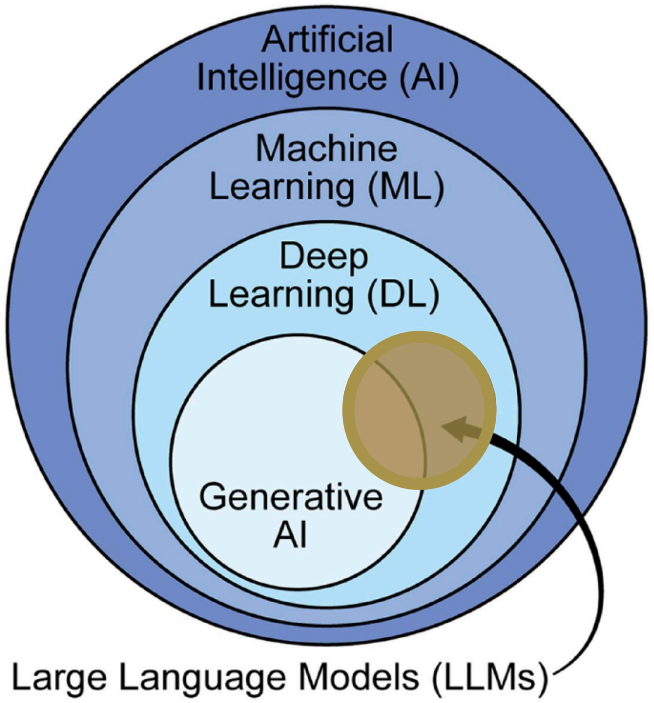
Beyond the Hype:
Trust, Evaluation, And Human Judgement In Aerospace Engineering

Olivia J. Pinon Fischer, Ph.D.

Assistant Professor
Daniel Guggenheim School of Aerospace Engineering
Georgia Institute of Technology

Georgia Tech

Large Language Models (LLMs) in the AI Landscape



The term **"Artificial Intelligence"** encompasses systems that process information, recognize patterns, and generate outputs such as predictions, recommendations, or decisions [3].

1980s–2000s

Rule-Based NLP

TECHNIQUES

Regex patterns, handcrafted grammars, keyword matching

LIMITATION

Brittle — can't handle ambiguity or variation in phrasing

2000s–2015

Statistical / ML

TECHNIQUES

Naive Bayes, SVM, CRFs, TF-IDF, Word2Vec

LIMITATION

Better generalization but still limited contextual understanding

2015–2017

Deep Learning

TECHNIQUES

RNNs, LSTMs, CNNs for sequence modeling

LIMITATION

Sequential processing — slow, struggles with long-range dependencies

2017–Present

Transformers

TECHNIQUES

Self-attention, parallelism BERT, GPT, T5, LLaMA

BREAKTHROUGH

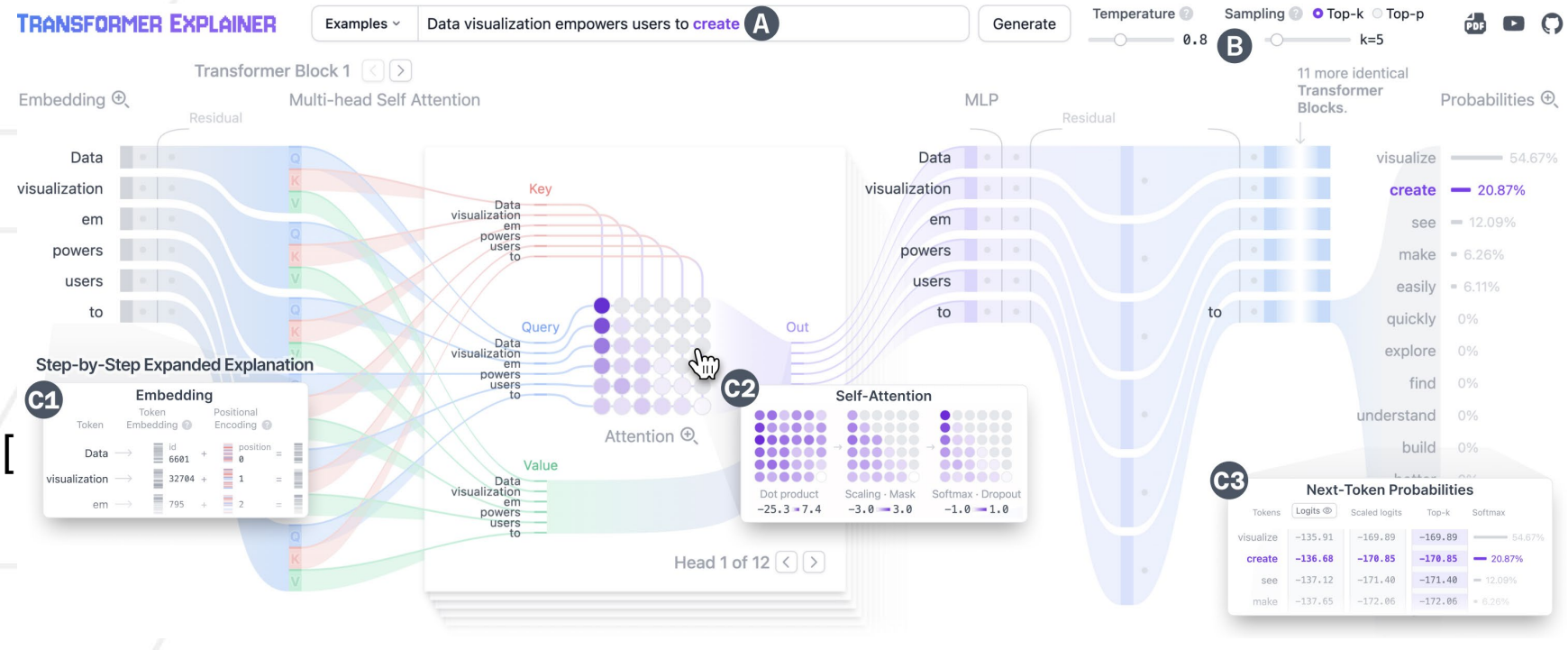
Self-attention processes sequences in parallel, capturing long-range dependencies far more effectively

LLMs are **one class** of AI model.

[1] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
[2] Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.
[3] J. McCarthy, M. Minsky, N. Rochester, and C. Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," 1955.

Georgia Tech
Olivia J. Pinon Fischer, Ph.D.

LLMs… in a Nutshell



[1]

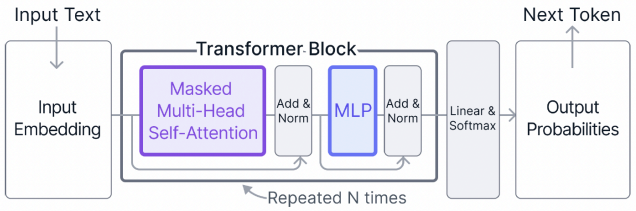


Figure 2: Diagram illustrating the data flow of the Transformer architecture (decoder-only), which consists of many components: input text is converted to input embeddings that pass through repeated identical Transformer blocks, each containing masked multi-head self-attention mechanisms, followed by multi-layer perceptron (MLP) layers with residual connections (Add) and normalization (Norm). The final linear and softmax layers output probabilities for the next-token prediction.

[2]

Autoregressive LLMs generate text by predicting the next token.

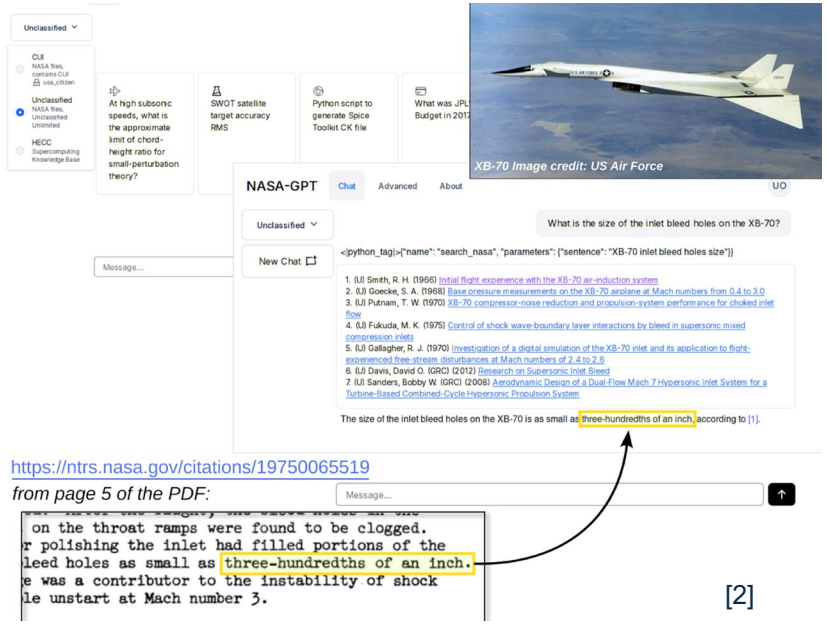
[1] https://poloclub.github.io/transformer-explainer/
[2] Cho, A., Kim, G. C., Karpekov, A., Lee, S., Helbling, A., Hoover, B., ... & Chau, D. H. (2026, April). Transformer Explainer: Learning LLM Transformers with Interactive Visual Explanation and Experimentation. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing*

Georgia Tech
Olivia J. Pinon Fischer, Ph.D.

“LLMs are extraordinary librarians…..” [1]

“They have read almost everything
Useful for breath, retrieval, analogy, and first drafts”

Ken Ono's talk at ***Organizing Mathematical Knowledge in the Age of AI and Formalization****, National Academy of Sciences, Washington, DC, July 20, 2026.*



[2]

[1] *Ken Ono's talk at Organizing Mathematical Knowledge in the Age of AI and Formalization, National Academy of Sciences, Washington, DC, July 20, 2026.*
[2] Dalle, D. (2024). NASA-GPT: Searching the Entire NASA Technical Reports Server Using AI. NASA@SC24 project 09. NASA Ames Research Center.
[3] Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in neural information processing systems* 33 (2020): 9459-9474.

Georgia Tech
Olivia J. Pinon Fischer, Ph.D.

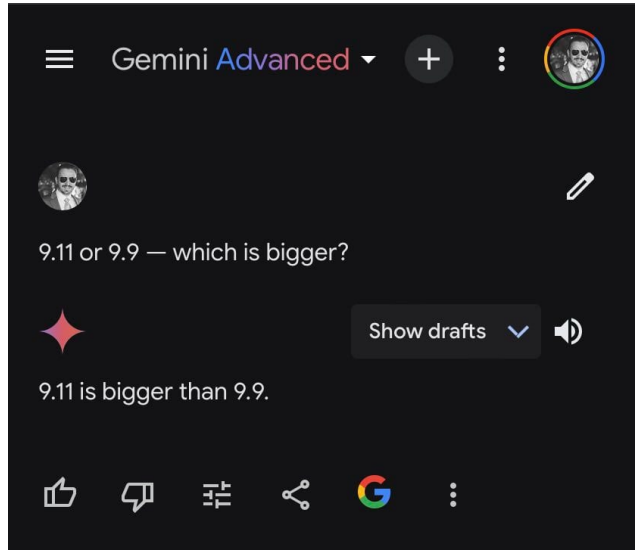
“…. but in Mathematics they can still…” [1]

- *“Invent lemmas*
- *Omit hypotheses*
- *Give plausible but invalid proof steps”*

*Ken Ono's talk at **Organizing Mathematical Knowledge in the Age of AI and Formalization**, National Academy of Sciences, Washington, DC, July 20, 2026.*

“Large language models remain unreliable for exact quantitative computation as they may produce incorrect or hallucinated results when executing mathematical and domain-specific quantitative calculations.”

Madachy, Raymond. "Trustworthy AI for Quantitative Systems Engineering via Deterministic Delegation to Verified Tools, *INSIGHT*, 2026

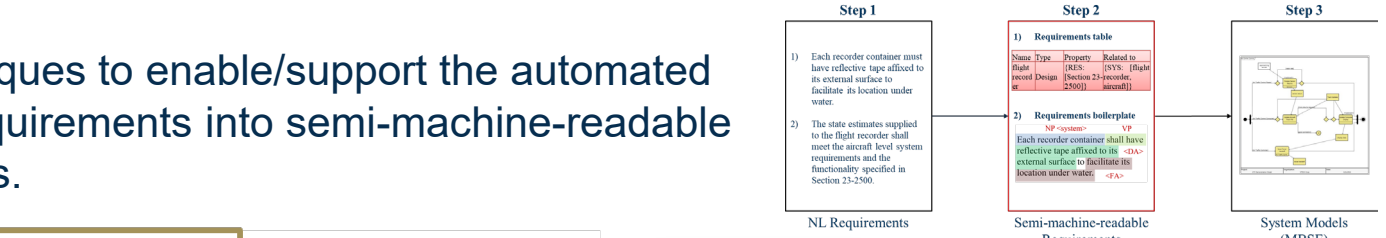


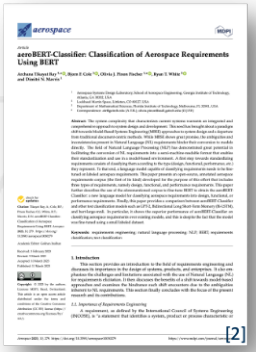
Do you want your librarian to be your systems engineer?

[1] *Ken Ono's talk at Organizing Mathematical Knowledge in the Age of AI and Formalization, National Academy of Sciences, Washington, DC, July 20, 2026.*
[2] Madachy, Raymond. "Trustworthy AI for Quantitative Systems Engineering via Deterministic Delegation to Verified Tools." *INSIGHT* 29, no. 3 (2026): 58-66.

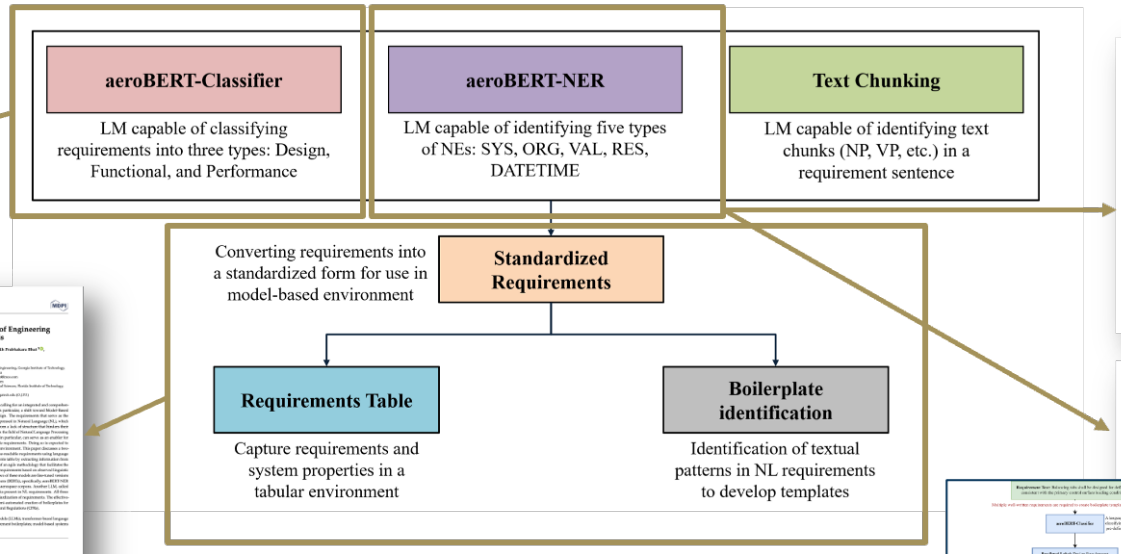
Examples of applications of AI in systems engineering

Identify, develop, and implement tools and techniques to enable/support the automated translation of natural language (NL) aerospace requirements into semi-machine-readable requirements.




[2]


[1]

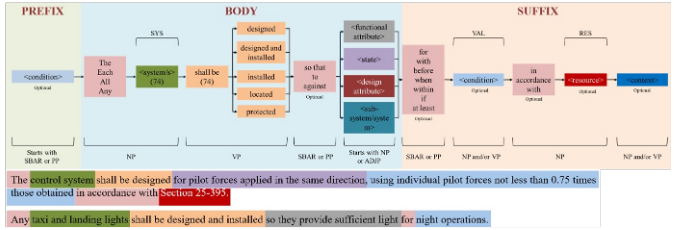



[4]


[5]

Improving LLM performance in Aerospace NER task with Entity-adaptive data augmentation and fine-tuning strategies


[3]



Standardization of requirements for their use in a model-based environment:

- -Extraction of information from free-form NL requirements
- -Agile methodology **for creation of boilerplate templates**

[1] Tikayat Ray, A., Cole, B. F., Pinon Fischer, O. J., Bhat, A. P., White, R. T., & Mavris, D. N. (2023). Agile methodology for the standardization of engineering requirements using large language models. *Systems*, *11*(7), 352.
[2] Tikayat Ray, A., Cole, B. F., Pinon Fischer, O. J., White, R. T., & Mavris, D. N. (2023). aerobert-classifier: Classification of aerospace requirements using bert. *Aerospace*, *10*(3), 279.
[3] Tikayat Ray, A., Pinon-Fischer, O. J., Mavris, D. N., White, R. T., & Cole, B. F. (2023). aeroBERT-NER: Named-entity recognition for aerospace requirements engineering using BERT. In *AIAA SciTech 2023 Forum* (p. 2583).
[4] Tikayat Ray, A., Pinon Fischer, O. J., White, R. T., Cole, B. F., & Mavris, D. N. (2024). Development of a language model for named-entity-recognition in aerospace requirements. *Journal of Aerospace Information Systems*, *21*(6), 489-499.
[5] Cho, K., Pinon-Fischer, O. J., & Mavris, D. (2025). Improving LLM performance in aerospace NER task: Entity-adaptive data augmentation and fine-tuning strategies. In *AIAA AVIATION FORUM AND ASCEND, AIAA-2025-3246*

AI for Systems Engineering

The number of examples of applications of AI in Systems Engineering is exploding

Engineering knowledge

- Search technical reports and project records
- Answer questions with links to sources

Requirements

- Identify requirements and possible defects
- Suggest traceability links

System models

- Create and revie SysML models
- Query an existing system architecture

Design and analysis

- Explore mission designs and their tradeoffs
- Run analyses through engineering tools

Testing and assurance support

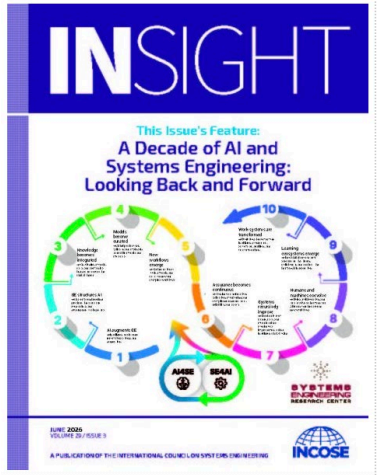
- Find failures in simulation
- Draft provisional requirements and diagrams from code

Benchmarks and evaluation

- Test knowledge, artifacts and engineering workflows
- Establish task criteria and check how scores are assigned

Human–AI teaming

- Capture the engineer's objectives and constraints
- Study review effort, error detection and learning



[1] Poulsen, V. V., Guertler, M., Eisenbart, B., & Sick, N. (2025). Advancing systems engineering with artificial intelligence: a review on the future potential, challenges and pathways. Proceedings of the Design Society, 5, 359–368.
[2] Dalle, D. (2024). NASA-GPT: Searching the Entire NASA Technical Reports Server Using AI. NASA Ames, SC24 research demonstration.
[3] Dogru, M. S., Penkert, J., Mahr, F., & Düser, T. (2026). AI applications in requirements engineering: a systematic mapping study. Proceedings of the Design Society, 6, 2263–2272.
[4] Esho, T., Hoyt, C., Marshall, J., & Gadewadikar, J. (2026). Artificial Intelligence Enabled Systems Engineering Modeling With Retrieval Augmented Generation. Systems Engineering, 29, 444–457.
[5] Quast, V., Jacobs, G., Dehn, S., & Höpfner, G. (2026). Enabling Humans and AI Systems to Retrieve Information from System Architectures in Model-Based Systems Engineering. Systems, 14(1), 83.
[6] Viros i Martin, A., & Selva, D. (2020). Daphne: A Virtual Assistant for Designing Earth Observation Distributed Spacecraft Missions. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 30–48.
[7] Qin, S., et al. (2026). StructureClaw: Traceable LLM Agents and an Executable Benchmark for Structural Engineering Workflows. arXiv:2607.14896v2.
[8] Kambhampaty, J. D., Wong, N. H., Theriot, E. M., Davis, M. M., Pinon Fischer, O. J., Mavris, D. N., & Souaya, M. (2026). Reverse engineering provisional DO-178C certification artifacts with large language models. DASC 2026
[9] Bell, R., Longshore, R., & Madachy, R. (2024). Introducing SysEngBench: A Novel Benchmark for Assessing Large Language Models in Systems Engineering. Acquisition Research Program, SYM-AM-24-072.
[10] Topcu, T. G., Husain, M., Ofsa, M., & Wach, P. (2025). Trust at Your Own Peril: A Mixed Methods Exploration of the Ability of Large Language Models to Generate Expert-Like Systems Engineering Artifacts and a Characterization of Failure Modes. Systems Engineering, 28, 583–604.
[11] Selva, D. (2026). On (Human) Learning in Design Space Exploration and the Impact of AI Assistants. INCOSE INSIGHT, 29(3), 43–47.
[12] Zhao, L., Alhoshan, W., Ferrari, A., Letsholo, K. J., Ajagbe, M. A., Chioasca, E.-V., & Batista-Navarro, R. T. (2021). Natural Language Processing (NLP) for Requirements Engineering (RE): A Systematic Mapping Study. ACM Computing Surveys.
[13] Lee, R., Mengshoel, O. J., Saksena, A., Gardner, R., Genin, D., Silbermann, J., Owen, M., & Kochenderfer, M. J. (2020). Adaptive Stress Testing: Finding Likely Failure Events with Reinforcement Learning. Journal of Artificial Intelligence Research, 69, 1165–1201.
[14] Chaudhari, A. M., & Selva, D. (2023). Evaluating Designer Learning and Performance in Interactive Deep Generative Design. Journal of Mechanical Design, 145(5), 051403.
[15] Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten Challenges for Making Automation a "Team Player" in Joint Human-Agent Activity. IEEE Intelligent Systems, 19(6), 91–95.
[16] McDermott, T., DeLaurentis, D., Beling, P., Blackburn, M., & Bone, M. (2020). AI4SE and SE4AI: A Research Roadmap. INCOSE INSIGHT, 23(1), 8–14.
[17] Zhang, W., Cockburn, C., Henshaw, M., Douglas, P., Palmer, P., Olivier-Myall, J., & Ji, S. (2026). MBSE Co-Pilot: A Research Roadmap. Systems Engineering, 29, 20–33.

Georgia Tech
Olivia J. Pinon Fischer, Ph.D.

Evaluating LLM-generated SE artifacts

AI-generated results must meet the **same engineering acceptance criteria** as results produced by other methods.

Evaluation should cover **both** the output and the process that produced it.

> **How do we measure artifact quality and evaluate the work of producing and reviewing the artifacts?**

What should we measure? What are appropriate metrics?

How do we evaluate generation?

Examine source use and uncertainty. Test the effects of prompts, missing context and repeated runs.

How do we validate the scores?

How do we study human–LLM teams?

Measure errors detected and missed, review effort, and how engineers reach and revise their judgments.

Early evaluations exist.

More work is needed on measures of artifact quality and on evaluation of the human-LLM workflow.

Related work in other disciplines


Articles

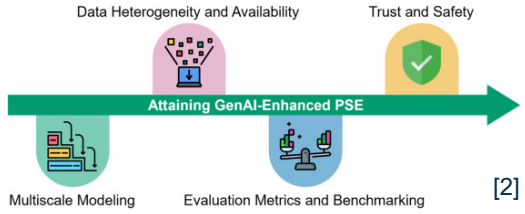
What generative AI means for trust in health communications

Adam G. Dunn, Ivy Shih, Julie Ayre & Heiko Spallek

Pages 385-388 | Published online: 03 Nov 2023

Cite this article

[1]



[2]

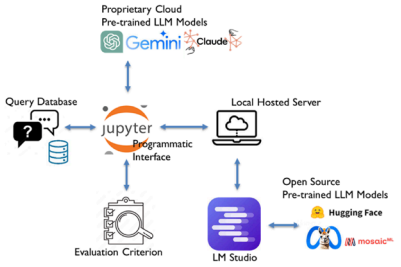
Generative AI challenges in Process Systems Engineering (PSE) [2]

[1] Dunn, Adam G., Ivy Shih, Julie Ayre, and Heiko Spallek. "What generative AI means for trust in health communications." *Journal of Communication in Healthcare* 16, no. 4 (2023): 385-388.
[2] Decardi-Nelson, Benjamin, Abdulelah S. Alshehri, Akshay Ajagekar, and Fengqi You. "Generative AI and process systems engineering: The next frontier." *Computers & Chemical Engineering* 187 (2024): 108723.

Georgia Tech

Olivia J. Pinon Fischer, Ph.D.

Preliminary Benchmarks and Evaluations



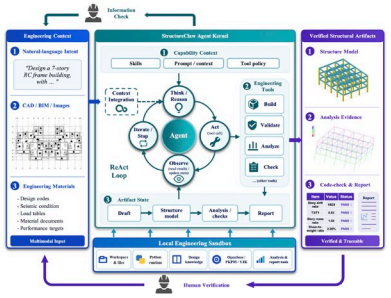
SysEngBench [1]

Focus

116 multiple-choice questions from one introductory systems engineering course. Scores the percentage answered correctly.

Limitation

No generated artifacts or professional workflows. Comparison with practicing engineers was proposed as future work.



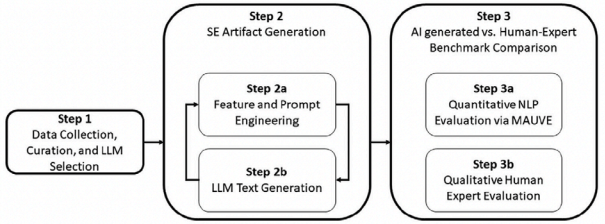
StructureClaw [2]

Focus

Compares geometry, connectivity, supports, loads and solver results with references. Tests requests for missing information and no analysis with invalid inputs.

Limitation

The benchmark checks against a predefined model. It does not determine whether a different way of modeling the same structure is also acceptable. The language judge was not calibrated against independent engineers.



Trust at Your Own Peril [3]

Focus

52 prompts from a UGV teaching case. MAUVE compares generated and expert text. Qualitative review found premature requirements, unsupported estimates and overspecification.

Limitation

One case; review of the best-scoring Claude set only. No tests of variability or team performance.

[1] Ryan Bell, Ryan Longshore, and Raymond Madachy. *Introducing SysEngBench: A novel benchmark for assessing large language models in systems engineering.* Acquisition Research Program, 2024.
[2] Qin, Sizhong, Yi Gu, Yao Jiang, Ao Cai, Changjian Zhou, Shaoxuan Shuai, Jiachang Wang et al. "StructureClaw: Traceable LLM Agents and an Executable Benchmark for Structural Engineering Workflows." *arXiv preprint arXiv:2607.14896* (2026).
[3] Topcu, T. G., Husain, M., Ofsa, M., & Wach, P. (2025). Trust at Your Own Peril: A Mixed Methods Exploration of the Ability of Large Language Models to Generate Expert-Like Systems Engineering Artifacts and a Characterization of Failure Modes. Systems Engineering, 28, 583–604.

Georgia Tech

Olivia J. Pinon Fischer, Ph.D.

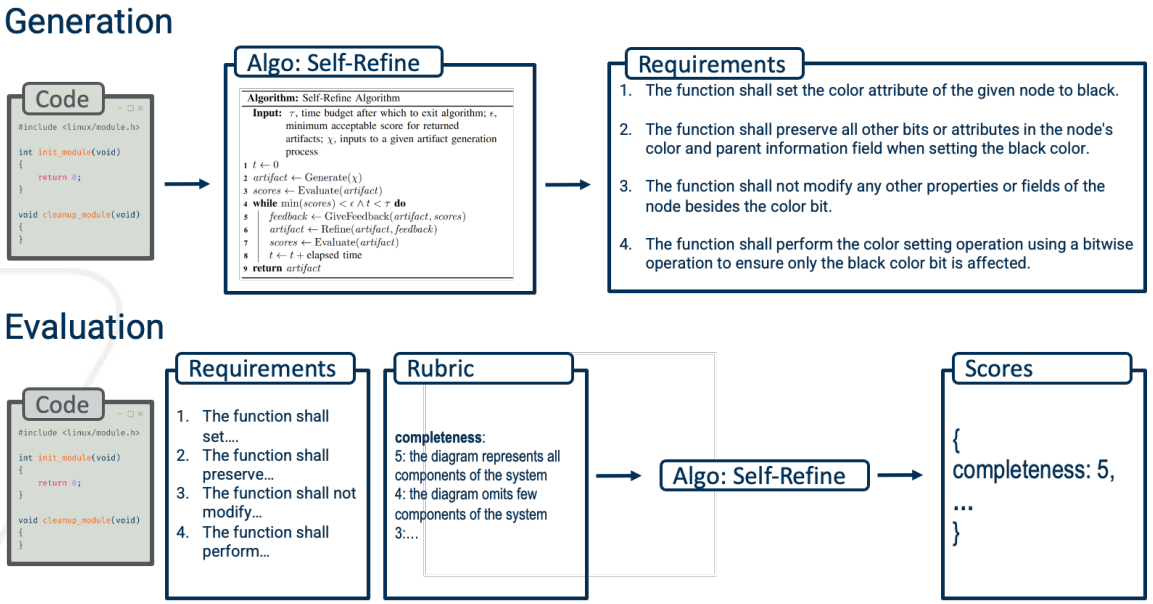
Preliminary Benchmarks and Evaluations

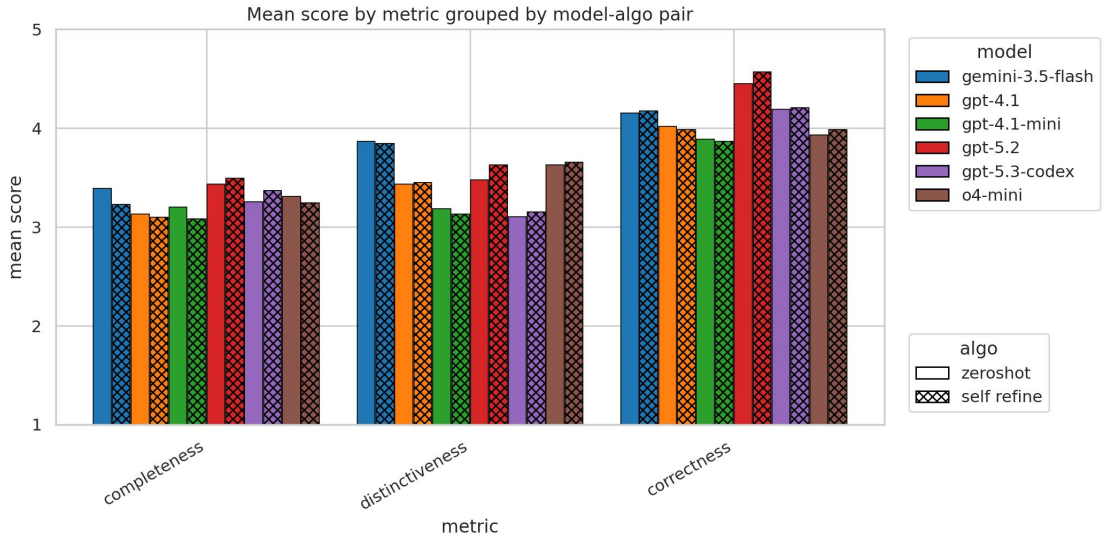
Reverse engineering provisional DO-178C certification artifacts with large language models

Focus: Generating provisional DO-178C certification artifacts with large language models

Research Objective: build experimental infrastructure to assess the feasibility of using LLMs in a reverse engineering context

- Use case: red-black-tree code from the Xen hypervisor.
- Outputs: low-level requirements, function groupings, sequence and class diagrams.
- Provisional: candidates for engineering review, **not evidence of certification acceptance**.





[1] J. D. Kambhampaty , N. H. Wong , E. M. Theriot , M. M. Davis, O. J. Pinon Fischer, D. N. Mavris, and M. Souaya, *Reverse engineering provisional DO-178C certification artifacts with large language models*, Digital Avionics Systems Conference (DASC), September 2026, Orlando , FL

Georgia Tech
Olivia J. Pinon Fischer, Ph.D.

Preliminary Benchmarks and Evaluations

Reverse engineering provisional DO-178C certification artifacts with large language models

Comparing LLM configurations

Prompts, context, models and tools can change the result. Compare them on shared tasks with reproducible evaluation procedures.

Defining artifact quality

Standards provide relevant criteria. Their use in a reverse-engineering rubric needs justification; some judgments depend on task and context.

Evaluating engineers using the tool

Measure missed errors, corrections, review time and decision quality. Did not study human-tool interaction.

Checking the LLM evaluator

Compared LLM scores with one human evaluator. Agreement was mixed; systematic calibration and inter-rater studies remain necessary.

[1] J. D. Kambhampaty , N. H. Wong , E. M. Theriot , M. M. Davis, O. J. Pinon Fischer, D. N. Mavris, and M. Souaya, *Reverse engineering provisional DO-178C certification artifacts with large language models*, Digital Avionics Systems Conference (DASC), September 2026, Orlando , FL

Preliminary Benchmarks and Evaluations

Reverse engineering provisional DO-178C certification artifacts with large language models

Comparing LLM configurations

Prompts, context, models and tools can change the result. Compare them on shared tasks with reproducible evaluation procedures.

Defining artifact quality

Standards provide relevant criteria. Their use in a reverse engineering rubric needs justification; some judgments depend on

We need benchmarks that test engineering intent, artifact quality and engineers' ability to review and use the result.

Measure missed errors, corrections, review time and decision quality. Did not study human-tool interaction.

Checking the LLM evaluator

Compared LLM scores with one human evaluator. Agreement was mixed; systematic calibration and inter-rater studies remain necessary.

[1] J. D. Kambhampaty , N. H. Wong , E. M. Theriot , M. M. Davis, O. J. Pinon Fischer, D. N. Mavris, and M. Souaya, *Reverse engineering provisional DO-178C certification artifacts with large language models*, Digital Avionics Systems Conference (DASC), September 2026, Orlando , FL

Georgia Tech

Olivia J. Pinon Fischer, Ph.D.

Additional Challenges

VOLUME

thousands of lines of code, entire documentation volumes, ….

SPEED

…. generated in minutes

POLISH

…. and delivered in confident, fluent language.

[1] Garousi, Vahid. "Human Oversight and Overload: Two Hidden and Costly Burdens of AI-Assisted Software Engineering." arXiv preprint arXiv:2606.05770 (2026).
[2] Bansal, Gagan, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. "Does the whole exceed its parts? the effect of ai explanations on complementary team performance." In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1-16. 2021.
[3] Lee, John D., and Katrina A. See. "Trust in automation: Designing for appropriate reliance." *Human factors* 46, no. 1 (2004): 50-80.

When generation outpaces validation

The cost of review

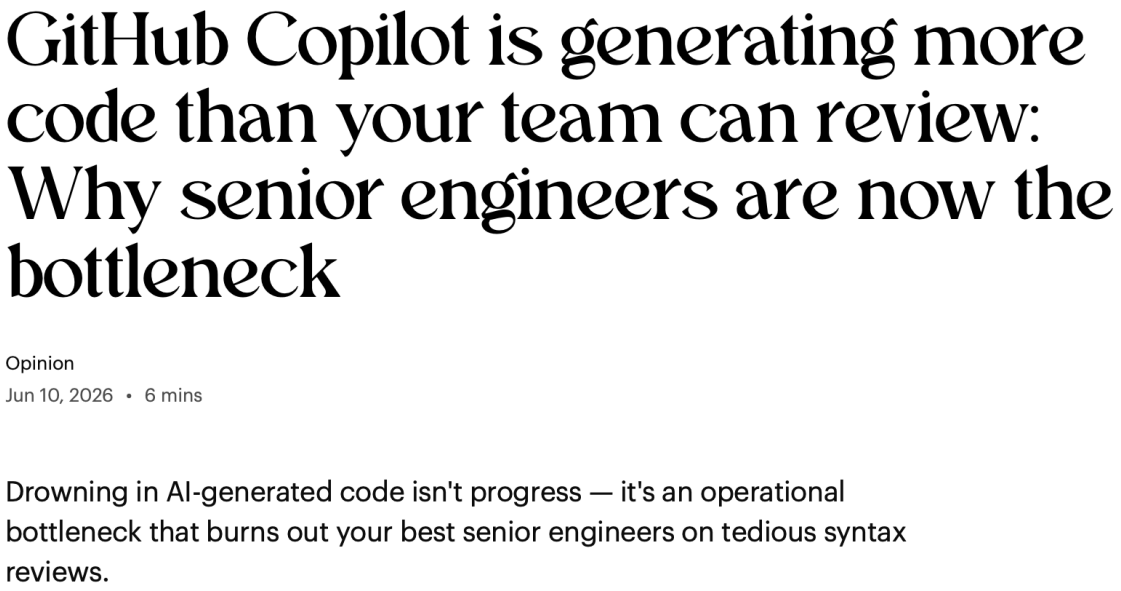
“a wealth of information creates a poverty of attention”
~Herbert A. Simon, 1971

Structural tension

Generation is fast; validation stays bounded by human review capacity. Review and rework can absorb time saved in generation [1].

Mental demands of AI oversight

Intensive use and monitoring of AI tools can contribute to mental fatigue [2].



GitHub Copilot is generating more code than your team can review: Why senior engineers are now the bottleneck

Opinion

Jun 10, 2026 • 6 mins

Drowning in AI-generated code isn't progress — it's an operational bottleneck that burns out your best senior engineers on tedious syntax reviews.

[1] Garousi, Vahid. "Human Oversight and Overload: Two Hidden and Costly Burdens of AI-Assisted Software Engineering." *arXiv preprint arXiv:2606.05770* (2026).
[2] Bedard, J., Kropp, M., Hsu, M., Karaman, O. T., Hawes, J., & Kellerman, G. R. (2026, March 5). When using AI leads to “brain fry.”

Georgia Tech
Olivia J. Pinon Fischer, Ph.D.

Automation changes the work left to people

“For the first time in history, technology is not automating what leaders do with their hands or their schedules, or their arithmetic. **It's automating what they do with their minds.”** [1]

Engineers and organizations remain responsible for decisions supported by AI.

“If the computer is being used to make decisions because human judgement and intuitive reasoning are not adequate in this context, then which of the decisions is to be accepted? The human monitor has been given an impossible task.” [2]
~Lisanne Bainbridge, 1983

Excessive automation may hinder learning [3]. However, whether long-term AI use strengthens or weakens engineering competence needs direct longitudinal study.

[1] T.R. Clark, Leading Through AI: Who You Are When Machines Can Think, April 2026
[2] Bainbridge, Lisanne. "Ironies of automation." In Analysis, design and evaluation of man–machine systems, pp. 129-135. Pergamon, 1983.
[3] Selva, Daniel. "On (Human) Learning in Design Space Exploration and the Impact of AI Assistants." INSIGHT 29, no. 3 (2026): 43-47.
[4] Lovett, Nolan. "The tragedy of the cognitive commons: How AI could disrupt the regeneration of professional expertise." Human Resource Development Review (2026): 15344843261470602.

Georgia Tech
Olivia J. Pinon Fischer, Ph.D.

The role of human expertise

. . . and the risk of over-reliance

Interventions which require the person to form their own judgment before seeing the suggestion, or which delay the suggestion, reduce overreliance compared with simply explaining the AI's answer [1].

The reviewer needs to understand the system well enough to spot missing requirements and questionable design choices. This requires **engineering knowledge and engineering intuition.**

Novices tend to over-rely on what an AI system tells them.

> "Effective AI oversight depends on the expertise AI adoption may undermine." [2]

[1] Buçinca, Zana, Maja Barbara Malaya, and Krzysztof Z. Gajos. "To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making." *Proceedings of the ACM on Human-computer Interaction* 5, no. CSCW1 (2021): 1-21.
[2] Lovett, Nolan. "The tragedy of the cognitive commons: How AI could disrupt the regeneration of professional expertise." Human Resource Development Review (2026): 15344843261470602.

The impact of AI on learning

The role of Academia

- **Keep a strong foundation in engineering fundamentals.**
- Assess the reasoning and process by which students obtain the answer, not only the final answer.

Train students in identifying what problem to solve and what questions to ask:

Formulation	Method	Verification	Understanding
Problem definition	Choice of model	Independent checks	Explain the analysis
Assumptions	Role of AI	Errors and corrections	Defend the conclusion
Initial estimate	Justified decisions	Limitations	Predict a new case

Students should graduate able to formulate a problem, judge a result and defend an engineering decision.

With each new student generation entering college, the over-reliance on AI for everything will be more pronounced (middle schoolers already use AI to write their essays and do their homework)

Georgia Tech

Olivia J. Pinon Fischer, Ph.D.

“These days we adopt innovations in large numbers, and put them to extensive use, faster than we can even hope to know their consequences… which tragically removes our ability to control the course of events.”

~ Patric Lagadec, La civilisation du risque: Catastrophes technologiques et responsabilité sociale. (1981).

Thank you

Olivia Pinon Fischer, Ph.D. (olivia.pinon@ae.gatech.edu)
Assistant Professor
Daniel Guggenheim School of Aerospace Engineering
Georgia Institute of Technology

Georgia Tech