

# WPI

# Leveraging AI to Manage Technical Debt During Test & Evaluation in Aerospace Systems Engineering

Aerospace Technical-debt Learning and Assessment System (**ATLAS**)

Zak Ouzzif — Enterprise Architect

Dr. Shamsnaz Bhada — PhD. Advisor

AI4SE/SE4AI Workshop 2025

September 2025

# The Scale of Modern Aerospace Complexity

## Programs at Scale

- Millions of LOC; thousands of verification artifacts per milestone.
- Hundreds of interfaces and cross-dependencies. ▶ Dynamic baselines across IPTs and suppliers.

## Resulting Challenge

- Human-only reviews cannot keep up with volume + variance.
- Late discovery drives rework, delay, and risk.
- Debt patterns often span documents and tools.

# Why Technical Debt Matters — Hubble (TD Trade-offs, not “error”)

HST Spherical Aberration (1990): Flawed null-corrector setup passed V&V. Lesson: verification-system debt—accepted compromises not retired pre-launch. [1]

## Verification Debt (Conscious Trade-off)

- Limited redundant optical testing (cost/schedule).
- Single calibration path lowered defect observability.

## Documentation Debt (Latent Liability)

- × Calibration records/anomaly rationale incomplete/siloed.
- Metrology tool error risk not tracked as 'debt'.



## Accumulation → Manifestation

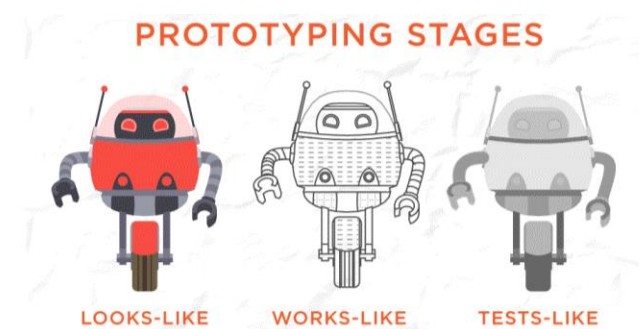
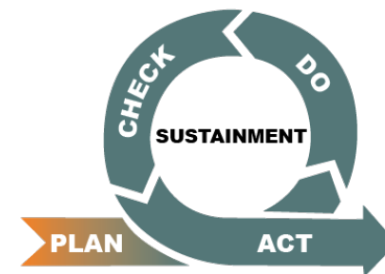
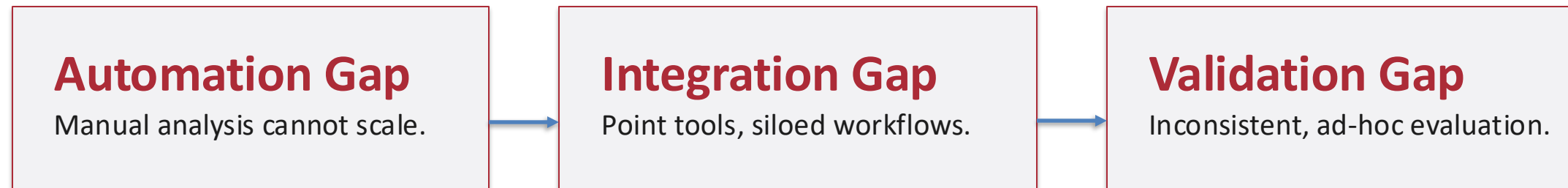
- Debts remained hidden in test infrastructure.
- Surfaced as mission-level risk once in orbit.

## TD ≠ Poor Quality

- TD = known compromise; not equal to failure.
- Failure was not retiring the debt before launch.



# Critical Gaps in Current TD Management



- Most work targets code; SE documentation under-served. [3][4][5]
- Cross-domain dependencies create unique TD patterns. [4][5]

# Research Objectives: From Problem to Solution: TDMF

## • Research Question:

“How can AI, particularly Large Language Models (LLM), during the test and evaluation phase automate Technical Debt identification and integrate it into systems engineering workflows?”

### Quantitative Targets

**Detection Rate:** >95% TD identification [1][4]

**Time Savings:** 75% reduction in review cycles [4]

**Cost Impact:** \$15-25M per program [2]

**Schedule Recovery:** 20% improvement

### ATLAS Capabilities

LLM-powered document analysis [3], [1]

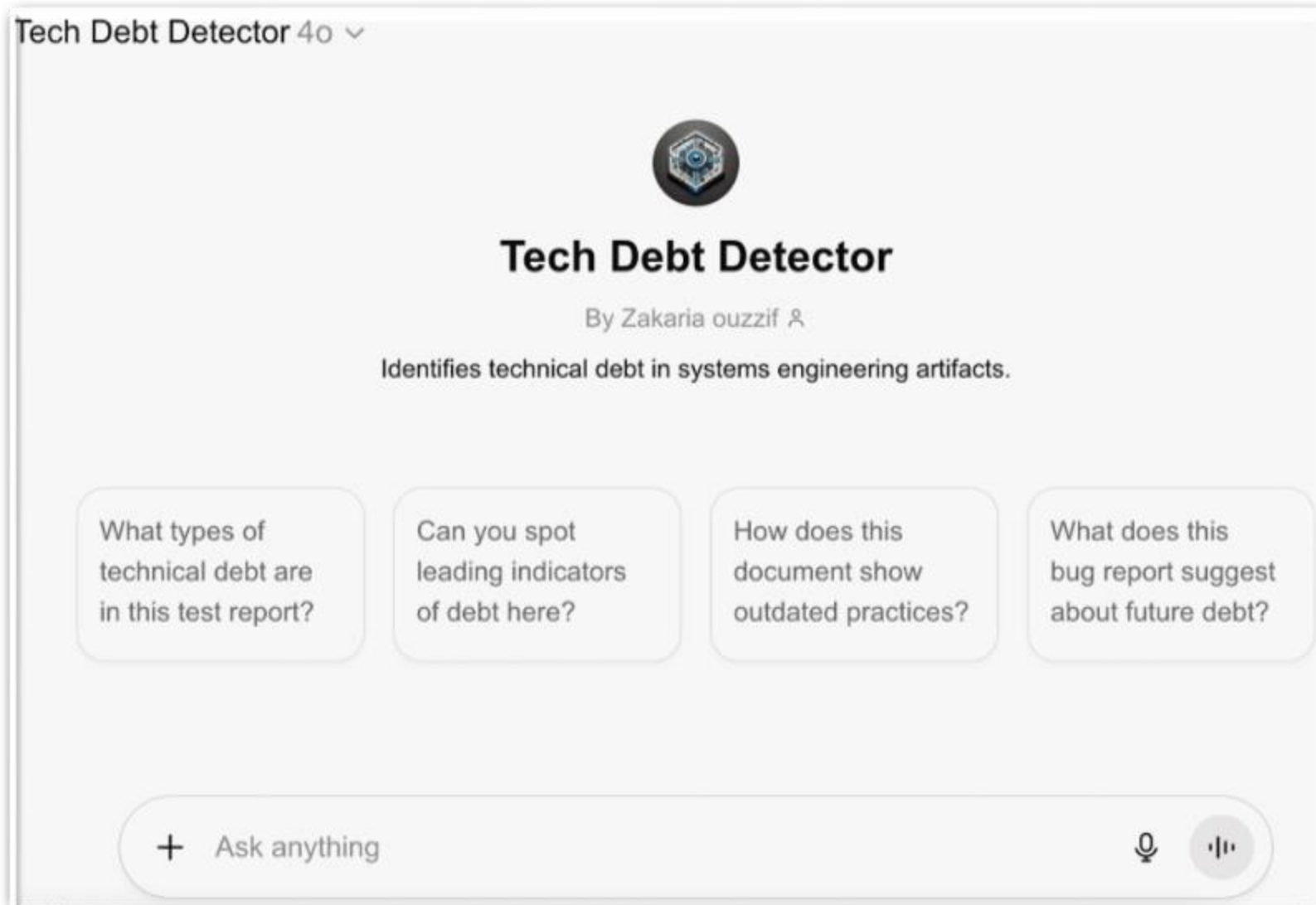
Automated TD categorization & severity [1][6]

Real-time dashboard & alerts [3][2]

Predictive risk modeling [2]

**Personal Driver:** After witnessing 6-month delays from undetected interface TD in missile defense programs, I committed to automating what humans consistently miss

# Aerospace Technical-debt Learning and Assessment System (ATLAS)



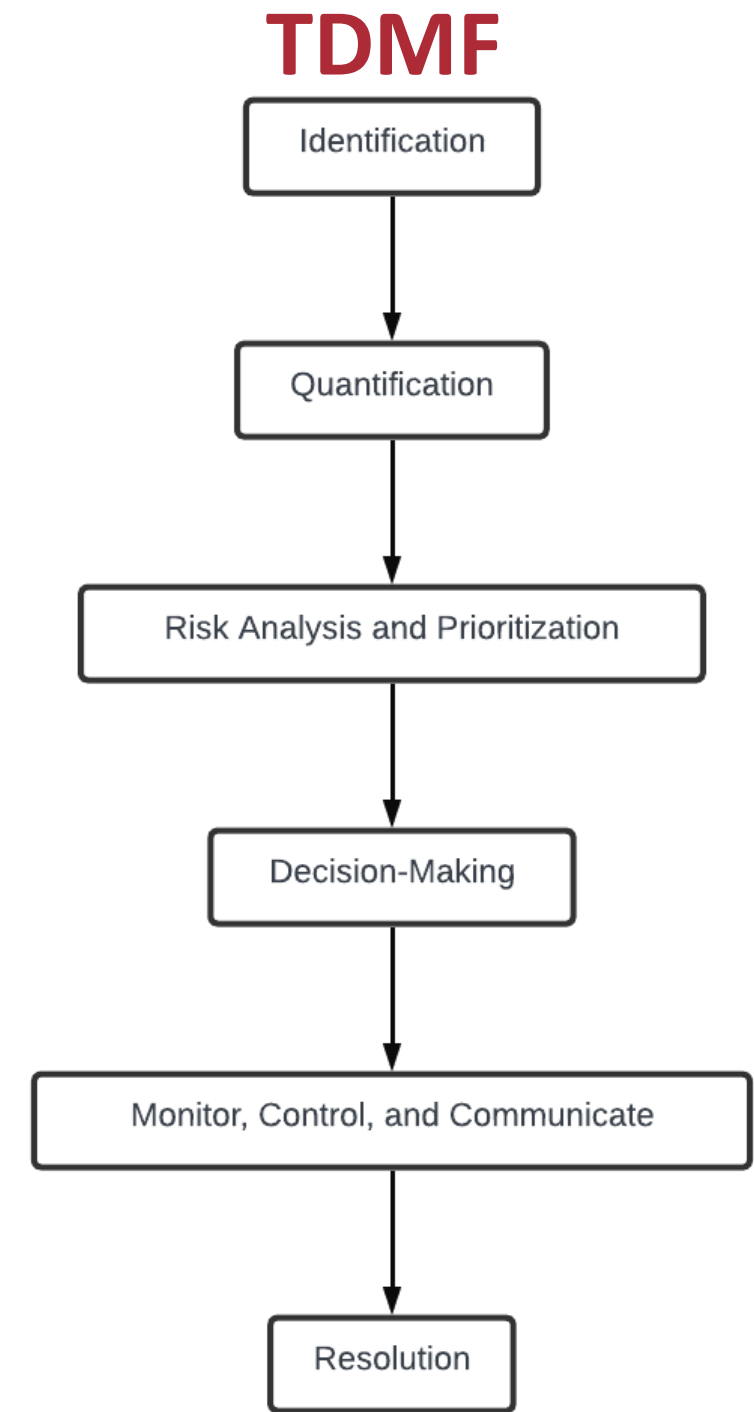
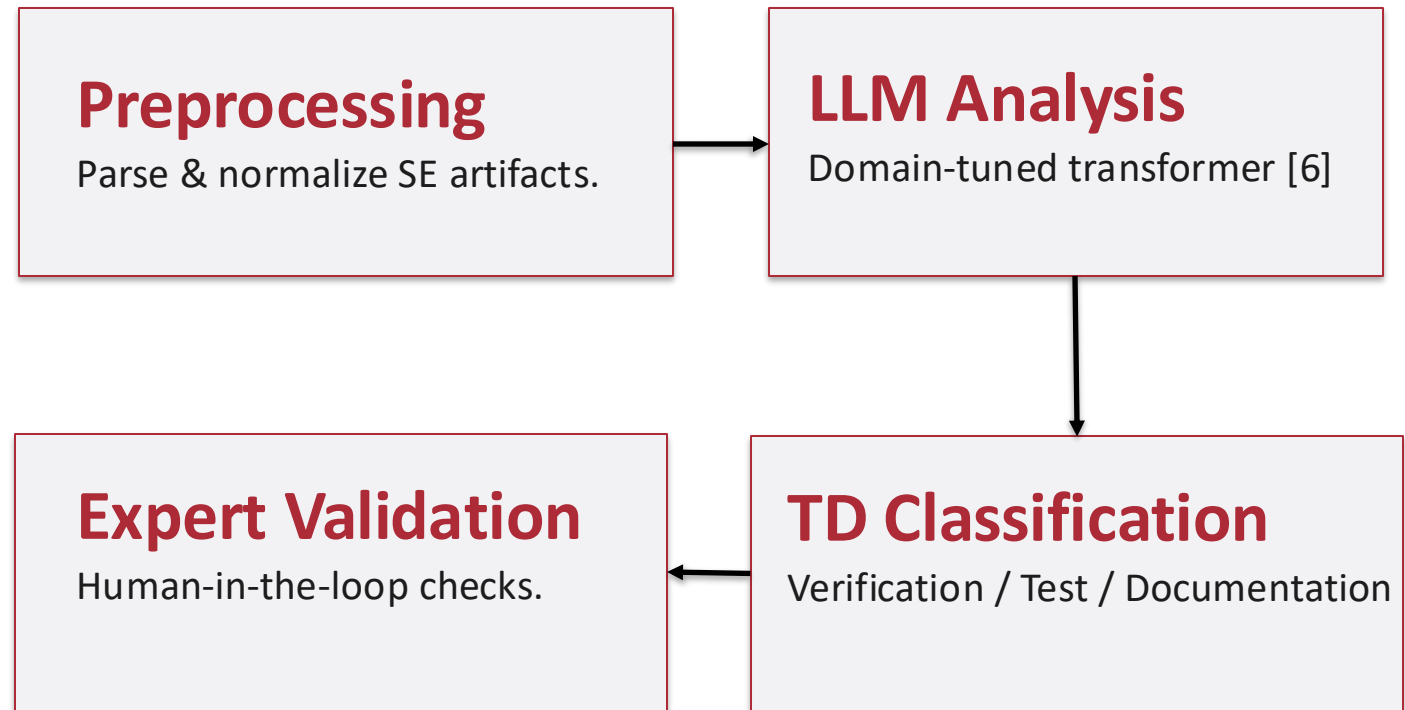
**Automated Analysis:** Parses verification & test artifacts to detect *verification, test, and documentation debt*.

**Expert-Aligned:** Provides context-aware classifications validated by aerospace engineers.

**Workflow Integration:** Exports insights directly into SE tools (DOORS, JIRA, custom databases).

Methodological underpinnings: Powers; LLM-in-SE surveys. [6][12]

# Technical Debt Management Framework (TDMF)



# Three TD Types in T&E

## Verification Debt

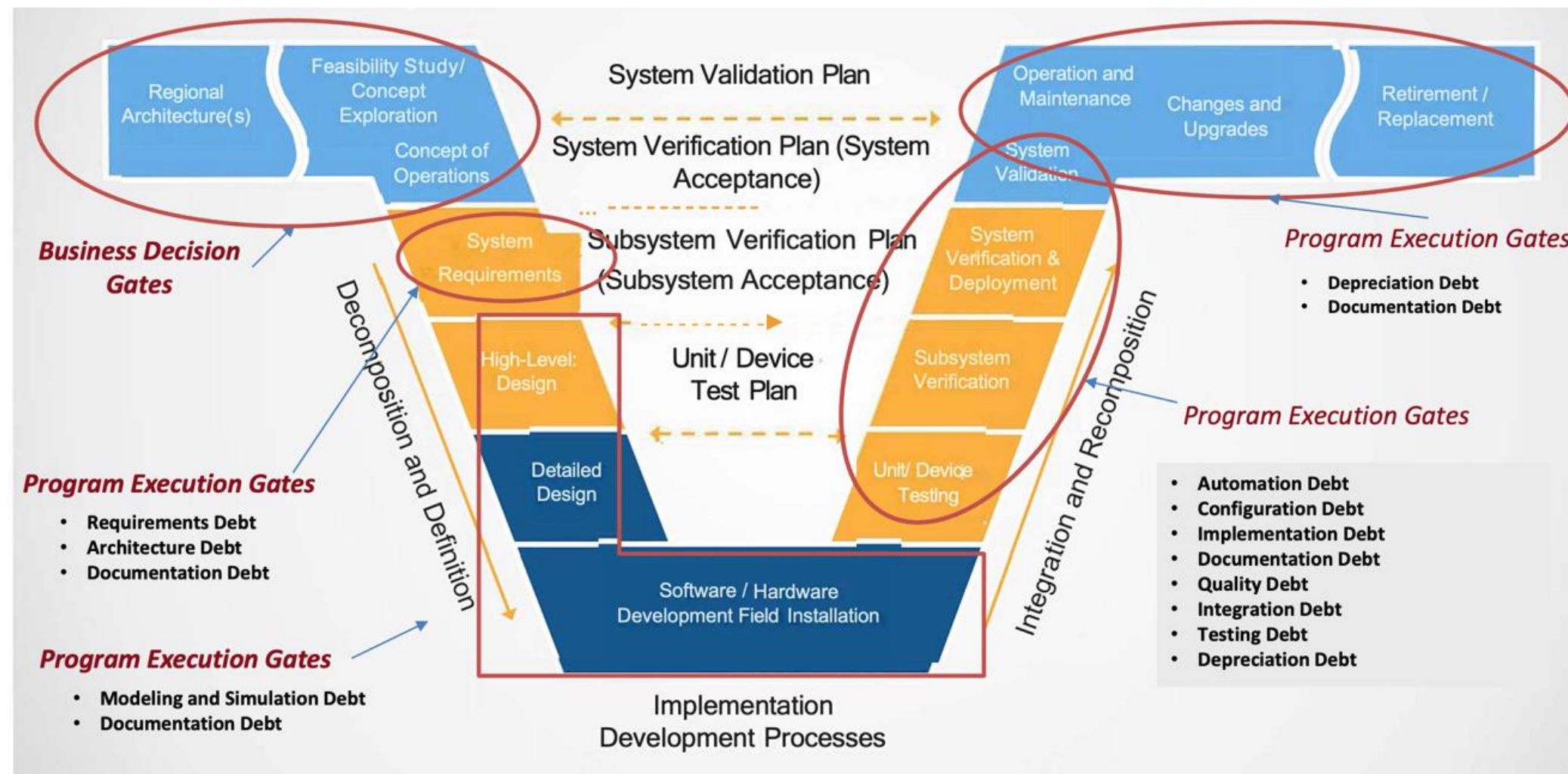
- ▶ Incomplete evidence, missing coverage, unresolved anomalies.

## Documentation Debt

- ▶ Outdated, missing, or inconsistent specs/IFs/acronyms.

## Test Debt

- ▶ Procedural inconsistencies; undocumented equipment substitutions.



▶ Informed by Kruchten et al., Li et al., Kleinwaks et al. [3][4][5]

# Technical Debt Accumulation in V-Model Lifecycle

## Left Side (Design)

### Requirements TD: 15%

Ambiguous specs, missing traces

### Design TD: 25%

Interface mismatches, assumptions

### Implementation TD: 20%

Workarounds, shortcuts

## Right Side (V&V)

### Test TD: 30%

Incomplete coverage, deferred tests

### Verification TD: 35%

Unresolved anomalies, waivers

### Documentation TD: 40%

Missing updates, version conflicts

Key Insight: TD compounds exponentially from left to right in V-Model — ATLAS catches it early when remediation costs are 10x lower [2]

# Validation Methodology

Component	Details
Experts	25 SE professionals (avg 12 yrs)
Agreement	$\kappa = 0.82$ (substantial)
Historical	NASA mission documentation with known TD manifestations
Contemporary	eVTOL Vehicle-Agnostic IFR test plans
Metrics	F1, precision/recall, efficiency

Methodological underpinnings: Powers; LLM-in-SE surveys. [6][12]

# Quantitative Performance

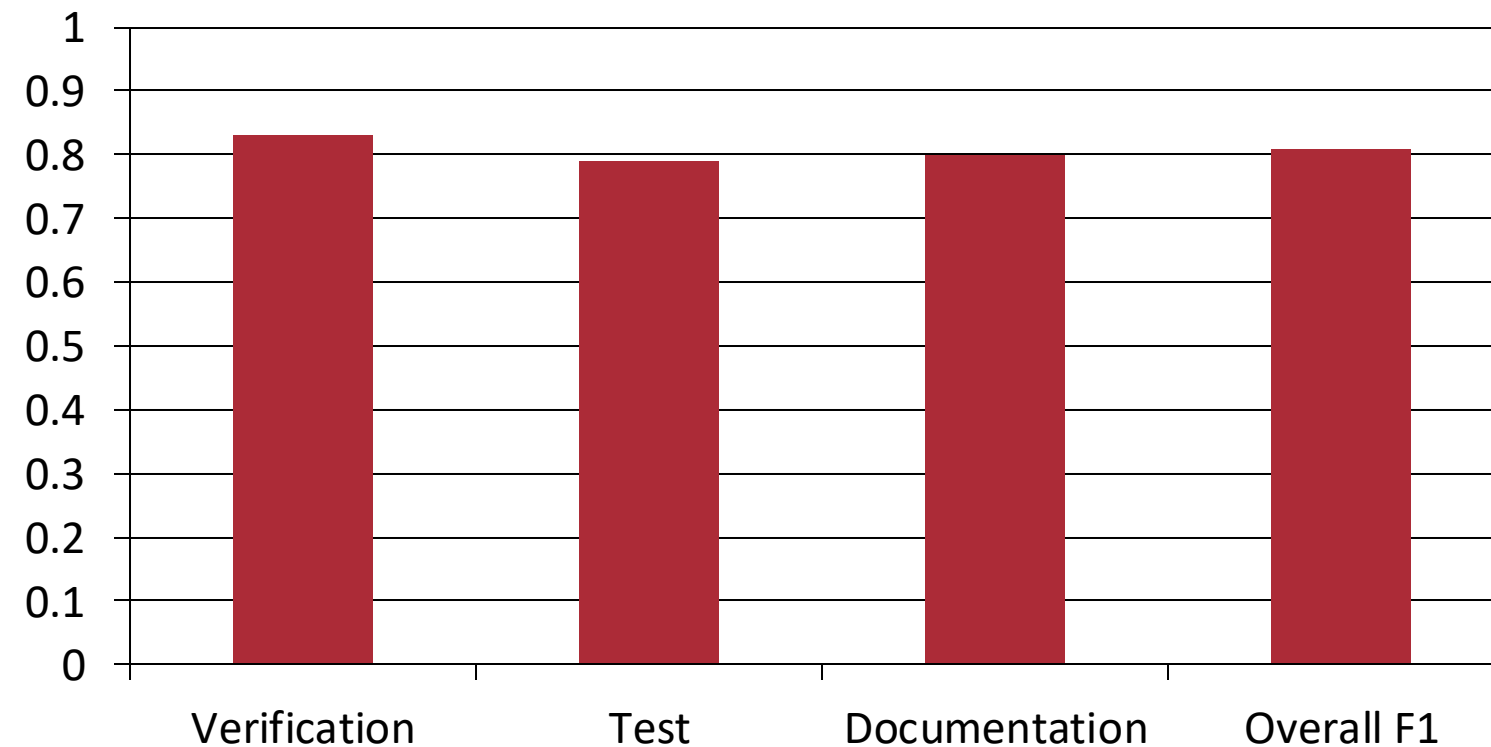
## Efficiency

- ▶ 45% time reduction vs. manual review.

## Expert Agreement

- ▶ 87% validation agreement.

F1 Score



All metrics from this work. [11]

# Integration with V-Model SE



- ▶ Seamless at decision gates; minimal process change.
- ▶ Compatible with DOORS, JIRA, custom DBs. [8]
- ▶ Aligned with DoD Digital Engineering & NASA NPR 7123.1B. [9][10]

# Industry Impact & Relevance



## Cost-Benefit

- 15–25% rework reduction;
- \$7.5–25M savings per major program.



## Risk

- Earlier detection of integration issues.
- Fewer late-stage defects.
- Improved schedule predictability.

See GAO; DoD Digital Engineering. [2][9]

# Key Research Contributions

## Novel Domain

- Human-AI collaboration in SE workflows. [3][9][10]
- LLM-based document analysis [3], [1]
- Context-aware TD classification [1]
- Expert-validated ground truth [1]
- Practical deployment pathways [2]

## Validated Framework

- Empirical results with expert ground truth . [11]
- Integrate with existing SE workflows [2]
- Validate against real aerospace data [1], [2]

## Integration Method

- LLM applied to SE T&E artifacts (beyond code). [5][6]

## Open Questions

- Calibration across programs; auditing AI suggestions.

# Future Directions

## Technical

- Real-time TD monitoring during test execution
- Predictive models for TD accumulation
- Cross-program learning capabilities
- Multi-modal analysis (diagrams + text)

## Applications

- Defense systems integration
- Autonomous vehicle certification
- Medical device verification
- Critical infrastructure systems.

## Vision:

Establish TDMF as standard practice for AI-assisted verification in safety-critical systems engineering

# Key Takeaway

AI-driven TD management empowers aerospace programs to move from reactive fixes to proactive, system-wide risk mitigation, reducing rework, cost, and mission risk

Contact: [zouzzif@wpi.edu](mailto:zouzzif@wpi.edu) — References next slide

# References

- [1] T. Bahill and S. J. Henderson, “Requirements development, verification, and validation exhibited in famous failures,” *Systems Engineering*, vol. 8, no. 1, pp. 1–14, 2005.
- [2] U.S. Government Accountability Office, “F-35 Joint Strike Fighter: DOD Needs to Update Modernization Schedule and Improve Data on Software Development,” GAO-21-226, 2021.
- [3] P. Kruchten, R. L. Nord, and I. Ozkaya, “Technical debt: From metaphor to theory and practice,” *IEEE Software*, vol. 29, no. 6, pp. 18–21, 2012.
- [4] Z. Li, P. Avgeriou, and P. Liang, “A systematic mapping study on technical debt and its management,” *J. Syst. Softw.*, vol. 101, pp. 193–220, 2015.
- [5] H. Kleinwaks, A. Batchelor, and T. H. Bradley, “Technical debt in systems engineering: A systematic literature review,” *Systems Engineering*, vol. 26, no. 4, pp. 428–440, 2023.
- [6] X. Hou et al., “Large language models for software engineering: A systematic literature review,” *ACM Trans. Softw. Eng. Methodol.*, 2024.
- [7] X. Zhang, A. E. Hassan, and Y. Zou, “An empirical study on using artificial intelligence to detect technical debt,” *IEEE Trans. Softw. Eng.*, vol. 48, no. 6, pp. 2053–2071, 2022.
- [8] IBM, “Rational DOORS Technical Specification,” 2023.
- [9] Department of Defense, “Digital Engineering Strategy,” 2018.
- [10] NASA, “NASA Systems Engineering Processes and Requirements,” NPR 7123.1B, 2013.
- [11] Z. Ouzzif, “Enhanced Technical Debt Management Framework (TDMF) — Extended Abstract,” AI4SE/SE4AI Workshop, 2025. (This work.)
- [12] D. M. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation,” *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.