

## Large Language Model Enabled Generation of Systems Engineering Artifacts

Paul Wach, PhD, Research Assistant Professor,

Virginia Tech National Security Institute

**Mohammed Husain** 

Copy

Virginia Tech Graduate Student, Scale AI Solutions Architect

Taylan G Topcu, PhD, Assistant Professor

Virginia Tech Grado Department of Industrial & Systems Engineering

	Presentation to.
	1 i contation to.
	IASE & SEAAL Workshop
	AI4SE & SE4AI WUIKSHUP
	2023
	2023
	ILC ADMV SYSTEMS
X	U.J. AKMY ENGINEERING
	RESEARCH CENTER
	CENTER
sht 2023 • Wach, Husain, Top	u • All Rights Reserved

### Motivation: Complex Engineered Systems

Complexity: an emergent property of systems that arises from parts interacting in non-simple ways [Simon '62, Moses '04, de Weck et al. '11]

#### Is often blamed for scope, schedule, and cost overruns

[Crawley et al. '04, Sheard and Mostashari '10, Summers and Shah '10, Collopy and Maddox 13', Sinha and de Weck '16, Locatelli '18]

# Tesla CybThe Air Force admits the F-35 fighter jet costsAfter Mustoo much. So it wants to spend even more.

Tesla's long-awaited ( from Elon Musk, it's u

BY JOSEPH MARING PUBLISHED APR 16, 2021 Developing and procuring a brand-new nonstealth plane to save money makes sense only if the Pentagon can defy its entire history of defense spending.



cope

VIRGINIA TECH

# The Discipline of Systems Engineering is Evolving with 3 Thrusts

Bringing forth opportunities and challenges ...

### **Policy Push**

- 2018 DoD Digital Eng. Strategy
- 2019 DoD AI Strategy, [C]JADC2
- 2020 Mission Eng. Guide

### **Adversary/Industry Push**

- 2nd Cold War
- Digital Transformation

AI designers & AI for design

### **Technology Push (Digital Transformation)**

**Systems** 

Engineering

- Increased interconnectedness & Cyber-physical Characteristics
- Increasing Autonomy & Intelligence
- Digital Twins, Cloud-based computing & execution, rapid prototyping



# AI and Digital Transformation in DoD

- Acquisition and program management are critical areas where clever use of Gen AI can realize significant cost savings and value
- Cost overrun can be inferred as a function of total lifecycle cost expenditure during the system design phase
- Generative AI can augment a technical cadre of systems engineers in an increasingly cross-functional landscape

Schedule Delays and Cost Overruns Plague DOD Automated Information Systems



T-IMTEC-89-8

# AI and Digital Transformation in DoD

- Generative AI is being rapidly adopted across the DoD and Intelligence Community
- Use cases are widely applicable across every component, from staff officer augmentation to RFI/RFP generation and deconfliction
- However, true adoption is hampered by multiple factors:
  - Inherent black box nature of LLMs
  - Vendor "vaporware"
  - Expectation management w/ openand closed-source model

FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects Americans' Rights and Safety

MAY 04, 2023

#### RELEASE

IMMEDIATE RELEASE

DOD Announces Establishment of Generative AI Task Force

Sen. Schumer's first AI insight forum focuses on 2024 election, federal regulators

More than 65 senators and top tech CEOs debated openness and transparency for AI systems at the first meeting, among other key issues.



Objective

- We want to expedite the digital transformation
- Accomplished through use of generative AI, such as LLM
- Conduct brief experiment on generation of model-based systems engineering (MBSE) artifacts
- Plausibility probe with a constrained environment that provided an idea of potential value (and risk) and leads to future endeavors





# Objective: Develop a test framework for quantitative analysis of LLM outputs

#### Test environment setup

1. Select LLMs for evaluation

2. Curate representative artifacts from Bulldog dataset

3. Transform the ~7-8 artifacts into ~50 "instances" by breaking each artifact up into several "chunks."

#### **Iterative testing**

4. For each "artifact chunk," write the "prompt" that would be used to generate that segment.

5. Using a system level prompt that describes the Bulldog system, run all 50 prompts. 6. Use the MAUVE framework to evaluate semantic similarity between model responses and ground truth.

### Leveraged Datasets

### Bulldog

 Bulldog is a set of synthetic full-text systems engineering artifacts for a hypothetical unmanned ground vehicle (UGV) program. The documents include a TMRR, a CDD, an ICD, and a PMRR.

#### Why was this useful?

 The full-text documents allow for evaluation of machine-generated SE artifacts against ground truth with minimal pre-processing.

### Silverfish

 Silverfish is a dataset of open-source systems engineering artifacts focused around the field deployment of vehicular "obstacles" for perimeter security. The artifacts live in GENESYS and are retrievable via Jupyter Notebook.

Why was this useful?

 Silverfish allows for evaluation of opportunities and challenges integrating LLMs within enterprise systems.



### Step 1: Select Models

Editor Few-Shot Examples

#### System Message ()

Create a system message or leave blank to exclude

User 🚯

Internal wireless communication system is used to support communication between the sensors and the operator, and also supports fire control communications between the operator and the obstacles. The sensors include obstaclebased seismic and acoustic sensors, infrared sensors and an unmanned aerial vehicle-based surveillance system to provide warning of potential adversaries approaching the protected area. The operator is located in a vehicle, and operates within visual range of the protected area. The operator is in communication with a higher-level command and control (C2) system for exchange of doctrinal-related and situation awareness information.

Your task is to support with the generation of systems engineering artifacts. Do not include any AI-specific language such as "as an AI language model...". Make sure to use DoD-specific language and terminology to the extent possible.

-

Begin:

Input: {{ Ground\_Truth }}
Output:

bulldog_ground_tru	ith 🔻
Variables	
Scround_Truth	Prompt]
Model	
GPT-4	
Temperature	0
Maximum Tokens	8192
Stop Sequence	

VIRGINIA TECH

### Step 2, 3: Curate and Transform Artifacts

11

Input 💸		New Variant				
Prompt	Based on the information provided, write me 1 paragraph on the Approach for Trade-off Analysis. You are writing this for the Technical Activities and Products section as part of a Milestone A document and a Technology Maturation and Risk Reduction Artifact.					
Prompt	Based on the information provided, write me a very brief Other System Attribute (OSA) for Training. Remember you are writing this for a Capability Description Document. Keep the response to one sentence total.					
Prompt	Based on the information provided, write me two very brief Other System Attribute (OSA) for Operational Environments, e.g., what environmental conditions it should survive in, and a note on early warning capabilities. Remember you are writing this for a Capability Description Document. Keep the response to about a paragraph length.					
	Based on the information provided, write me a very		/	V	T	7

### Step 4, 5: Write and Execute the Prompts

In	DH	ŧ	$\sim$	₽
	րս	÷	ゝ	ð .
				•

#### New Variant

Based on the information provided, write me 1 paragraph on the Approach for Trade-off Analysis. You are writing this for the Technical Activities and Products section as part of a Milestone A document and a Technology Maturation and Risk Reduction Artifact. The approach for Trade-off Analysis for the Silverfish program will involve a systematic evaluation of the various system components and their potential alternatives, with a focus on balancing performance, cost, schedule, and risk factors. This analysis will be conducted in accordance with the Department of Defense's (DoD) Defense Acquisition Guidebook (DAG) and will utilize the Analytic Hierarchy Process (AHP) to prioritize and make decisions about the competing alternatives. The trade-off analysis will consider factors such as the performance of the sensors, the reliability and security of the wireless communication system, the effectiveness of the fire control communications, and the overall system resilience against cyber threats. The analysis will also take into account the system's interoperability with the higher-level command and control (C2) system. The results of this analysis will inform the system design and development process, helping to ensure that the Silverfish program meets its operational requirements while staying within budget and schedule constraints.

Based on the information provided, write me a very brief Other System Attribute (OSA) for Training. Remember you are writing this for a Capability The Silverfish system requires comprehensive operator training to ensure proficient use of the ground-based weapon platforms, sensor systems, and secure wireless.

VIRGINIA TECH

Prompt

Prompt

12

### Step 6: Test and Evaluate

- While the LLMs can clearly generate coherent and high-quality content, the question is how to quantitatively assess how aligned their responses are to ground truth
- This approach uses the MAUVE<sup>1</sup> framework as opposed to classical NLP benchmarks such as ROUGE
- MAUVE uses the Kullback-Leibler divergence between the machine- and human-generated text to assign a scalar between 0 and 1



https://arxiv.org/pdf/2301.05402.pdf

## Results Analysis (1 of 4)

Consecutively iterated prompts get more specific

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

# Results Analysis (2 of 4)

Model Variant (Score)

Claude (0.003)

#### Prompt ("Features")

Based on the information provided, please write me a one paragraph Operational Context statement for a Capability Description Document. Make sure to mention how it impacts the Joint Force Commander and what the Bull Dog/UGV System could support in future battlespaces.

#### Human or AI?

The complexity and uncertainty of the current and future Operational Environment (OE) requires the Joint Force Commander (JFC) to extend the reach of their situational understanding to continuously adapt to a changing environment across the full spectrum of conflict, while facing hybrid threats. This extended battlespace awareness requires the collection of information into the integrated battle command systems network to enable informed decision-making. UGV Systems can support future forces and expanded battlespace concepts by serving as economy of force assets and enhancing force protection by providing standoff operational capabilities for many warfighter functions. UGV Systems must be expeditionary with lethal and non-lethal capabilities that are versatile, agile, sustainable, survivable, and capable to transition across the spectrum of operations. UGV Systems must provide persistent surveillance for situational awareness and can provide force application, including targeting, lethal fires, and non-lethal effects, while protecting the force through standoff from threat capabilities.

### Human or AI?

The Bull Dog Unmanned Ground Vehicle (UGV) System provides the Joint Force Commander (JFC) with an autonomous and deployable asset capable of conducting critical and high-risk warfighting missions in support of decisive action. Operating as an integrated node in the kill chain, the Bull Dog UGV System will provide unmanned capabilities across the spectrum of conflict to achieve overmatch in the multi-domain battlespace. The UGV System's modular, open architecture will enable rapid technological insertion to counter emerging threats and adapt to dynamic operational environments, supporting sustained joint force lethality and survivability. With modular weapon systems and the capacity for line-of-sight and beyond line-of-sight control, the Bull Dog UGV System could support key joint functions such as fires, movement and maneuver, intelligence, and command and control to shape future battlespaces.



# Results Analysis (3 of 4)

Model Variant (Score)

Claude (0.993)

#### Prompt ("Features")

Based on the information provided, please write me a one paragraph Operational Context statement for a Capability Description Document. Make sure to mention how it impacts the Joint Force Commander, the importance of extended battlespace awareness, and what specific warfighter functions the UGV could support as an economy of force asset

#### Human or AI?

The complexity and uncertainty of the current and future Operational Environment (OE) requires the Joint Force Commander (JFC) to extend the reach of their situational understanding to continuously adapt to a changing environment across the full spectrum of conflict, while facing hybrid threats. This extended battlespace awareness requires the collection of information into the integrated battle command systems network to enable informed decision-making. UGV Systems can support future forces and expanded battlespace concepts by serving as economy of force assets and enhancing force protection by providing standoff operational capabilities for many warfighter functions. UGV Systems must be expeditionary with lethal and non-lethal capabilities that are versatile, agile, sustainable, survivable, and capable to transition across the spectrum of operations. UGV Systems must provide persistent surveillance for situational awareness and can provide force application, including targeting, lethal fires, and non-lethal effects, while protecting the force through standoff from threat capabilities.

#### Human or AI?

The UGV system will provide capability to the Joint Force Commander enabling the extended battlespace awareness and freedom of maneuver necessary to decisively defeat adversaries across the spectrum of conflict. By reducing force structure requirements for critical yet high-risk warfighting functions, the UGV system allows the Joint Force Commander to employ a smaller force footprint while retaining operational dominance. Specifically, the UGV system can support economy of force RSTA, maneuver, fire support, and force protection missions, allowing the Joint Force Commander to allocate a higher percentage of manned forces to decisive operations. With onboard sensors and weapons, the UGV system provides 24/7 persistent stare and precision engagement to detect, identify, and defeat threats within the extended battlespace. The UGV system's ability to conduct autonomous lethal and non-lethal effects and seamlessly interface with manned and unmanned systems empower the Joint Force Commander with flexible, scalable, and layered options for shaping the extended battlespace.

### Results Analysis (4 of 4) Qualitative Assessment

- The overall **strengths** of the LLM responses can be characterized as the following:
  - More targeted prompting tends to more closely match the content in the ground truth, i.e., non-specific prompting leaves too much room for the LLM to "explore."
  - Use of terminology tended to be more closely aligned with more specific prompting. Generally, more powerful models (as measured by parameter and training corpus size) can vocabulary match.
  - The LLM made several correct historical references to similar systems to assign system-level attribute requirements.

17

### However, there is still **room for improvement with LLMs**:

- The LLM responses tended to struggle on sections defining Development Threshold and Objectives; the use of LLMs for numeric operations continues to be an open field of research
- While the LLM did a great job on the Market Research Report, there is an open question as to the veracity of those responses (e.g., who in the commercial markets offer those services?)
  - Broadly speaking, when doing Analysis of Alternatives and other exercises, the LLM will need to refer to specific systems, vendors, and processes



## **Conclusions & Next Steps**

### Conclusions

- LLMs can generate reliable and useful SE artifacts but only in conjunction with an expert-in-the-loop
  - The ultimate value of the LLMs will be based on how much expert-editing of the responses are conducted before submission/acceptance
- Limitations
  - Only closed-source models were used
  - Pre-training corpus is limited
  - Limited sample sizes

### **Next Steps**

- Real world artifacts
- Leveraging other VT, SERC, and AIRC research
  - Mission engineering
  - Digital engineering curriculum
  - Test & Evaluation
  - Enterprise transformation
  - Defense Civilian Training Corp (DCTC)





Paul Wach, paulw86@vt.edu

Mohammed Husain, mhusain01@vt.edu

Taylan Topcu, ttopcu@vt.edu







#### Problem

 Can LLMs support an Acquisition Command with accelerating the Army Capabilities Integration and Development System (ACIDS) Process through the generation of Capability Description Documents (CDDs) and Analysis of Alternatives (AoA)?

#### **Completed Work**

- Used retrieval from foundational documents (FARs, DODIs, and CDD Writer's Guide)
- Leveraged fine-tuned open-source models as well as targeted prompt engineering to generate respective artifacts

#### **Lessons Learned**

- Prompt engineering plays a **significant role** in the value of answers
- SFT and RLHF were not explicitly explored but cannot be overlooked

Provide a technology readiness level (TRL) assessment on hypersonics based on the information provided.

Based on the provided excerpts, it is difficult to directly assess the Technology Readiness Level (TRL) of hypersonics.<sup>[1]</sup> However, we can infer that research and development in the field of hypersonics is actively ongoing.<sup>[2]</sup>

The excerpts mention various research areas and application directions related to hypersonics, such as advanced propulsion, computational tools, testing techniques, aerodynamics, and flight experiments.<sup>[3]</sup> This indicates that there is ongoing work to develop and optimize materials, predict material responses, simulate multi-scale coupling, assess flight capabilities, and evaluate performance.<sup>[4]</sup>

Additionally, the excerpts mention the need for ground testing, real-time monitoring of flight experiments, data acquisition, and analysis for

Using OpenAl GPT-3.5 Turbo

🖺 Hide Attachments 🛛 🔁 Reset Chat

Enter a question



Scale knows how to responsibly deploy AI systems to overcome key bottlenecks.



### **Customization & fine tuning**

Base models are outperformed by fine-tuned ones

Edge cases Models invariably encounter edge cases and grow stale

### Safety & security

Models contain PII, proprietary data, or classified information

### Talent bottleneck True end-to-end Al expertise is difficult to scale

Ability to fine-tune models on your proprietary data Data engine for fine-tuning

Human feedback maintains accuracy and prevents model drift Scalable, integrated RLHF



Enterprise architecture and human-in-the-loop safeguards keep data protected Enterprise-grade guardrails



Systems that supplement, not replace, teams at all levels of expertise **Talent Augmentation** 

