Towards Trustworthy Integration of Generative AI in the MBSE Development Lifecycle

Collins Aerospace
An RTX Business

Amer Tahat, Isaac Amundson, David Hardin, Darren Cofer

25th Annual High Confidence Software and Systems (HCSS 2025)

May 14, 2025

DARPA PROVERS

Pipelined Reasoning of Verifiers Enabling Robust Systems

Develop automated, scalable **formal methods** tools that are integrated into *traditional* development pipelines

Enable *traditional* product engineers to incrementally produce and maintain **high-assurance** national security systems

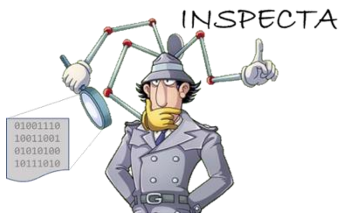
Barriers to formal methods adoption

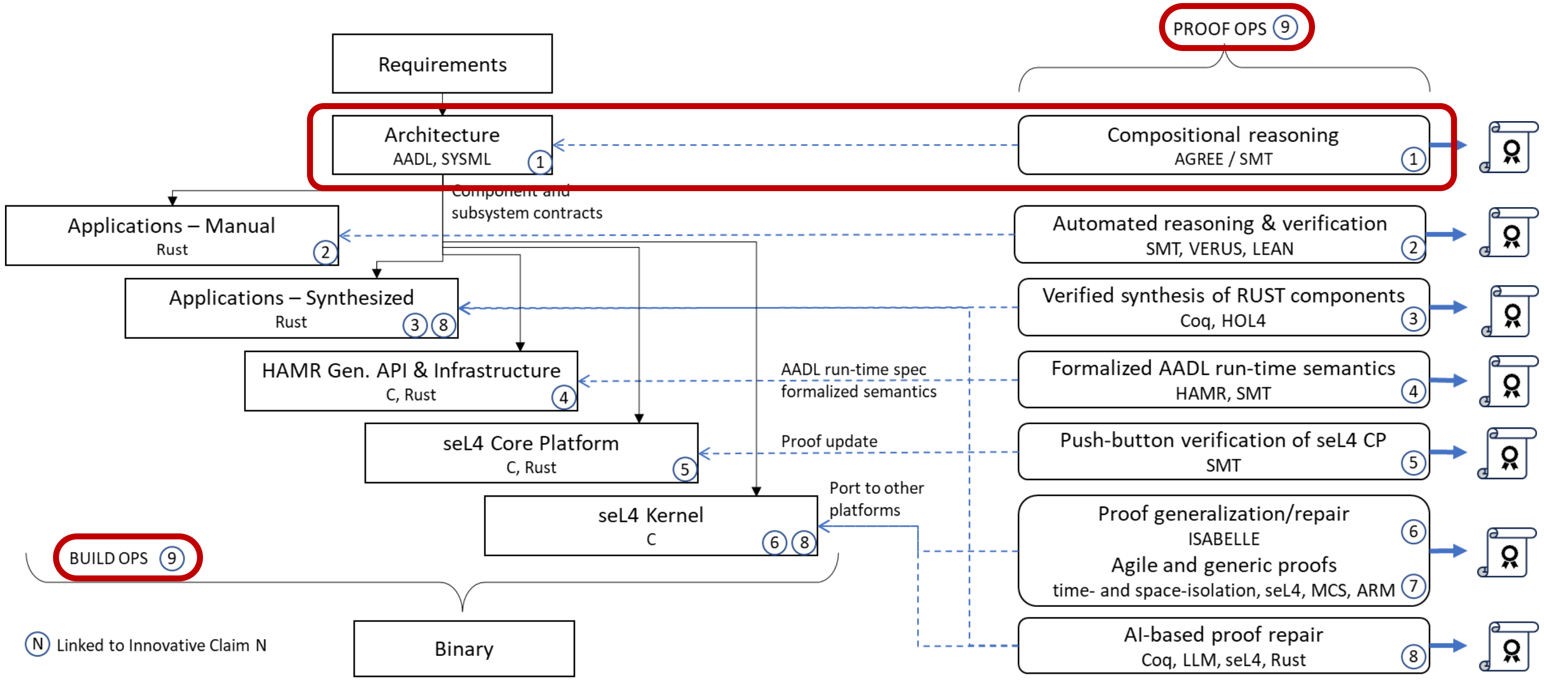
- Scalability to support real-world systems
- The UI targets formal-methods experts, creating a learning curve for others.
- Lack of commercially licensed / supported tools
- Formal methods skillset required
  - Property specification language
  - Explainable counterexamples

GenAI's Potential: Accelerates system development — but trust, correctness, and consistency must be non-negotiable.

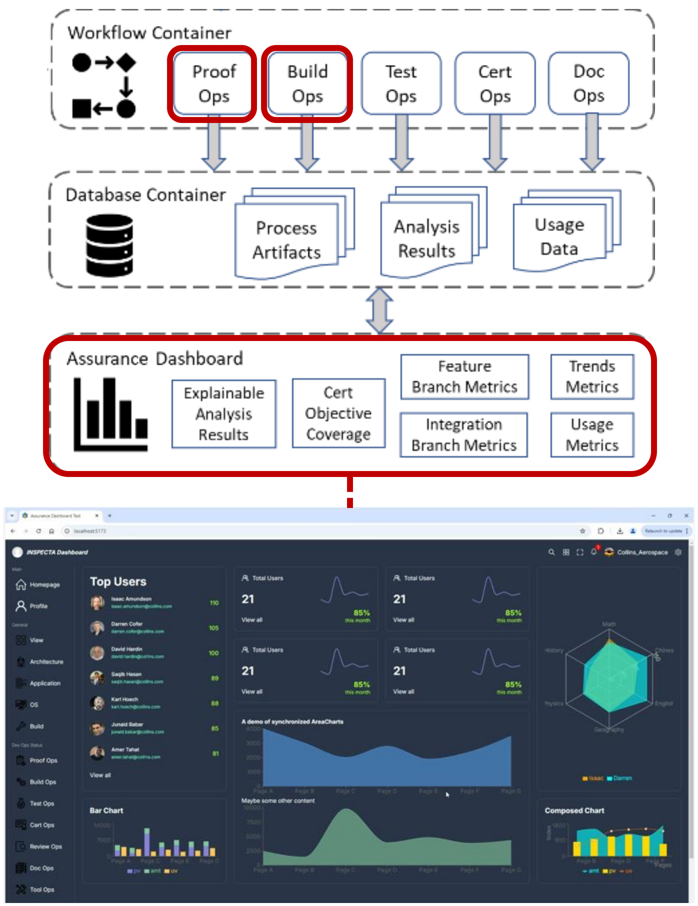
INSPECTA

Industrial-Scale Proof Engineering for Critical Trustworthy Applications





Cloud-based DevOps



Role of AGREE Copilot (AGREE-Dog):

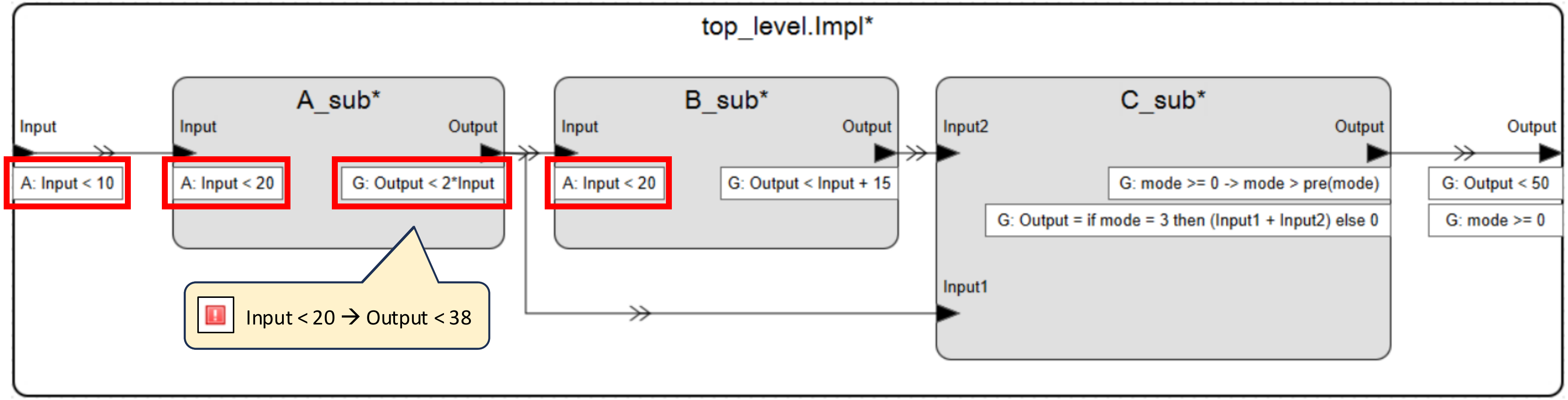
→ Creates a trust fabric for AI-accelerated **MBSE** engineering.

Collins Aerospace
An RTX Business

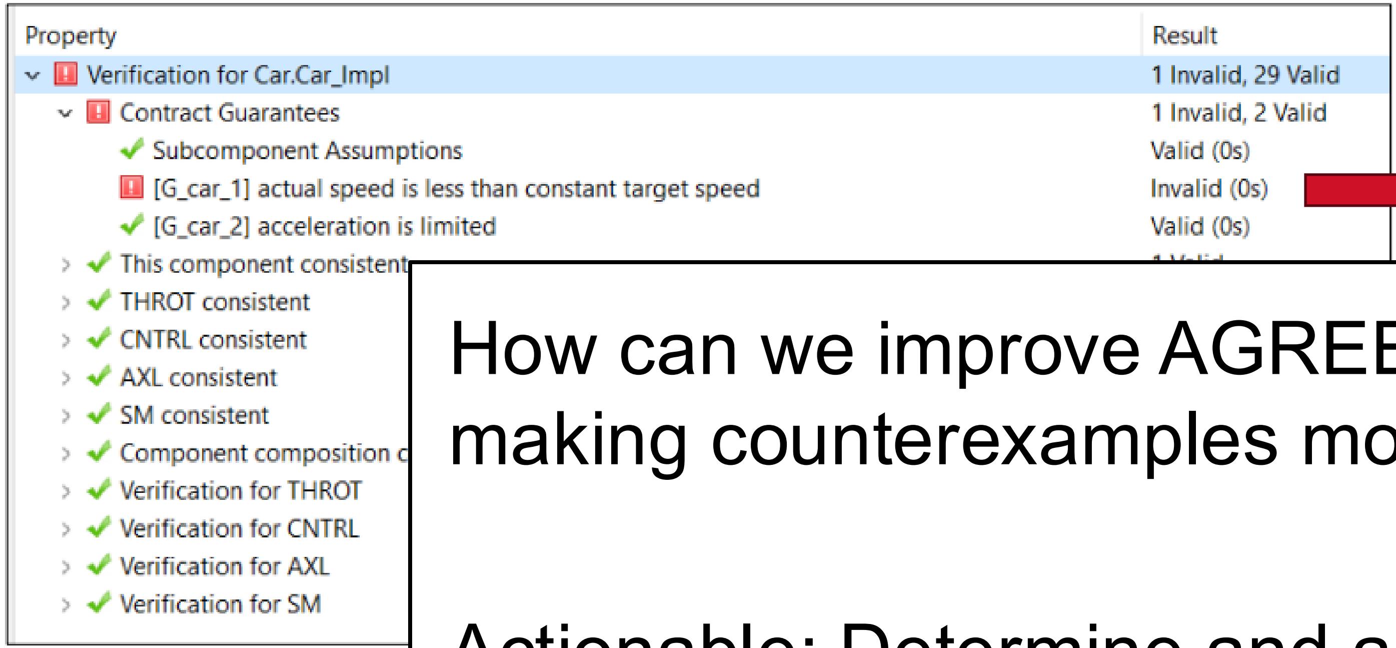
Compositional Reasoning with AGREE

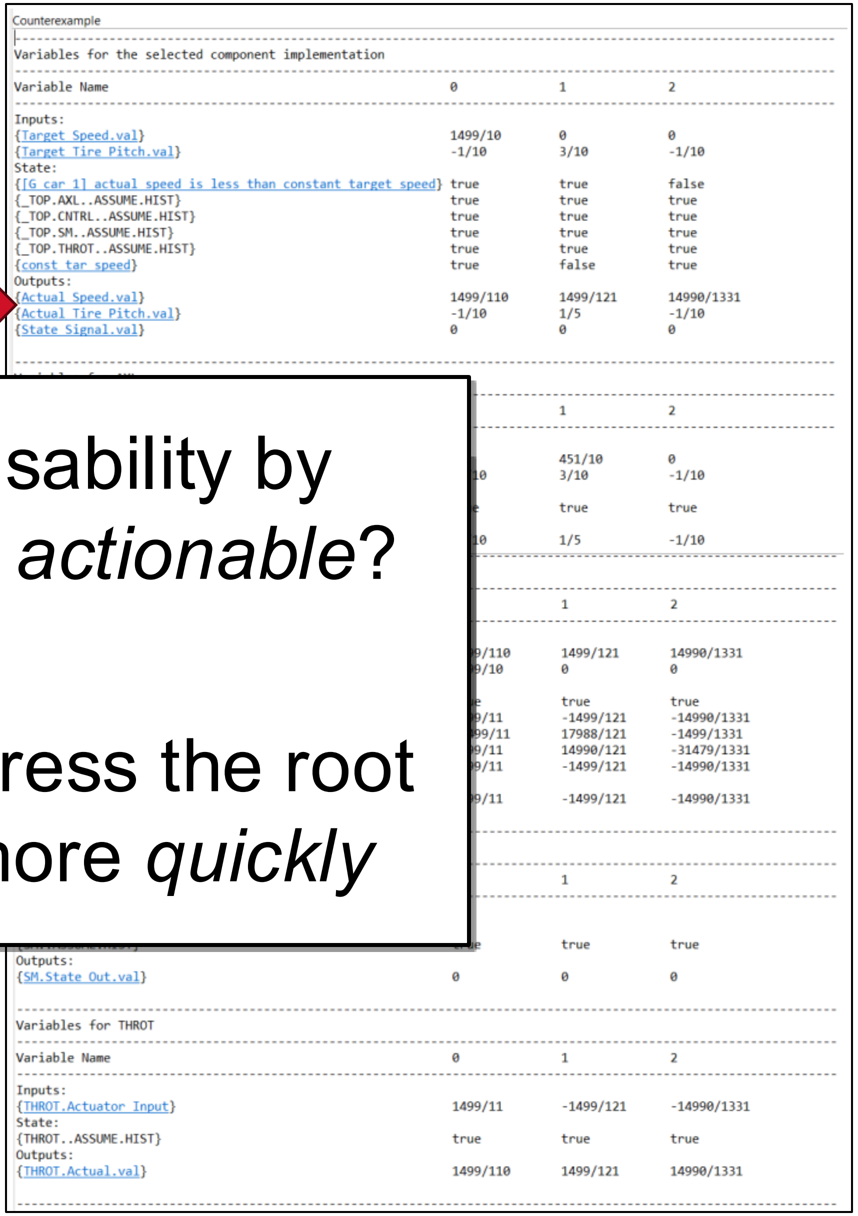
Assume Guarantee Reasoning Environment

- Assume-Guarantee annex for AADL architecture models
  - Assumptions describe the expectations that a component has on the environment
  - Guarantees describe bounds on the behavior of the component when assumptions are valid
- Compositional analysis to prove correctness of:
  - Component interfaces (component assumptions are satisfied by upstream guarantees)
  - Component implementations (component assumptions and subcomponent guarantees satisfy guarantees)



AGREE Counterexamples





How can we improve AGREE usability by making counterexamples more *actionable*?

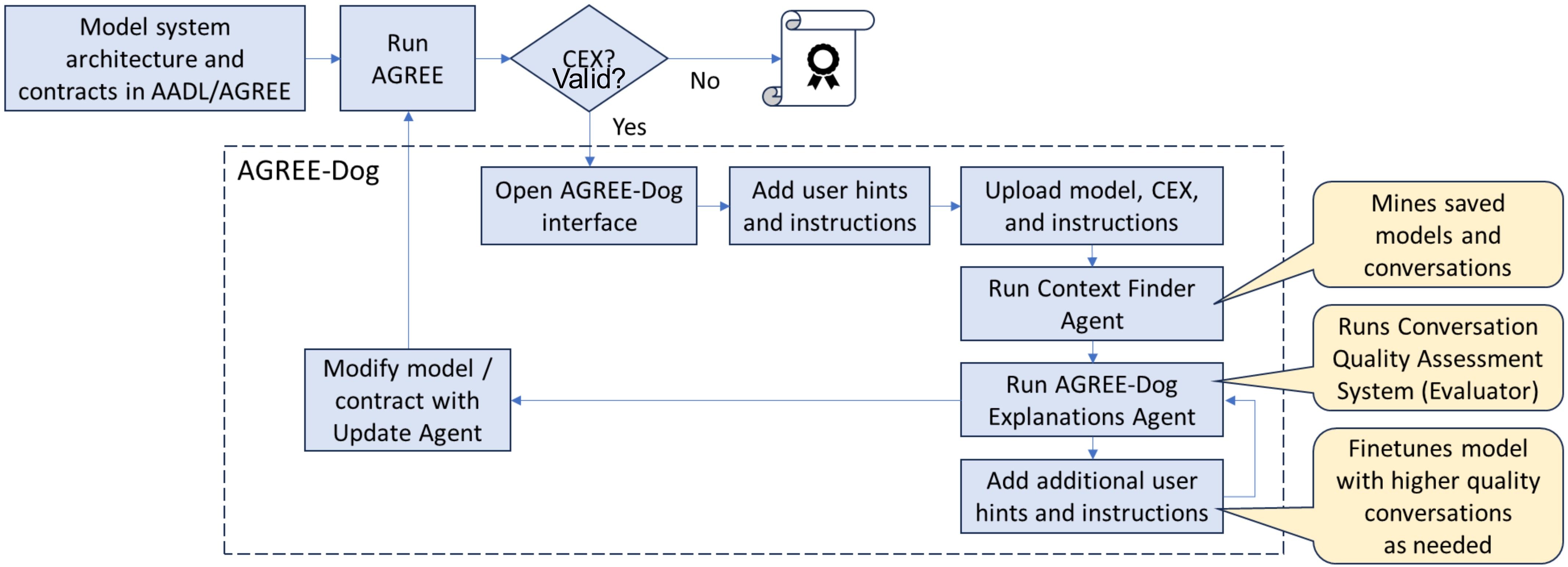
Actionable: Determine and address the root cause of the counterexample more *quickly*

Before AgreeDog: costly and error-prone manual interpretation of counterexamples, requirements, logs, and models — sometimes taking **days or weeks** to complete repairs!

Collins Aerospace
An RTX Business

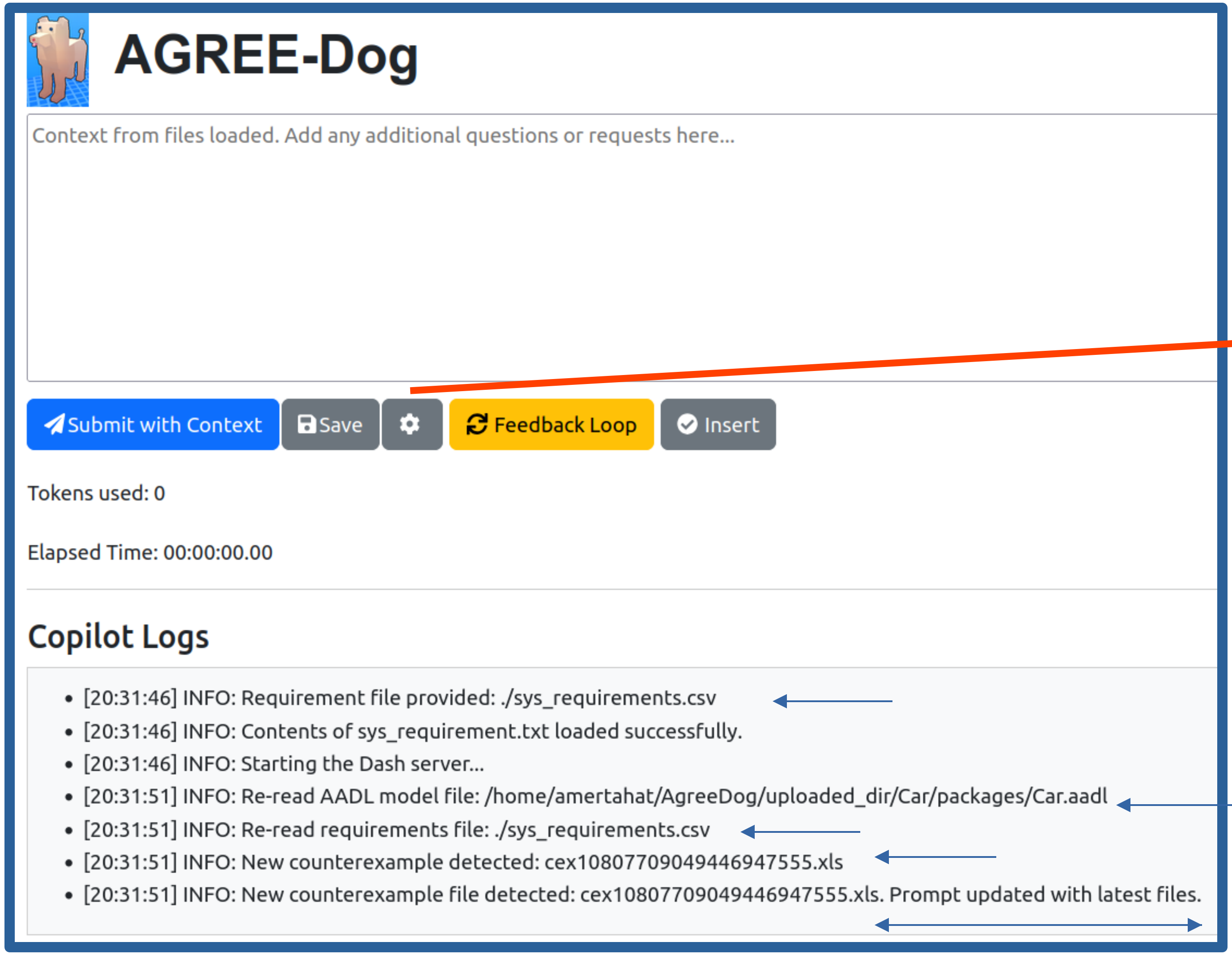
Toward Trusted Gen AI in the MBSE Lifecycle

- Early trials of **AGREE-Dog**—have reduced system-level repair cycles from **days/weeks** to **minutes/seconds**, delivering substantial time savings.
- Powered by GPT-4o, O3, and GPT-4.1 multimodal models.



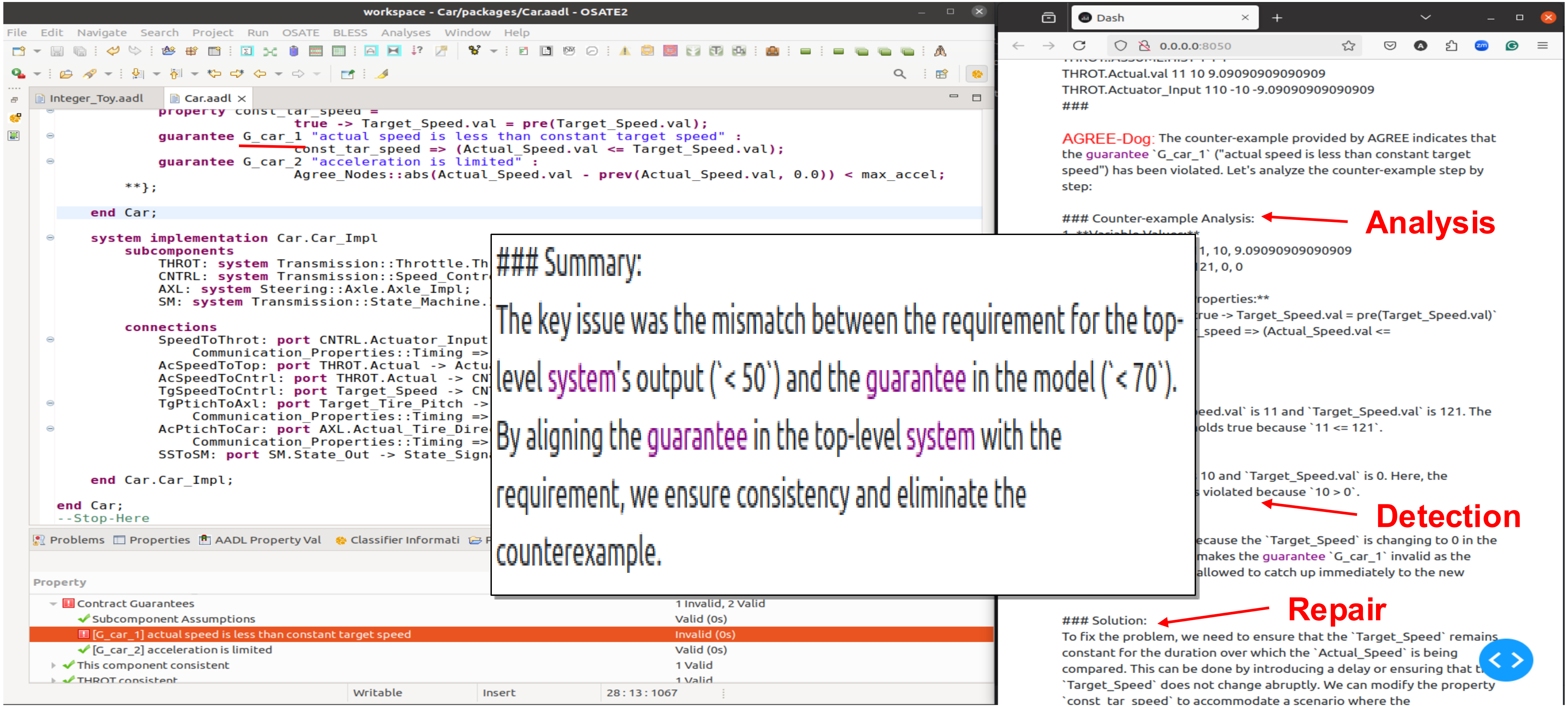
Collins Aerospace
An RTX Business

Toward Trusted Gen AI in the MBSE Lifecycle



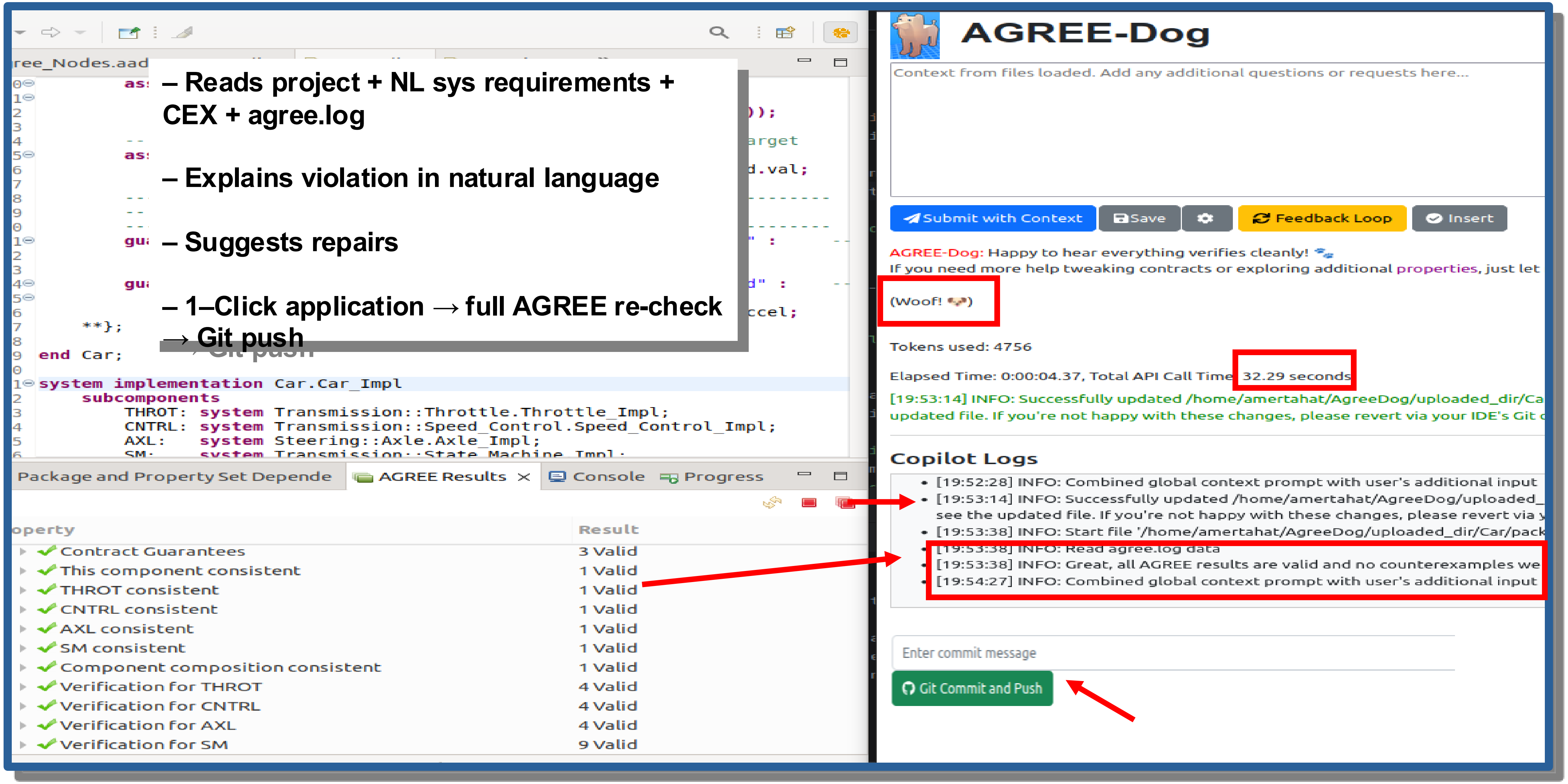
Advanced settings menu

AGREE-Dog: Counterexample Analysis and Resolution



Toward Trusted Gen AI in the MBSE Lifecycle

- Reads project + NL sys requirements + CEX + agree.log
- Explains violation in natural language
- Suggests repairs
- 1–Click application → full AGREE re-check → Git push



Collins Aerospace
An RTX Business

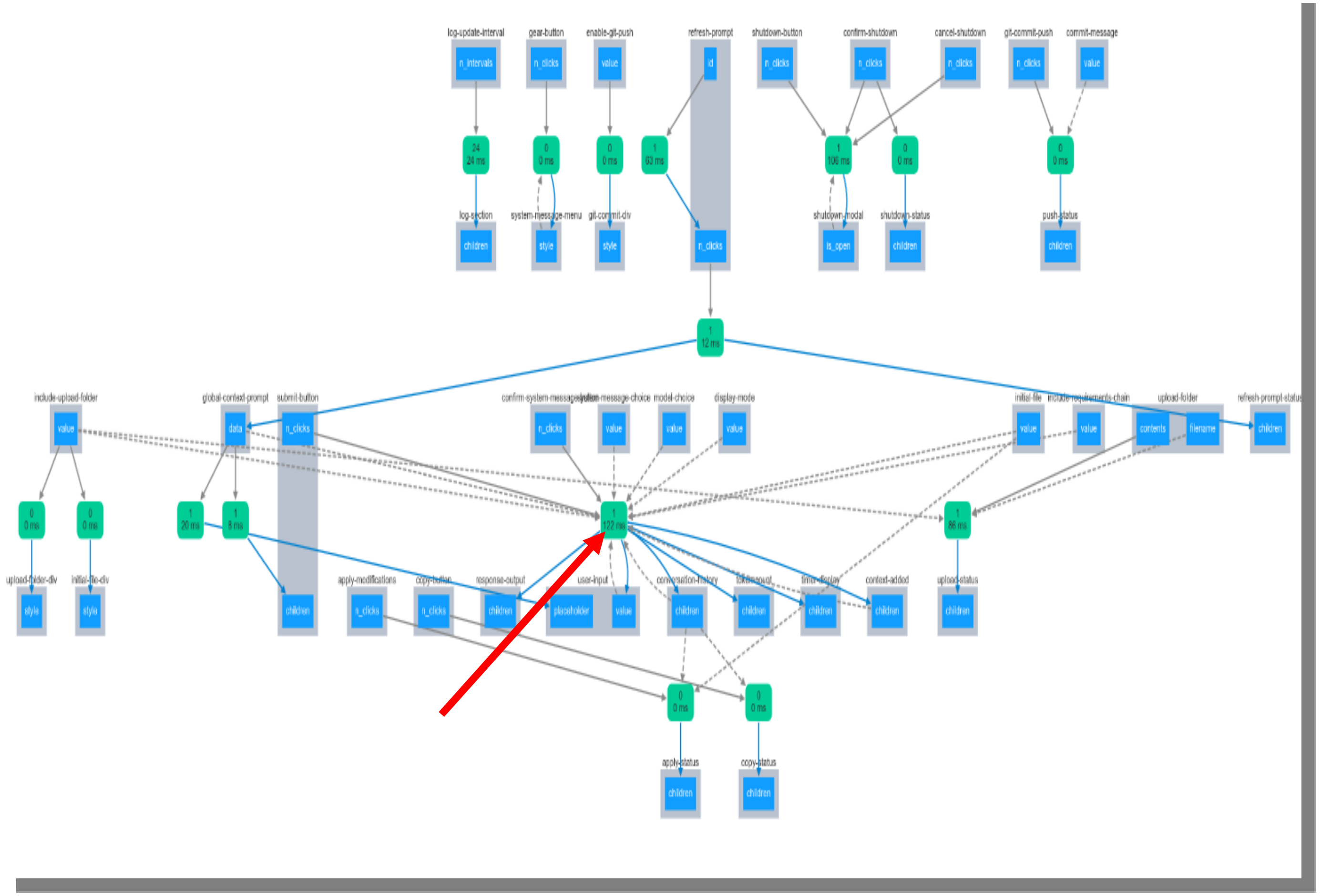
AgreeDog System Call-Graph Complexity

Traditional Workflow: Tens of interdependent calls & manual back-and-forth

AGREE-Dog Approach:

- ✓ Collapses complex interactions into a single recommendation step
- ✓ Minimizes manual navigation of call-graphs and dependencies

→ Dramatic reduction in user effort via push-button UI.



AGREE-Dog significantly automates **16** DevOp and ProofOp ➔ drastic user effort reduction

Ongoing Experimental Evaluations:

- **Test bed**: We created 13 test scenarios for an autonomous-vehicle (Car) AADL model—each scenario comprising 400 LoC models, <100 LoC CEX, <100 LoC sys requirements, and agree Logs. Total ∼**10,000** LoC

- **Fault Injection:** We injected faults into three key components (Top-level Control System, Steering, and Transmission) and ran a full **AGREE-Dog** repair “conversation” for each.

Key Impact & Scalability

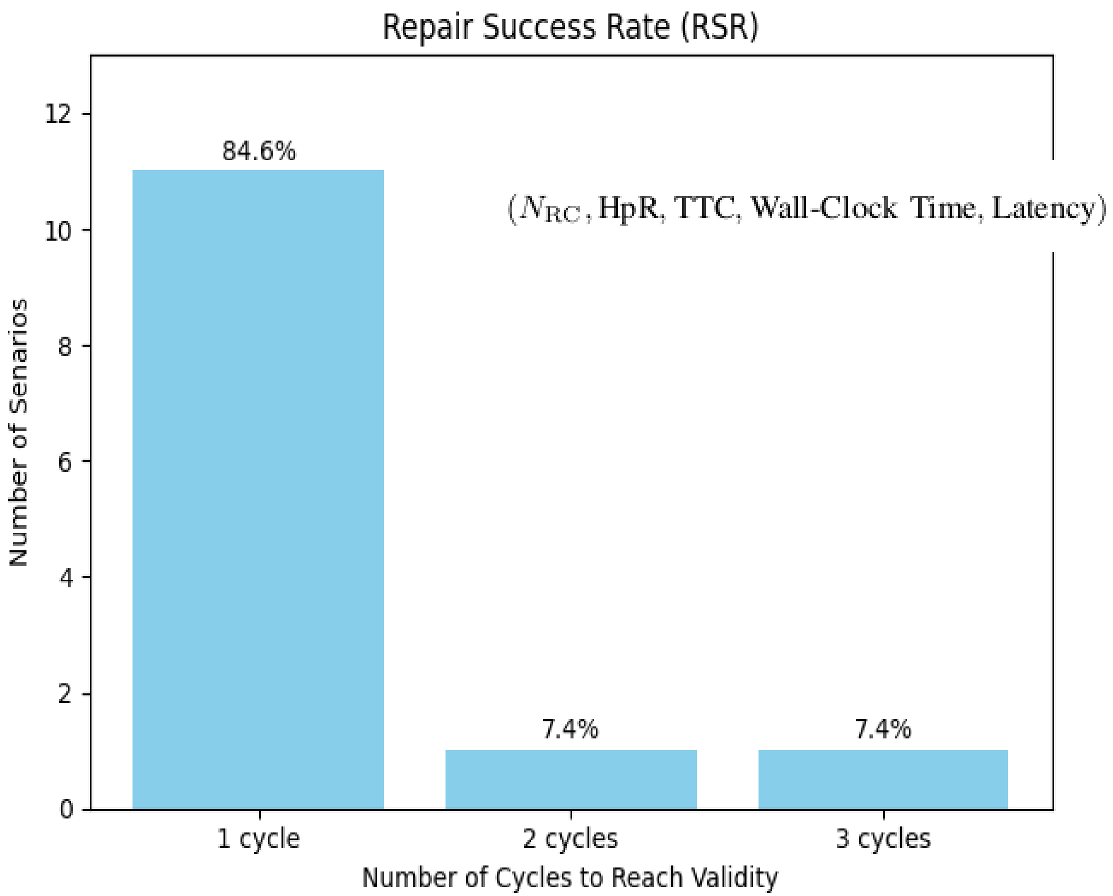


TABLE I: **Structural Metrics Summary**. Token values rounded to the nearest hundred.

Metric	Result
System Validity	100%
Repair Success Rate (Fig. 5)	11/13 (84.6%) in 1 cycle; 1/13 in 2 cycles; 1/13 in 3 cycles
Human Input Ratio (HpR)	$<$ 0.1% of total tokens
AGREE-Dog Input	$>$ 99.9% of total tokens
Token Use (per test suite)	4.8k, 5.5k, 22k tokens

TABLE II: **Temporal Metrics Summary**. Time values rounded to the nearest second.

Metric	Result
Wall-Clock Time (WCT)	Mean: 2:09 min; Median: 1:39 min
Latency (per cycle)	Mean: 22 s; Range: 4–33 s

- ✓ **Key Observation**: Human input dropped below 0.1% of the workflow—yet we still achieved 100% system-wide validity, demonstrating a dramatic reduction in manual effort.

Collins Aerospace
An RTX Business

Challenges

- Contextual constraints
  - GenAI struggles with **accurate** AGREE model explanations due to **limited** context, causing hallucinations
  - ➢ AGREE-Dog Retrieval-Augmented Generation (RAG) System **adjusts context dynamically**, reducing hallucinations and improving explanation accuracy.
- Model repair
  - Repairs can generate multiple solutions, some of which violate system requirements
  - ➢ AGREE-Dog **deploys Agree.log validity analysis** and follows predefined requirements to ensure repairs respect system constraints. (**MMU is useful**)
- Evaluation
  - Manual evaluation limits scalability for large models
  - ➢ A Conversational Quality Assessment System (CQAS) automates performance tracking

Current Limitations

- Manual Solver Selection
  - Users must still choose between JKind, SMT solvers.
  - ➢ We are developing a **multi-agent router** that intelligently selects solvers based on context, performance, and explainability.
- Model repair
  - We are testing multiple-root causes for cex generation, across more complex systems.
  - ➢ The AGREE-Dog more automated navigation between system files.
  - ➢ Judge Worker orchestration: feasibility study.
- RL Integration In Progress
  - Repair adaptation via reinforcement learning is in early stage.
  - ➢ We are incorporating conversation history into our recommendation system and fine-tuning datasets.

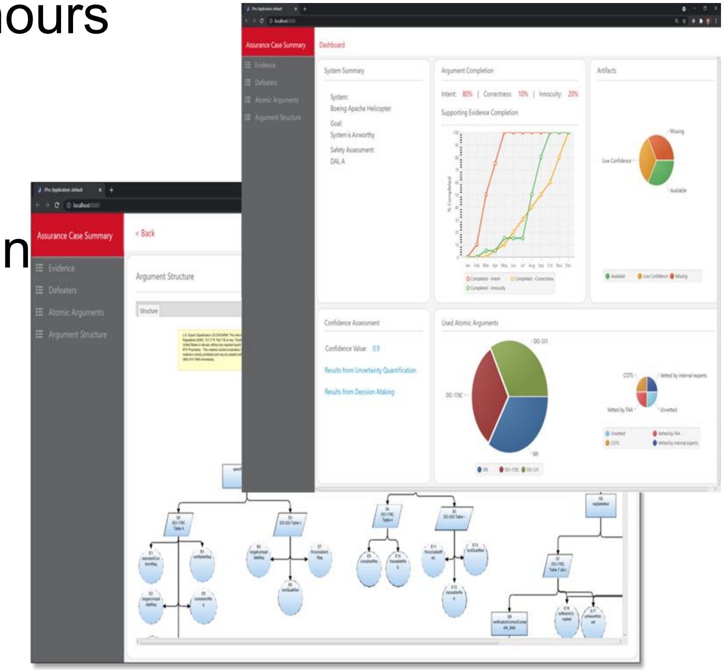
Collins Aerospace
An RTX Business

Conclusions

- **Summary**
  - AGREE-Dog: Push-Button • System-Wide Formally Valid Repairs • Explainable
  - Traceable • Rapid • Human-in-the-Loop
  - Validation cycles reduced from days/weeks to minutes/hours

- Up Next:
  - Federated/**Organizational agents** & solver orchestration
  - CQAS (evaluation metrics)
  - SysMLv2 migration: Embed AGREE-Dog

    like Copilot into next-gen modeling environments
  - Assurance patterns dashboard



Resources

- AGREE-Dog Video Demo:
  - https://www.youtube.com/watchv=2cKRm9fuifs
- INSPECTA project page:
  - https://loonwerks.com/projects/inspecta.html
- AGREE-Dog source code:
  - https://github.com/loonwerks/AgreeDog
- AGREE-Dog-Plugin source code:
  - https://github.com/loonwerks/AgreeDog-plugin.git

Thank You!

This work was funded by DARPA contract FA8750-24-9-1000.

The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

Collins Aerospace
An RTX Business