MAVS-GC: Governance-First AI for Failure-Mode Control

A concise research overview for external review and potential support

Prepared by Saif Malik - MAVS Research Program

Mathematical Formulation (MAVS-GC)

System tuple:

$$M = (X, Phi, F, G, A, W, P, Theta, Pi)$$

Core pipeline:

$$x \to Phi(x)=phi \to \{f_i(phi)\} \to r_i \to z \to a \to w_{i,m} \to theta \to R \to Pi$$

Definitions:

$$s_i = f_i(phi) \text{ in } [0,1]$$

$$r_i = 2s_i - 1 \text{ in } [-1,1]$$

$$z = (g_1(phi),...,g_k(phi)) \text{ in } R_{>=0}^k$$

$$a = A(z), \text{ where A is monotone}$$

$$w_i = W(i, phi, z)$$

$$m = max_l \; p_l(phi) \text{ in } [0,1]$$

$$theta = Theta(a,m) = theta_0 + lambda\, a - delta\, m$$

$$R(x) = sum_i \; w_i \; r_i$$

$$Pi(R,theta,a) = 1[a < tau\_hard] * 1[R >= theta]$$

$$\text{Equivalently: } Pi = 0 \text{ if } a >= tau\_hard; \text{ otherwise } Pi = 1[R >= theta]$$

Auditable trace:

$$(r, w, z, a, m, theta, tau\_hard, R, Pi)$$

Key properties:

All-speak evaluation: every specialist evaluates every input.

Monotone safety: higher diagnostic severity cannot make acceptance easier.

Bounded mitigation: mitigation cannot override hard veto because hard veto is inside Pi.

Governance separability: A or Theta can change acceptance without retraining specialists.

Executive Summary

MAVS, the Multi-Adaptive Vetting System, is a governance-first AI architecture. Its core claim is not that a single model becomes universally more accurate, but that explicit governance over specialist outputs can improve how systems behave when evidence becomes uncertain, contradictory, corrupted, or unstable.

The MAVS-GC formulation separates specialist prediction from output governance. All specialists evaluate every input; diagnostics produce red flags; severity is aggregated; contextual weights and bounded mitigation influence a governed threshold; and the final decision is made through an auditable consensus trace. This distinguishes MAVS from static ensembles and routing-based Mixture-of-Experts systems.

The current research program has completed a Foundation Arc, a synthetic validation chapter, and three real-benchmark programs: Chapter 10A for clean predictive correctness, Chapter 10B for robustness under corruption, and Chapter 10C for reproducibility and stability. The strongest empirical result is Chapter 10B: under specialist-failure and high-corruption regimes, Pure MAVS-GC substantially reduced unsafe acceptance while maintaining high accuracy.

Chapter	Question	Result
1-8	Can MAVS be specified as a formal governance architecture?	Foundation Arc completed: mission, thesis, identity, formal objects, metrics, theorem roadmap, cost model.
9	Do MAVS-GC mechanisms behave according to intended formal semantics?	Synthetic validation completed; governance altered decisions, preserved hard-veto behavior, and produced complete traces.
10A	Does governance improve clean predictive correctness?	Negative-to-mixed: competitive, but not universally superior.
10B	Does governance fail more safely under adverse conditions?	Supported and verified; strongest evidence layer, especially under specialist failure.
10C	Does governance preserve reproducibility and stability?	Clean effects limited; corruption-condition stability evidence supported.

1. What MAVS-GC Is

A MAVS system is represented as M = (X, Phi, F, G, A, W, P, Theta, Pi). X is the input space, Phi is a shared feature map, F is a set of always-on specialists, G is a diagnostic system, A aggregates diagnostic flags into severity, W rebalances specialist influence, P provides bounded mitigating evidence, Theta maps severity and mitigation into an acceptance threshold, and Pi produces the final decision.

The key architectural distinction is regulated consensus. Specialists provide calibrated scores s_i in [0,1], converted into supports r_i = 2s_i - 1. The system computes a governed consensus R = sum_i w_i r_i and accepts only when R meets the governed threshold theta. The trace exposes r, w, z, a, m, theta, R, and the final decision.

This means MAVS-GC is better interpreted as a governance framework than as a fixed detector set. The current benchmarks evaluate one concrete configuration, not the entire space of possible diagnostics, severity operators, mitigation strategies, rebalancers, and threshold policies.

2. Evidence Path: From Formalism to Benchmarks

The Foundation Arc establishes the research identity: intelligence generation and intelligence governance are distinct concerns. MAVS elevates governance into a first-class computational object rather than treating final acceptance as a static threshold after model scoring.

Chapter 9 then isolates governance behavior using synthetic specialists instead of trained models. This design removed model-training artifacts and tested whether diagnostics, severity, mitigation, hard veto, and trace generation behave as intended. In the false-positive trap, mean aggregation accepted 100% of unsafe cases, static weighted aggregation accepted 85%, while governance mechanisms materially reduced unsafe acceptance. Hard-veto compliance reached 100%, with no observed violations.

Chapter 10 moved to real benchmark datasets: Breast Cancer Wisconsin, Adult Income, Credit Card Fraud, and Bank Marketing. The comparison systems were Single Model, Mean Ensemble, Static Weighted Ensemble, Veto MAVS, and Pure MAVS-GC. This produced three subprograms: 10A for clean accuracy, 10B for robustness under corruption, and 10C for reproducibility and stability.

Program	Main interpretation
10A - Accuracy	MAVS-GC is competitive but does not establish universal clean-condition predictive superiority. It changes the error profile: often increasing precision or reducing false positives while reducing recall/F1.
10B - Robustness	MAVS-GC demonstrates the clearest empirical advantage: it frequently fails more safely under corruption, especially under specialist failure.
10C - Reproducibility	Clean-condition reproducibility improvements are limited, but stability preservation becomes stronger as corruption increases.

3. Chapter 10B: Highlighted Robustness Results

Chapter 10B is the main empirical signal to highlight. The hypothesis was that MAVS-GC fails more safely under adverse conditions. The verified result supports this hypothesis across multiple datasets, corruption families, benchmark splits, governance traces, robustness curves, and evaluation metrics. The strongest pattern appeared under specialist-failure corruption.

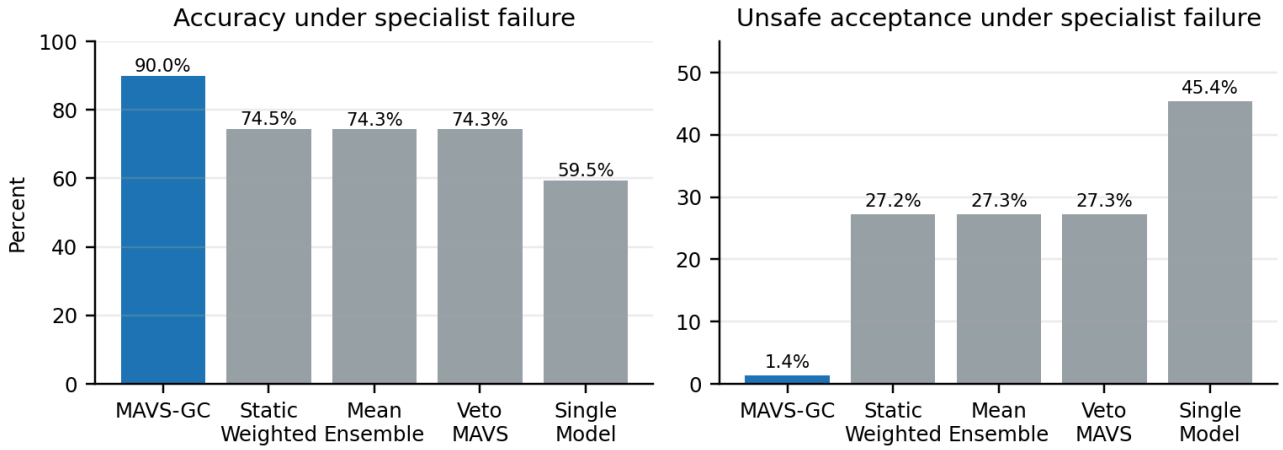

Figure 1. Under specialist failure, Pure MAVS-GC maintained 89.95% accuracy and 1.35% unsafe acceptance. Ensemble baselines clustered near 74% accuracy and roughly 27% unsafe acceptance, while the single-model baseline had 59.46% accuracy and 45.42% unsafe acceptance.

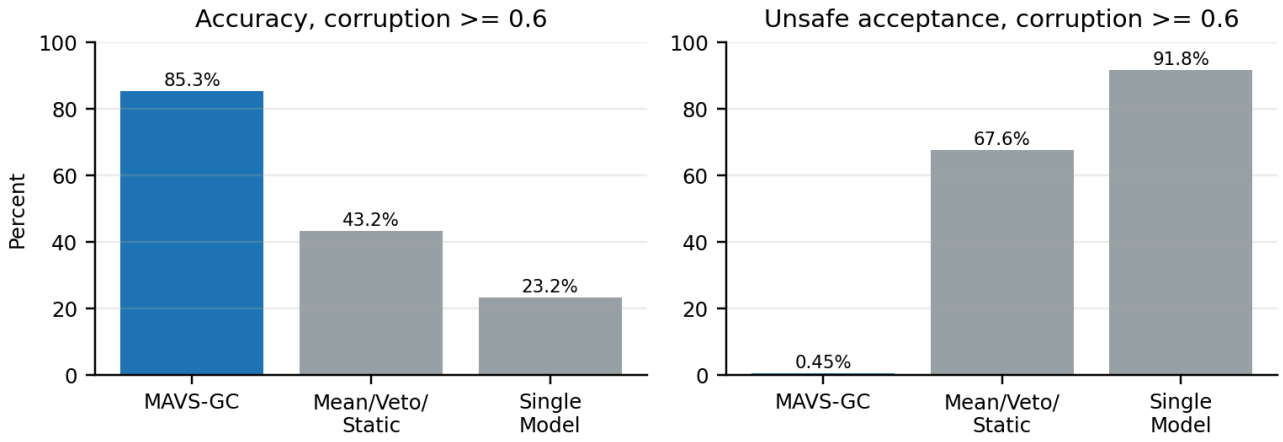

Figure 2. Under high corruption (corruption >= 0.6), Pure MAVS-GC maintained 85.30% accuracy with 0.45% unsafe acceptance. Mean/Veto/Static baselines were near 43.24% accuracy with 67.61% unsafe acceptance; the single model dropped to 23.22% accuracy with 91.78% unsafe acceptance.

Regime	Pure MAVS-GC	Baselines	Relative signal
Specialist failure	89.95% accuracy; 1.35% unsafe acceptance	Mean/Veto ~74.31% accuracy; ~27.29% unsafe acceptance; Single Model 59.46% / 45.42%	Unsafe acceptance ~20.2x lower than Mean/Veto and ~33.6x lower than Single Model.
High corruption >= 0.6	85.30% accuracy; 0.45% unsafe acceptance	Mean/Veto/Static 43.24% accuracy; 67.61% unsafe acceptance; Single Model 23.22% / 91.78%	Unsafe acceptance ~149x lower than ensemble-like baselines and ~202x lower than Single Model.

The central scientific interpretation is not that MAVS-GC is universally more accurate. It is that governance changes failure behavior. Under stress, MAVS-GC rejects more cautiously, suppresses unsafe acceptance, and preserves decision quality more effectively than the evaluated aggregation baselines. The Chapter 10B report therefore identifies MAVS-GC as a failure-management, robustness, and safety-oriented governance architecture rather than a pure accuracy-maximization architecture.

4. What 10A and 10C Add to the Interpretation

Chapter 10A is important because it prevents overclaiming. Under clean benchmark conditions, MAVS-GC did not demonstrate universal predictive-correctness superiority. It improved accuracy over Veto MAVS in 2 of 8 comparisons, over Static Weighted

Ensemble in 0 of 8 comparisons, and produced positive metric deltas in 79 of 288 benchmark comparisons. This positions MAVS-GC as competitive under normal conditions, not dominant.

Chapter 10C adds the stability story. Clean-condition reproducibility gains were limited, but corruption-condition stability effects were stronger. Mean corrupted prediction stability was 0.971615 for Pure MAVS-GC versus 0.952713 for the comparison mean; decision stability was 0.975770 versus 0.958762; consensus stability was 0.979332 versus 0.963946; and trace stability was 0.967976 versus 0.959693 for Veto MAVS. The chapter concludes that MAVS-GC preserves behavioral consistency under stress rather than universally reducing variance.

5. Current Limitations and Support Needed

The current evaluation is rigorous but bounded. It uses four tabular datasets, a fixed suite of corruption families, controlled split and audit structure, reproducibility manifests, and verified artifact trails. It does not establish production-scale behavior, LLM-agent behavior, universal robustness superiority, or cross-domain generalization beyond the tested benchmark suite.

The most valuable next step is external-scale validation: larger datasets, additional modalities, LLM/agent specialist settings, adversarial expansions, ablation matrices, and independent replication. Support from a research lab could help test whether the observed failure-management and stability-preservation effects survive under larger-scale and more realistic AI systems.

Specific support requested: research feedback, compute credits, review of experimental design, guidance on scalable evaluation settings, and potential collaboration on applying governance-first evaluation to LLM agents or safety-critical multi-model systems.

One-Sentence Summary

MAVS-GC is a governance-first architecture whose current evidence suggests competitive clean-condition performance, strong failure-mode control under corruption, and stronger stability preservation under adverse conditions, with Chapter 10B providing the clearest verified empirical signal.

Repository Links

Chapter 9 Synthetic Benchmark Program:
https://github.com/MAVS-RESEARCH/MAVS-Chapter-9

Chapter 10A Accuracy Benchmark Program:
https://github.com/MAVS-RESEARCH/MAVS-Chapter-10A

Chapter 10B Robustness Benchmark Program:
https://github.com/MAVS-RESEARCH/MAVS-Ch10B

Chapter 10C Reproducibility and Stability Program:
https://github.com/MAVS-RESEARCH/MAVS-Ch10C

Source Artifacts Used

- MAVS Foundation Arc, Chapters 1-8.
- MAVS formal definition and MAVS-GC calculus.

- Chapter 9 Synthetic Benchmark Program and completion report.
- Chapter 10A Accuracy Benchmark completion report.
- Chapter 10B Robustness completion report, robustness highlights, and corrected final numbers report.
- Chapter 10C Reproducibility and Stability completion report.