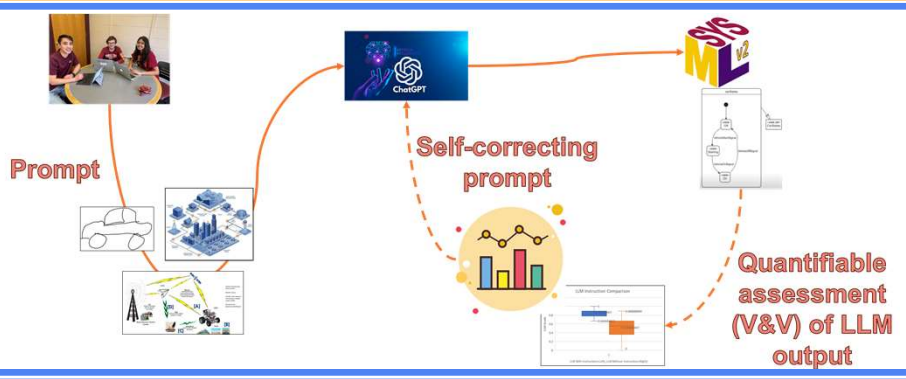
LLM



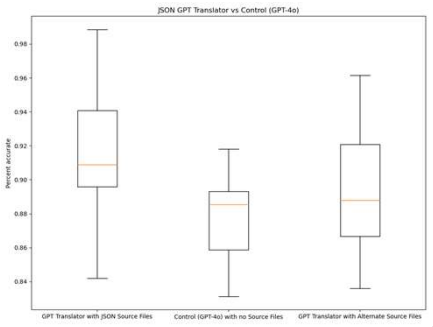
	MAUVE Scores		
Model	Prompt 1	Prompt 2	Prompt 3
GPT-4 (OpenAI)	0.0000	0.0000	0.9137
GPT-3.5 Turbo (OpenAI)	0.0000	0.0001	0.9749
Claude (Anthropic)	0.0000	0.0003	0.9932

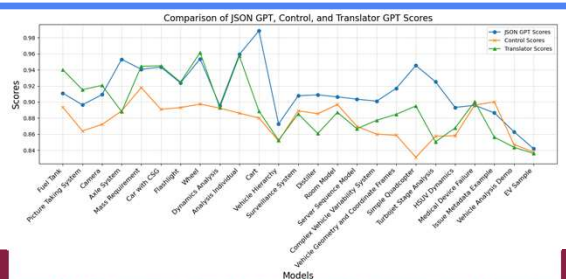
Text/Image to SysMLv2

Notional Prompt with self-correcting V&V of output



SysMLv2 image to text





NATIONAL SECURITY INSTITUTE
VIRGINIA TECH

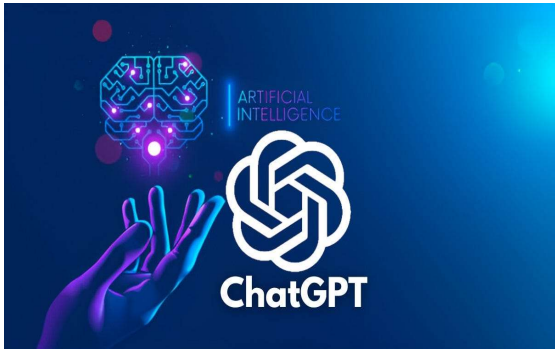
Methodology (Part 1 of 4) – Artifact Generation

LLMs used for Generation:

- ChatGPT – Highly customizable with lots of features to optimize and large number of parameters

Optimizations Performed:

- Fine tuning (SysMLv2 keywords to example diagrams from the SysMLv2 repo) using JSONL format
- Chain-of-Thought Prompting
- Knowledge base (txt files for context)



Gemini

Claude 3

Methodology (Part 2 of 4) – Artifact Analysis

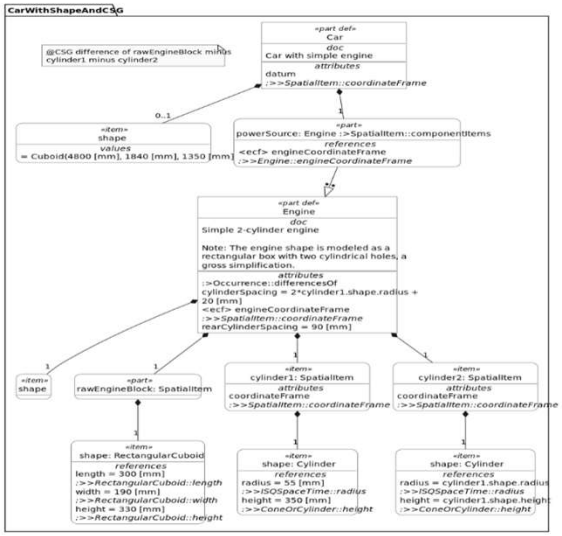
LLMs used for Generation:

- ChatGPT – Highly customizable with lots of features to optimize and large number of parameters
- Claude – Trained on a large number of parameters
- Gemini – Trained on a large number of parameter sand highly customizable

Requested GPT analyze artifacts to provide a textual description of the given system in a manner that is understandable to people who are not subject experts

Optimizations Performed:

- Fine tuning (Providing textual examples to LLM) using JSON format
- Knowledge base (txt or SysML files for context)



Describe this model in plain english and give as much detail about the relationships, attributes, etc. Do this in a paragraph format without using variable and package names. Avoid using SysMLv2 lingo altogether and translate the variable/relationship names into the context of the system description.

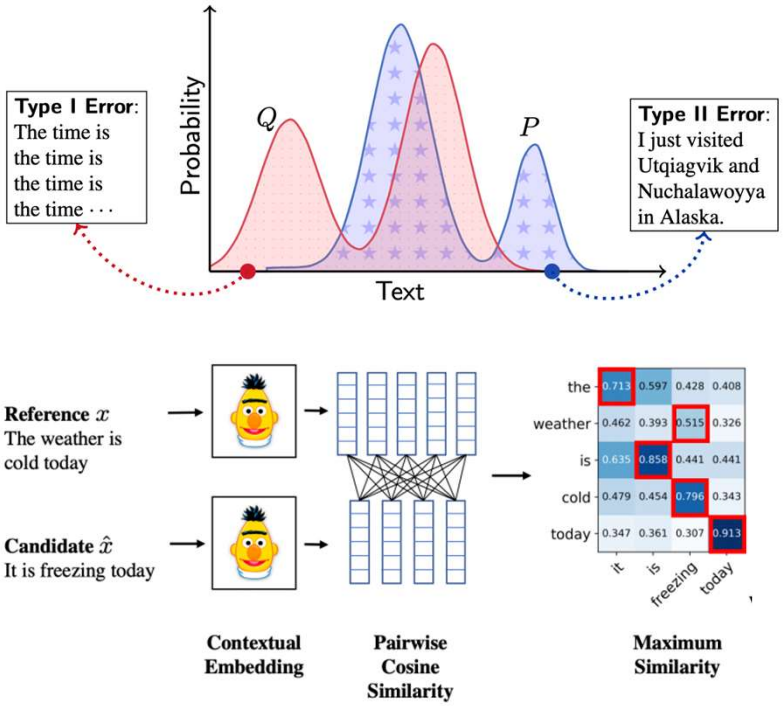
Methodology (Part 3 of 4) – Evaluation Metrics

For Artifact Generation:

- Syntax Score (Provided by Equations on the right)
- Logic Score (Qualitive human score with 3 quantified categories, 0 (insufficient system), 0.25 (partially sufficient system), 0.5 (sufficient system))
- Overall score, α, with syntax weighted double as much as logic

For Artifact Description:

- MAUVE Score
- BertScore
- 2 Sample T-Test



$$Syntax\ Score = 1 - \frac{Number\ of\ Syntax\ Errors}{Number\ of\ Lines\ of\ Code}$$

$$\alpha = \frac{Syntax\ Score\ + Logic\ Score}{1.5}$$

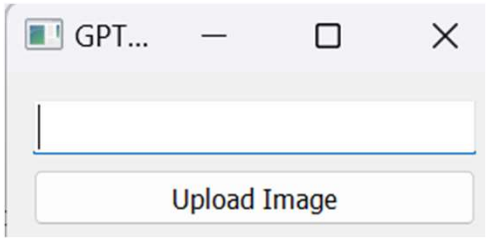
Methodology (Part 4 of 4) – Workflow Automation

For Artifact Generation:

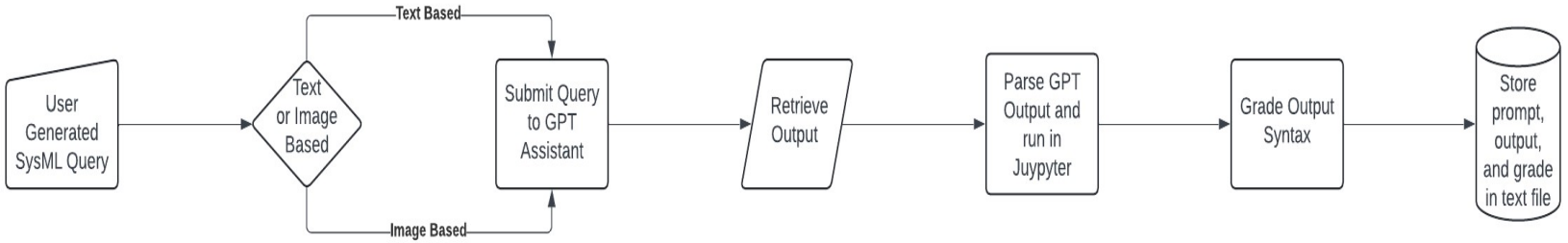
- Syntax Score (Provided by Equations on the right)
- Logic Score (Qualitive human score with 3 quantified categories, 0 (insufficient system), 0.25 (partially sufficient system), 0.5 (sufficient system))
- Overall score, α, with syntax weighted double as much as logic

For Artifact Description:

- MAUVE Score
- BertScore
- 2 Sample T-Test




```
----- TEST Run2 -----
Prompt: Generate SysMLv2 code that models a four-wheeled car in motion.
Code: package 'VehicleDefinition' {

        import ScalarValues::*;

        part def Vehicle {
                part def Transmission;
                part def CarAxle {
                        port leftAxleMounting;
                        port rightAxleMounting;
                }

                part def Wheel;

                // Specific vehicle contents
                part theCar :> Vehicle {

                        // Parts
                        part theTrans :>
                                theCar.Transmission;

                        // Completions
                        part rightFrontWheel : Wheel in
                                theCar.AxleMount.rightAxleMounting;
                }

        }

}

Errors: stderr ERROR:no viable alternative at input 'in' (1.sysml line : 22 column : 33)
ERROR:no viable alternative at input '.' (1.sysml line : 23 column : 11)
ERROR:no viable alternative at input '.' (1.sysml line : 23 column : 21)

Syntax Score: 89.0%
Image File:
```



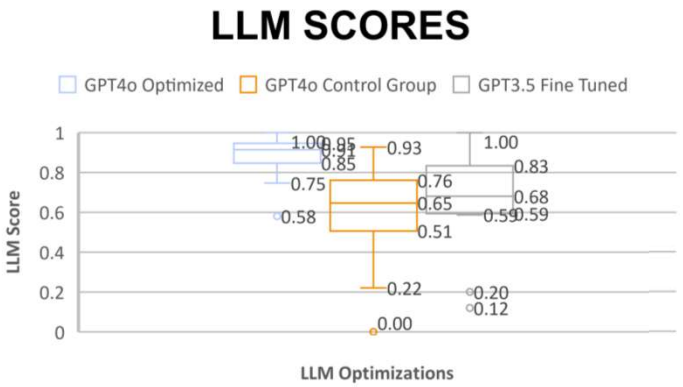
Results (Part 1 of 2) – LLM Artifact Generation

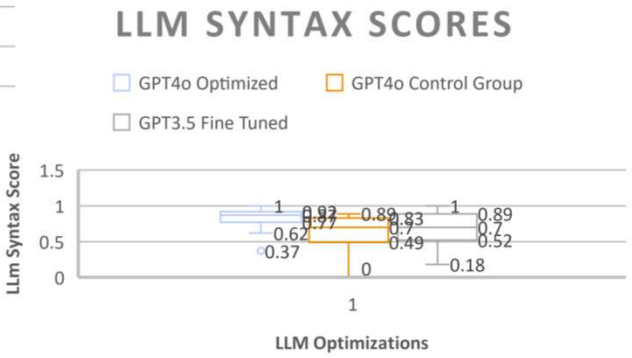
Statistics:

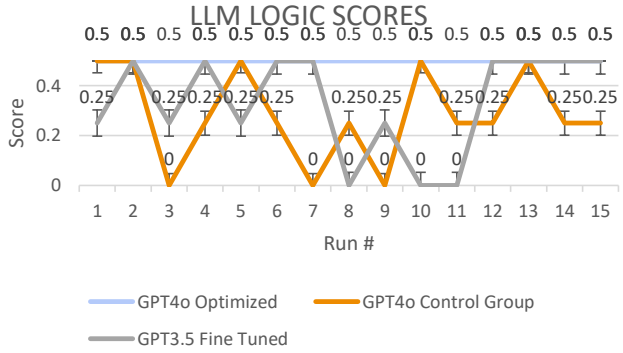
- 15 given text-based prompt querying for systems designed for a defense-specific audience
- Output syntax, logic, and overall score graded for each tested instance of GPT (GPT 4o optimized, GPT 4o control, and GPT3.5 fine tuned)

Results:

- The instance of GPT 4o given context, chain-of-thought prompting instructions, and a log of errors it historically made performed best
- GPT 3.5 when fine tuned performed second best
- The control instance of GPT 4o performed worst in all grading categories







Results (Part 2 of 2) – LLM Artifact Analysis

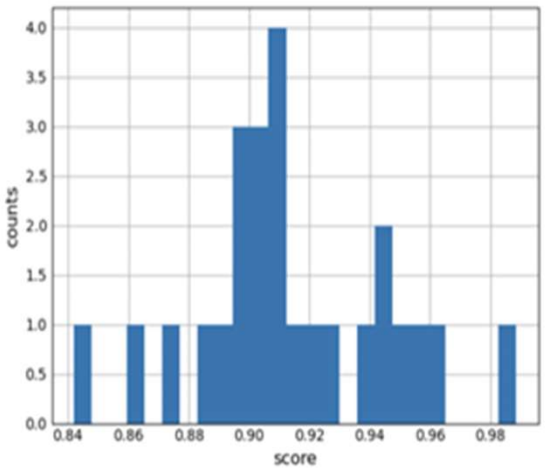
Statistics:

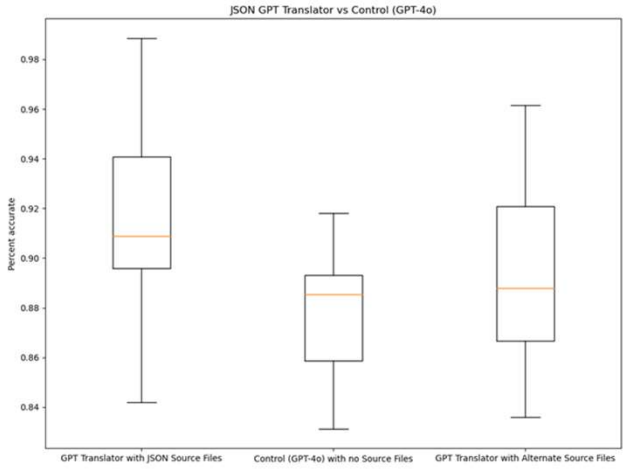
- 25 given text-based prompt querying for example system descriptions from the SysMLv2 repository
- Provided MAUVE and BertScore two sample t-test results

Results:

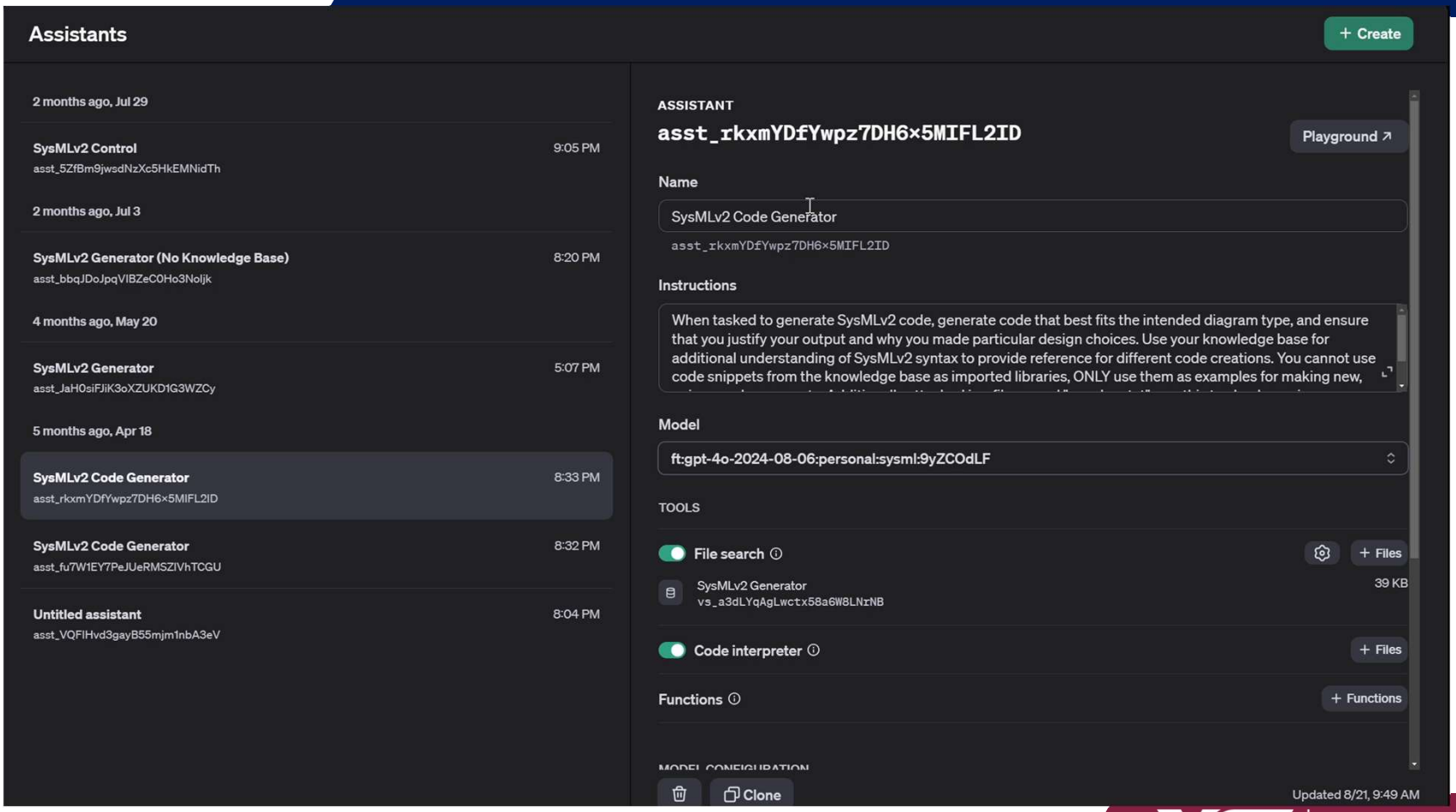
- Mean score of fine tuned GPT translator showed a statistically significant difference in that of the control group for the BertScore
- MAUVE showed indeterminate results

```
In [10]:
1 import torch
2 print(f"The JSON GPT on average was: {100 * torch.mean(F1):.2f} % accurate to the correct description.")

The JSON GPT on average was: 91.37 % accurate to the correct description.

In [11]:
1 import matplotlib.pyplot as plt

In [12]:
1 plt.hist(F1, bins=25)
2 plt.xlabel("score")
3 plt.ylabel("counts")
4 plt.show()
```





DEMO Video



NATIONAL SECURITY INSTITUTE
VIRGINIA TECH

What's on the Horizon?

Trajectory

- ❑ Application domain(s)
  - ➢ Mission engineering, predictive maintenance, secure energy, secure cyber resilient engineering (SCRE), smart cities
- ❑ Capability growth
  - ➢ Beyond SysML (CAD, etc.)
  - ➢ LLM vs SLM
  - ➢ Team of agents
  - ➢ Relational versus graph database
  - ➢ **Scaling**
    - ✓ CUBE (NSI GPU cluster)
    - ❖ CREATE (NSI cyber range)
  - ➢ Systems Theoretic Advisor

Scaling to CUBE environment

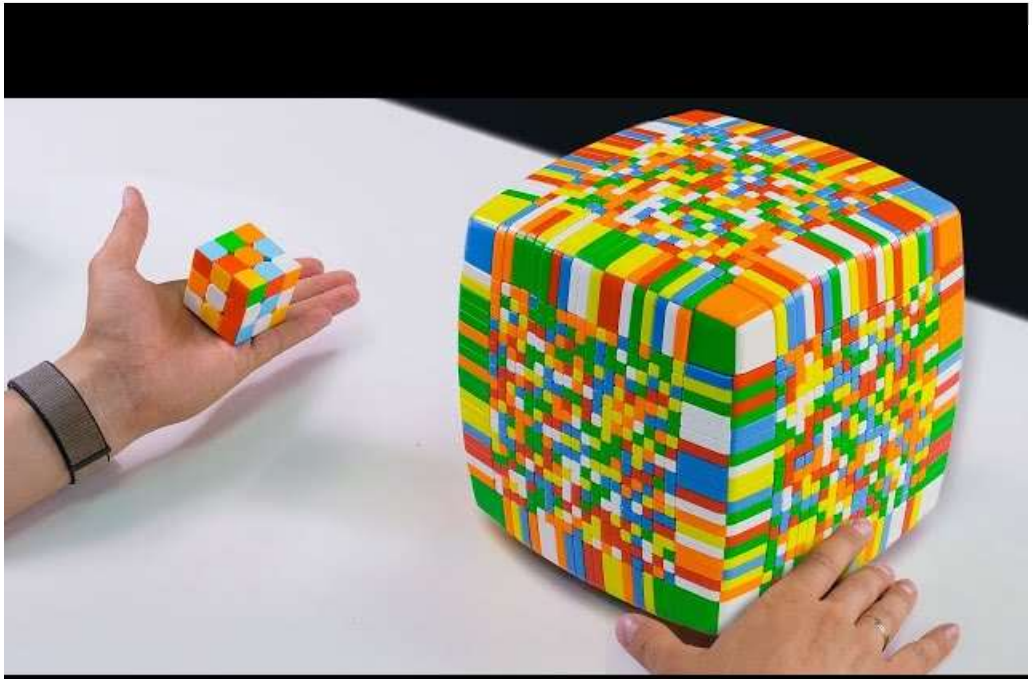
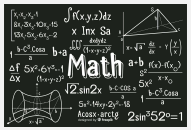


Image source credit

*CUBE = …
**CREATE = Cyber Research Environment and Threat Evaluation

Mathematical Underpinning to Digital Transformation (e.g., Digital T&E)

- Challenge:
  - Model-based systems engineering (**MBSE**) is qualitative (i.e., **lacking mathematical underpinning**)
- Goal:

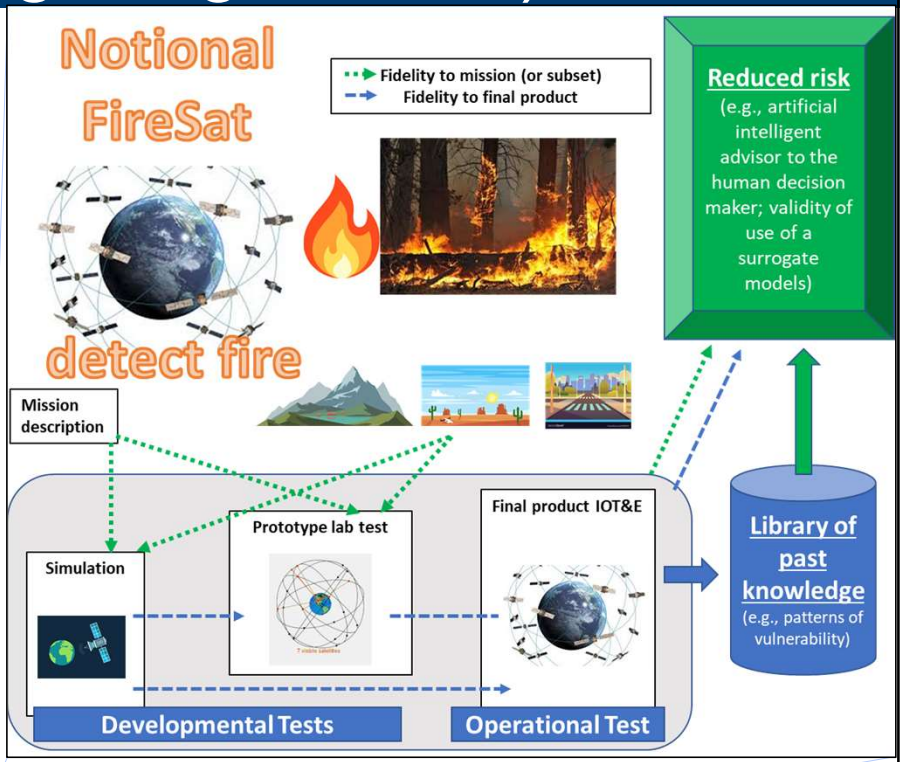


  - Develop methods and tools for Tier 3 of T&E
    - See article titled Positioning Test and Evaluation for the Digital Paradigm
- **Systems Theoretic Advisor**
  - Minimum viable product (MVP) developed by NSI funding completed in Aug 2024

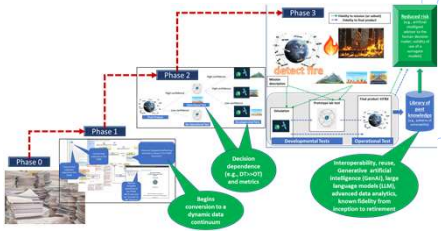
Harmonizes the data and models through mathematical underpinning

Sends that data to the machine learning (e.g., **LLM**) and computational backend to be processed

Collects that data and creates a visual of the solution to the query



Tier 3 of Digital T&E



NATIONAL SECURITY INSTITUTE
VIRGINIA TECH

Closing the Stage

"It is not necessary to change."

"Survival is not mandatory." - Deming

Current thrusts

1. System theoretic assistant
2. Text-2-text (Measured)
3. **Text-2-SysML**
4. **SysML-2-text**

Accepting nominations for naming of co-pilots

General
Leslie Groves
Director, Manhattan Project

Image from: Wikipedia

Introducing…

Image from: iStock

Questions?

Contact:
Paul Wach, paulw86@vt.edu
Brady Jugan, bradyj66@vt.edu
Scott Lucero, dslucero@vt.edu

Back-up

Virginia Tech – National Security Institute (NSI)

OUR MISSION

We meet the pressing needs of the defense and intelligence communities by developing their future workforce and advancing interdisciplinary research, technology, and policy.

NATIONAL SECURITY INSTITUTE
VIRGINIA TECH

Virginia Tech – National Security Institute (NSI)

Technical Divisions

Spectrum Dominance

- Assured and secure communications
- Advanced C4ISR and counter-C4ISR
- Quantum and heterogeneous computing
- RF machine learning
- Open Gen wireless innovation

Mission Systems

- Resilient, autonomous missions
- Remote & in-situ sensing
- Space situational awareness
- Marine autonomy and robotics
- Energetic materials

Intelligent Systems

- Data science, ML, AI
- Cyber security & complex systems
- Validation and test & evaluation
- Deep learning for sensor processing
- Data fusion and sensemaking

Facilities

Digital Test & Evaluation (T&E)

- Challenge:
  - To maintain and surpass the pace of the threat, new methods and tools are needed
- Goal:
  - Advance digital transformation of T&E
- Thrusts
  1. Model-Based Test & Evaluation Master Plan (MBTEMP)
  2. Integrated Decision Support Key (IDSK)
  3. Uncertainty propagation through the digital T&E pipeline
  4. Digital twin, connected and curated data
- Work in progress
  - Defined phases to measure progress
  - Creating methods and exemplars
  - Planning for mission assist