



International Council on Systems Engineering
A better world through a systems approach

The Cost of Expertise: Performance Trade-offs in LLMs for Systems Engineering

Presenter: Dr. Paul Wach, PhD

Co-authors:

Mr. Ryan Bell, Mr. Ryan Longshore, and Dr. Raymond Madachy

Naval Postgraduate School (NPS)

Mr. Brady Jugan, Ms. Mary Nerayo

Virginia Tech (VT)

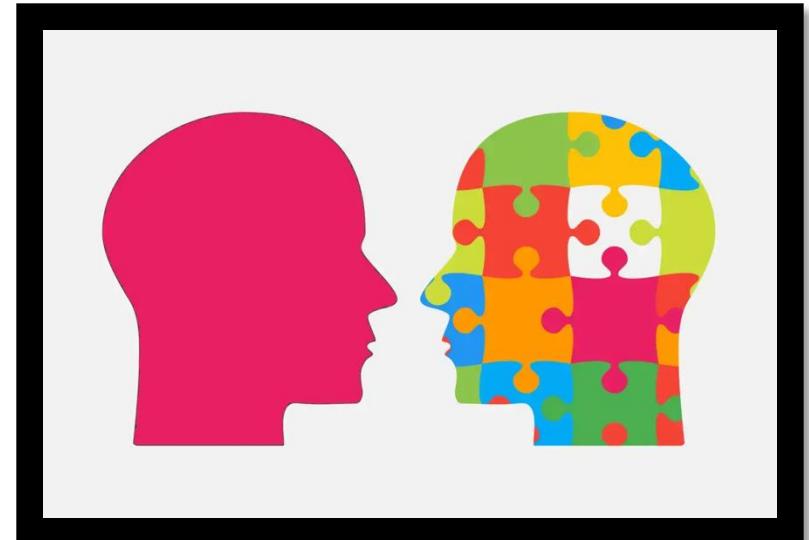


INCOSE International Symposium 2025 | Ottawa, Canada



Motivation

- Systems Engineering (SE) is inherently transdisciplinary.
- SE tools must be **both** deep and **broad**:
Can LLMs be both?
- Central question: *Should LLMs for SE be generalists or specialists?*



WWII - 1940s

Complex systems like radar and aircraft needed to work together.

Holistic thinking — seeing the whole, not just parts.

Post-Cold War - 1990s

Collapse of defense budgets after Cold War → SE lost institutional momentum.

Rise of AI and ML - 2010s-2020s

LLMs now offer new possibilities for accelerating SE tasks.

Semi-formalization: SE became more structured in large defense contractors.

Manhattan & Apollo Projects – 1950s-1960s

Move from document-based to **model-centric** engineering; SysML 1.0 (2006)
DoD Digital Engineering Strategy (2018).

MBSE & Digital Engineering – 2000s-2010s

Why LLMs in Systems Engineering?

Efficiency, Accessibility, and Reproducibility

Artifact Generation

- Requirements documents, ICDs, ConCops, etc.

SysML Translation and Model Manipulation

- SysMLv1 to SysMLv2, Text-to-SysML, etc.

Model Interpretation

- Explain structural/behavioral models to non-expert stakeholders

Knowledge Transfer

- Bridge communication between system modelers and SMEs

Concept Development

- Act as brainstorming partner using chain-of-thought prompts

Automation of Repetitive SE Tasks

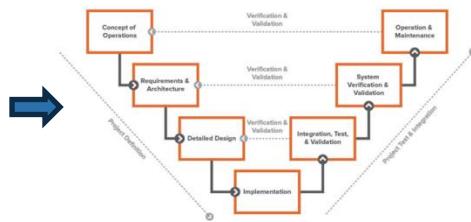
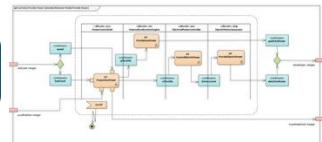
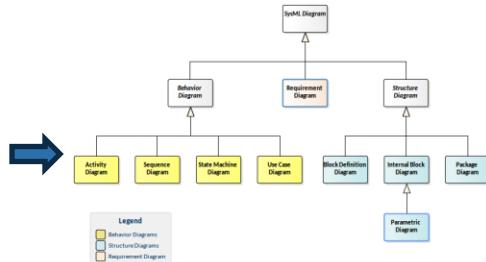
- Generating traceability matrices from requirements to design elements, etc.

SysMLv2

- Generation, manipulation, translation, etc.

Mackandal paragraph

Not much is known about Mackandal and he is almost considered legend. However, this is what we know about him. At age twelve, around 1742, he was taken from his homeland of Africa to Saint-Domingue so some stupid guy named Leoniard Da Mazy could make money off of him. Mackandal was a slave who became a leader of the slaves. He was captured and sold to more than most plantations. He ran away for twelve years, which is extremely impressive, seeing as if a slave can run away for three months or more, the punishment is death. He became an important leader of the black population of Saint-Domingue and later almost succeeded in poisoning the white slave owners. He was, unfortunately, burned at the stake.



Cafed Talh by Morocco and Tura by Spain, the island is claimed by both countries as their territory. Battened by strong winds and waves, and more than 80 km away from the nearest land, the island has only a handful of inhabitants. There are some fish stocks and hopes of natural resources, but the appeal for both countries is largely symbolic: a struggle of wills between independent Morocco and former ruler, Spain. Both Morocco and Spain insist they have long-standing claims to the island. Morocco says it has been occupied by its ancestors since 1492, when it was taken between Morocco and Spain from the Moors. The island was formally placed under the jurisdiction of Morocco in 1899 but was annexed by Spain in 1900, just before Spain's colonization of the Morocco peninsula. Morocco asserts Talh was rightly restored to Morocco after World War II, and a Moroccan contingent deployment has been stationed there since 1947. "Talh is an integral part of Morocco's territory historically, geographically, and under international law," a Moroccan government argues. However, Spain claims that it established sovereignty over the island by the mid 17th century when Spanish sailors used the zone as a port and a fishing ground. Spain incorporated the island in 1713. Spain claims that the Spanish Republics illegally because the island was not mentioned in the Algeria-Spanish Treaty after World War II. It is a land to be returned to Morocco. "The occupation of Tura by Morocco is an illegal behavior undertaken on no basis of international law," Spain's Foreign Ministry says.

Metric	What It Measures	Analogy	Used For
MAUVE	Style similarity (statistical distribution)	“Does it sound like a human wrote it?”	Comparing LLM output to human text
BERTScore	Semantic similarity (meaning overlap)	“Does it mean the same thing?”	Translating models to human-readable form
SysEngBench	Correctness in SE knowledge via MCQs	“Did it get the answer right?”	Measuring domain reasoning (SE skill)

*IS paper is on use of SysEngBench

Three Limited Experiments

- MAUVE
 - To what extent do LLMs achieve stylistic similarity of SE artifacts without specialization?
- BERTScore
 - To what extent do LLMs interpret a SysMLv2 diagram as accurately as a human?
- SysEngBench
 - **Are the more specialized language models more competent at SE tasks?**
 - *Hypothesis: Specialized LLM will perform better than foundation models (unaltered) at Systems Engineering tasks.*
 - NOTE: The term “specialized” is used in reference to common means of adjusting LLMs toward emphasis on a specific domain, such as fine-tuning

*IS paper is on use of SysEngBench

- 3 LLMs tested on 3 prompt styles
- Higher prompt specificity → better style matching with humans
- Baseline – high stylistic similarity achieved
- Begins to suggest : do we need costly specialization?

Table 1. Results across 3 LLMs and 3 prompt configurations.

Model	Prompt 1	Prompt 2	Prompt 3
GPT-4	0.0000	0.0000	0.9137
GPT-3.5 Turbo	0.0000	0.0001	0.9749
Claude	0.0000	0.0003	0.9932

- 3 LLMs:
 - A foundation model (GPT-4o, and two other fine-tuned GPT-4o models)
- Task: Describe SysMLv2 diagrams
- Suggests that LLMs may achieve a statistically significant equivalence to a human generated textual description of SysMLv2

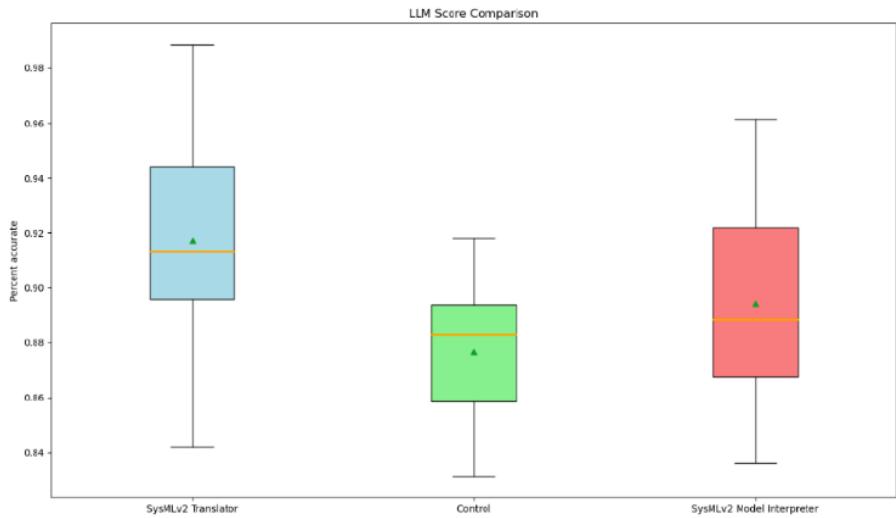


Figure 1: Results from t-test on BERTScore evaluation of the LLM conversion of SysMLv2 images to textual description

Background on SysEngBench

Major Categories	Subcategories	Question Count
Systems Engineering Overview	Systems Engineering Overview/System Concepts and Structures	68
	Systems Engineering Overview/Application and Value of Systems Engineering	18
	Systems Engineering Overview/Systems Science and Thinking	22
Lifecycle Stages	Lifecycle Stages/Generic Life Cycle Stages	35
	Lifecycle Stages/Defense Acquisition Life Cycle/Milestones and Reviews	57
	Lifecycle Stages/Defense Acquisition Life Cycle/Lifecycle Phases	95
Technical Processes	Technical Processes/Identifying Stakeholder Needs and Generating Requirements	88
	Technical Processes/Architecture Definition Process	18
Technical Management Processes	Technical Management Processes/Decision Management, Analysis of Alternatives, and Tradespace Analysis	39
	Technical Management Processes/ Risk and Opportunity Management Process	30
	Technical Management Processes/Configuration and Information Management	43
Cross-Cutting Systems Engineering Methods	Cross-Cutting Systems Engineering Methods/Modeling and Simulation	3
	Cross-Cutting Systems Engineering Methods/Model-Based Systems Engineering/Modeling Frameworks and Methods	92
	Cross-Cutting Systems Engineering Methods/Model-Based Systems Engineering/Modeling Language SysML v1.x	183
Specialty Engineering Activities	Specialty Engineering Activities/Cost Modeling	41
	Specialty Engineering Activities/Electromagnetic Compatibility	41
	Specialty Engineering Activities/Reliability, Availability, and Maintainability	139
	Specialty Engineering Activities/Supportability, Producibility and Disposability	93
	Specialty Engineering Activities/System Safety Engineering	9
	Specialty Engineering Activities/Usability and Human Systems Integration	96
*Total Questions		1144

*Common practice to 'rank' LLMs in specific domains through set of multiple choice questions

Input

Question

How is risk defined in systems engineering?

Choices

- A. The guaranteed outcome of a system's failure.
- B. The process of optimizing a system to avoid any failures.
- C. The potential for loss or an undesirable outcome, quantified by probability and severity.
- D. The act of integrating different system components to ensure safety.



Output

Answer

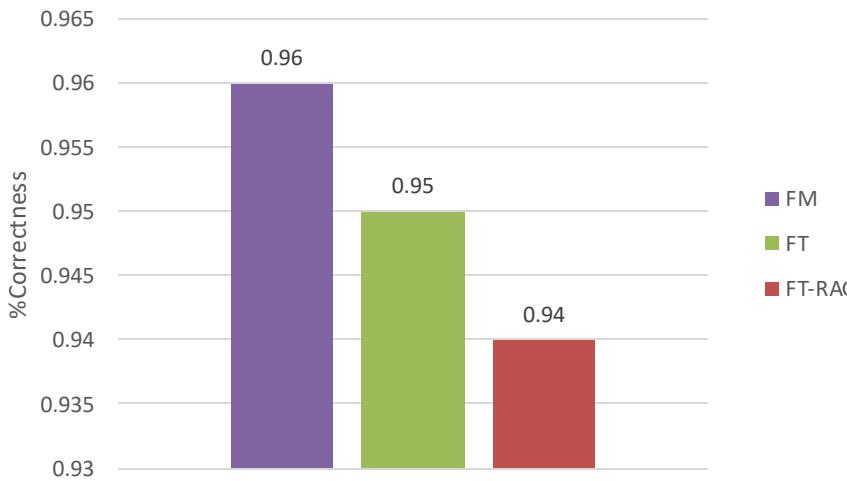
C

LLM

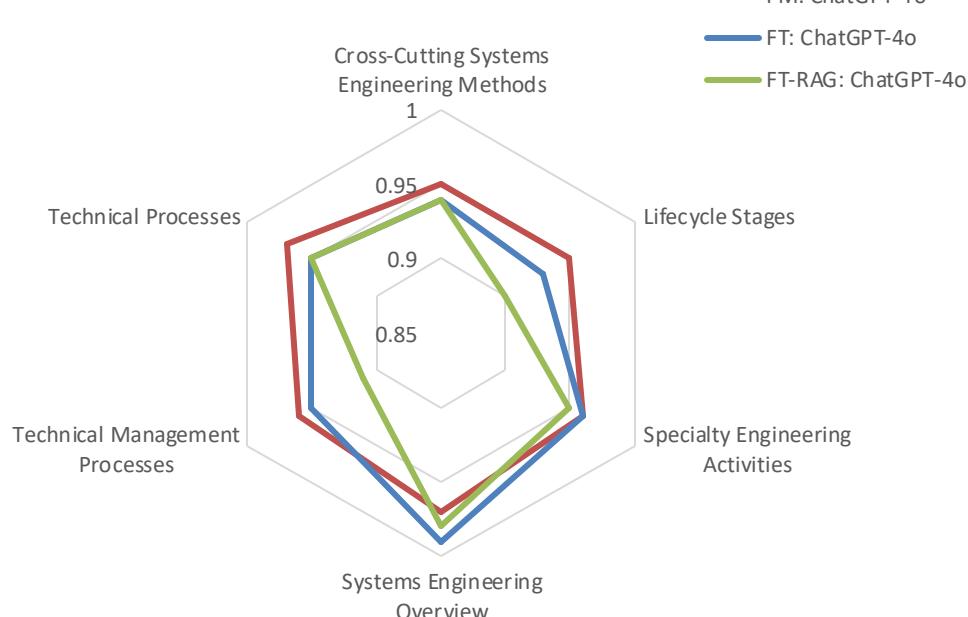


Results from SysEngBench

Identifier	Foundation Language Model	Model Setting(s)	Method of Specialization
Foundation model (FM)	ChatGPT-4o	Temperature = 0	None
Fine-tune model (FT)	ChatGPT-4o	Temperature = 0	Fine-tuned on SysMLv2
FT with RAG (FT-RAG)	ChatGPT-4o	Temperature = 0	Fine-tuned on SysMLv2 with provided SysMLv2 knowledge-base

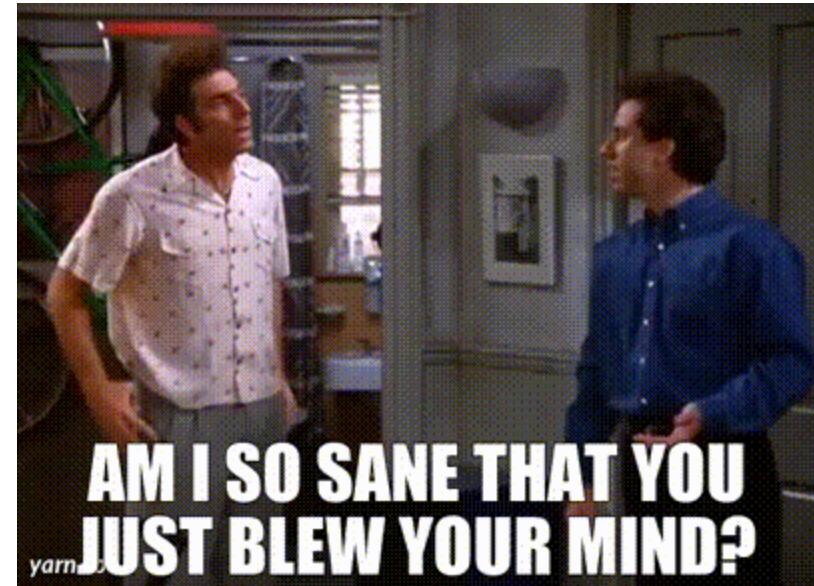


- FT and FT-RAG underperform in most areas, suggesting over-specialization harms general SE reasoning.
- Some gains in "modeling" and "SysML-related"



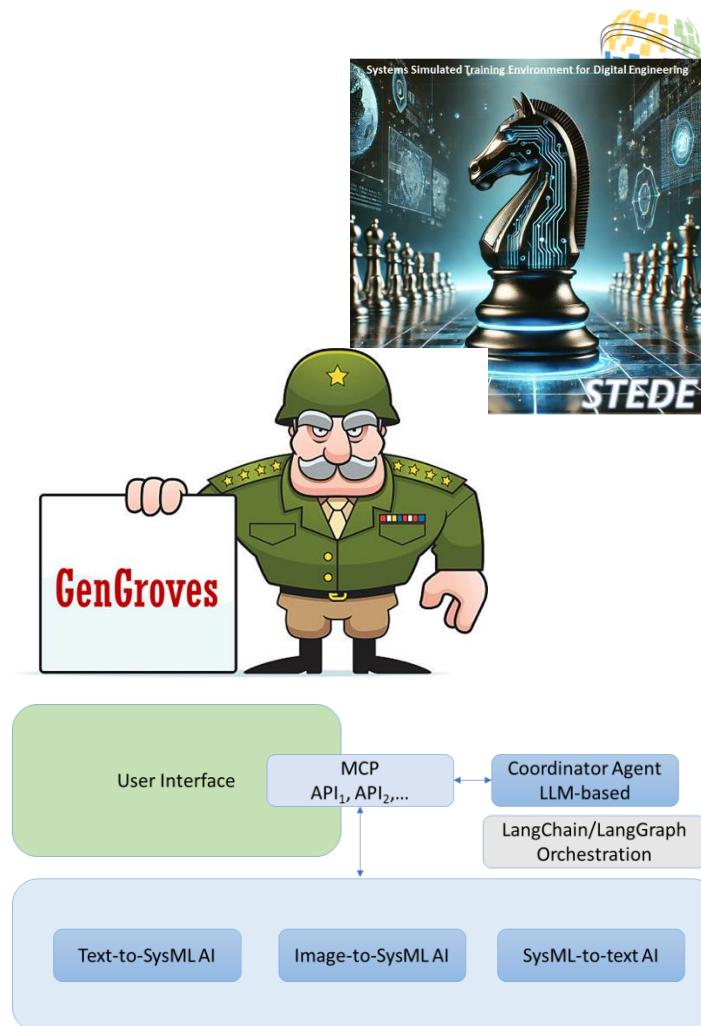
Why would LLMs specialized on SysMLv2 perform worse than unaltered LLMs?

- Overfitting:
 - Small, specific data (SysMLv2) hurts general SE understanding
- Loss of General Knowledge
 - a.k.a. "catastrophic forgetting"
- Task misalignment:
 - SysMLv2 \neq SE competence
- Maturity of SE:
 - Data currently may not fully represent SE discipline



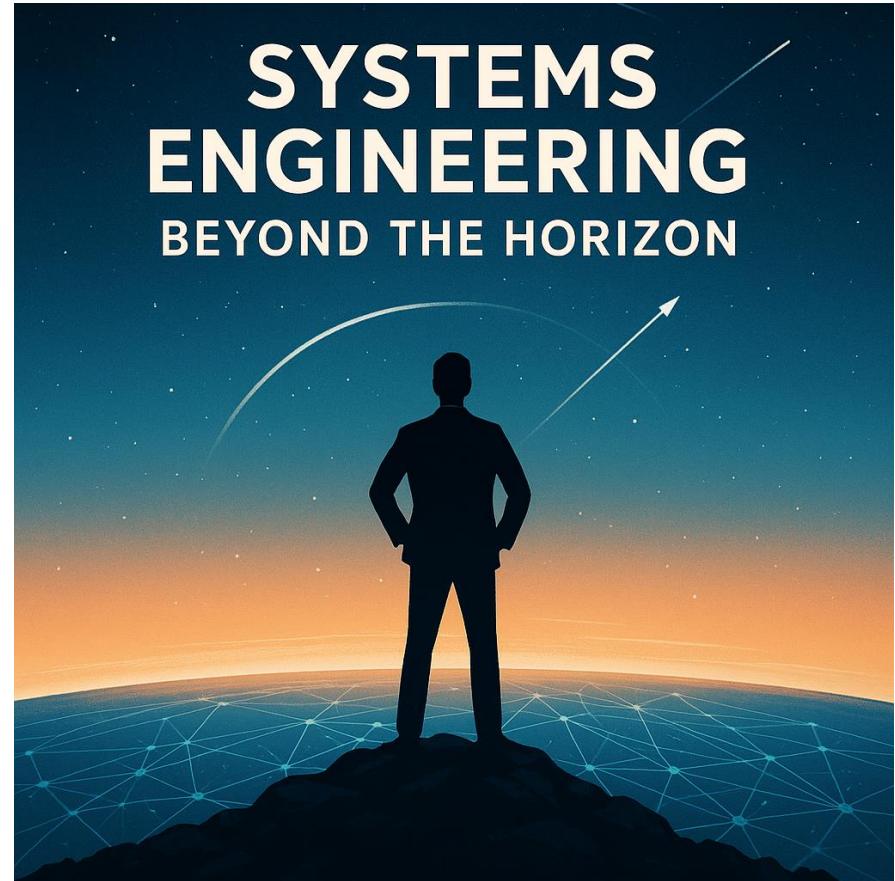
Work-in-progress

- LLMs show potential in generating & describing SE artifacts
- LLMs can **mimic** and **translate** - but may struggle with reasoning when over-specialized.
- Implication:
 - Is SE mature enough for LLM specialization?
 - What other LLM-based methods could change this outcome?
- Our next steps include:
 - Conduct further statistical analysis to **look for trends** within categories and subcategories
 - Conduct additional runs on the **same LLMs, additional LLMs**
 - Establishment of a **baseline method** and initial ranking for benchmarked performance of LLMs on SE tasks



Questions?

Paul Wach, paulw86@vt.edu



Backup