MATERIAL SIGNATURES FOR ANTINEUTRINO-BASED DETECTABILITY OF COVERT FISSILE PRODUCTION IN FUSION REACTORS

Anonymous authors
Paper under review

ABSTRACT

Antineutrino monitoring is a promising route for early safeguards signals, but fusion-adjacent deployment requires robustness to prior disagreement, detector nuisance variability, and transfer stress. We present a hybrid formal-empirical framework that combines contradiction-aware sequential calibration, an information-decomposed minimum detectable diversion criterion, and finite robust detector-material co-design. The method is grounded in source-lineaged reactor-spectrum and safeguards detection formalisms and extends them with manuscript-defined robust operators. We execute a CPU-only benchmark over 17,280 runs spanning prior families, drift rates, standoff distances, and detector resolutions with matched false-alarm operating points across comparators. Formal consistency checks pass for all theorem-linked symbolic obligations (4/4). Empirically, the robust method achieves strong delay robustness (delay win rate 0.9965 versus single-prior likelihood ratio; median delay ratio 0.6943 versus fixed-threshold test-statistic baseline), supporting the detectability-improvement claim under the tested open-parameterized regime. However, calibration and transfer closure remain conditional: calibration violation rate is 0.9524, leave-one-prior-out FAR inflation p95 is 1.2871, and hard-versus-easy transfer degradation ratio is 1.5402. These outcomes establish a defensible contribution boundary: robust detectability gains are supported, while policy-grade calibration and transfer claims require targeted recalibration and transfer-stability refinement. The manuscript reports both supported and mixed claims through explicit claim-evidence linkage, enabling reproducible iteration rather than optimistic overclaiming.

1 INTRODUCTION

Antineutrino monitoring has matured from a physics demonstration into a safeguards-relevant sensing modality because the signal is weakly shieldable, physically tied to fissile evolution, and now measurable with higher spectral and temporal fidelity than early deployment studies assumed (Bernstein et al., 2002; 2006; Oguri et al., 2014; Boireau et al., 2016; Kneale et al., 2023; Alekseev et al., 2025). At the same time, fusion-adjacent fuel-cycle pathways have reopened an old but unresolved question: whether material and blanket choices that alter neutron economy can be detected early enough to constrain covert fissile production scenarios under realistic detector noise and calibration drift (Ball et al., 2024; Ruegsegger et al., 2023; Shi & Peng, 2016; Zhao et al., 2013; Zheng et al., 2012; Reed et al., 2012; McCowan, 1978). The practical issue is not whether one can separate idealized signal classes in simulation; the issue is whether operational false-alarm control, delay, and transfer stability survive contradictory source priors and nuisance misspecification in the regimes where decisions are made.

The source lineage used here is explicit: we build on remote-monitor sensitivity analyses for detector realism (Kneale et al., 2023), high-statistics reactor flux/spectrum calibration for baseline priors (An et al., 2025a), anomaly synthesis for contradiction-aware prior families (Zhang et al., 2024), and fusion-blanket proliferation studies for material-parameterized covert-production pathways (Ball et al., 2024).

This work studies that operational regime. We treat detector efficiency, background level, energy resolution, and standoff as explicit nuisance quantities; we keep reactor and fusion-adjacent settings open and parameterized; and we evaluate detection performance under matched false-alarm operating points so comparator differences are not driven by threshold mismatch. The manuscript is intentionally hybrid: it combines formal guarantees for the calibration and ranking operators with executed simulation evidence and symbolic theorem checks. This combination is necessary because purely empirical gains can be unstable under prior contradiction, while purely formal guarantees can fail to capture transfer degradation in high-drift conditions.

The broader relevance extends beyond one safeguards niche. The methodological pattern, namely contradiction-aware prior families, minimax threshold calibration, and explicit robustness accounting, is applicable to other domains that combine sparse physics observability with high-consequence early warning decisions, including process monitoring and infrastructure anomaly detection under domain shift. In that sense, this manuscript contributes both a domain-specific antineutrino result and a generalizable uncertainty-governed detection design template.

Our contributions are:

- We define a contradiction-aware robust threshold operator for sequential antineutrino detection and prove existence with conservative false-alarm control under explicit regularity assumptions.
- We derive an information-decomposed minimum detectable diversion criterion showing when joint count-shape inference strictly improves detectability over count-only information.
- We formalize a finite robust detector-material co-design objective with a stability functional and prove optimizer existence plus bounded stability score.
- We execute a CPU-feasible benchmark over 17,280 runs across prior families, drift regimes, standoff distances, and detector resolutions, with matched-false-alarm comparator policy and bootstrap uncertainty.
- We report claim-wise support status with negative-result logging: one core detectability claim is supported, while calibration and transfer-stability claims remain conditional and are bounded by explicit caveats.

The evidence summary is concrete. At matched operating points, the robust method improves delay substantially in most regimes (delay win rate 0.9965 and median delay ratio 0.6943 against a fixed-threshold test-statistic comparator), but calibration robustness remains incomplete (calibration violation rate 0.9524, leave-one-prior-out FAR inflation $p95 = 1.287$), and worst-case transfer degradation is elevated (hard-over-easy ratio 1.540). Symbolic checks for the theorem-critical algebraic and monotonicity constraints all pass $(4/4)$. These combined results motivate a bounded conclusion: robust detectability gains are credible under the tested open-parameterized assumptions, but policy-grade calibration and transfer closure require targeted follow-up experiments.

2 Related Work and Novelty Boundary

2.1 Reactor Spectrum Baselines and Anomaly Tensions

Modern reactor antineutrino baselines are anchored by high-statistics flux and spectrum measurements, especially recent Daya Bay analyses and related instrumentation programs (An et al., 2025a;b; 2023a; Lombardo, 2023; An et al., 2023b). These studies improved precision and clarified how isotopic evolution and detector calibration shape inferred spectra. At the same time, anomaly-centered literature emphasizes unresolved model discrepancy across conversion and aggregate-beta pipelines (Zhang et al., 2024; Sonzogni et al., 2023; Mention et al., 2011; Mueller et al., 2011; Huber, 2011). The agreement is strong on one point: isotope fractions and detector systematics materially influence predicted observables. The disagreement remains consequential: competing anomaly attributions imply different prior families and therefore different thresholds for controlling false alarms under misspecification.

Our approach adopts this contradiction as a first-class design input rather than suppressing it. Instead of choosing a single prior lineage, we use a bounded prior family and calibrate for robust control. This differs from common single-prior detector tuning and is the primary formal bridge from anomaly debates to operational alarm design.

2.2 Safeguards Detection and Sequential Decision Rules

Operational safeguards studies established that antineutrino streams can track reactor state changes and support diversion-sensitive inference under explicit nuisance handling (Bernstein et al., 2002; 2006; Oguri et al., 2014; Christensen et al., 2015; Boireau et al., 2016; Stewart et al., 2019; Ishizuka et al., 2023; Kneale et al., 2023; Alekseev et al., 2025). Methodologically, likelihood-ratio and sequential detection frameworks are recurrent, often with count-dominant statistics and site-specific assumptions. More recent analyses show that spectral information can improve discrimination when calibration is controlled (Kneale et al., 2023; Stewart et al., 2019; Oguri et al., 2014). However, the magnitude and persistence of gains vary with background drift, resolution, and transfer regime, and not all studies enforce comparator fairness through matched false-alarm targets.

This manuscript contributes an explicitly matched operating-point comparison protocol and a claim-evidence closure strategy that separates supported gains from conditional ones. That distinction is important because apparent improvements can disappear once threshold mismatch is removed.

2.3 Burnup Uncertainty and Fusion-Adjacent Transfer

Burnup-aware analyses and isotopic evolution studies provide the mechanistic basis for time-varying signal formation (Kopeikin et al., 2024; Barresi et al., 2024; An et al., 2023b; Sonzogni et al., 2023; Jaffke, 2015; Hayes et al., 2012; Alekseev et al., 2025). In parallel, fusion and hybrid blanket studies characterize material-dependent breeding pathways and neutron-economy effects that are relevant to covert production risk (Ball et al., 2024; Ruegsegger et al., 2023; Shi & Peng, 2016; Zhao et al., 2013; Zheng et al., 2012; Reed et al., 2012; Sheng-chun, 2010; McCowan, 1978). The central limitation is data realism: many fusion-adjacent studies are conceptual or parameterized, whereas safeguards deployment evidence is largely reactor-domain specific.

The novelty boundary of this paper is therefore precise. We do not claim plant-specific predictive validity for proprietary fusion operations. We claim a reproducible, open-parameterized framework that can rank and stress-test detector-material strategies while preserving explicit caveats on external validity. This boundary is scientifically defensible and prevents over-interpretation.

2.4 Gap Statement

Prior literature leaves three linked gaps that motivate this work. First, there is no unified calibration theorem for contradiction-aware prior families integrated with operational FAR constraints. Second, many detectability claims rely on empirical separation without a formal criterion that attributes gains to count and shape information channels. Third, co-design recommendations are frequently nominal and lack explicit robustness objectives with stability auditing under drift. The technical core of this manuscript addresses exactly these gaps while preserving source lineage at each borrowed formal component.

3 Research Questions and Evaluation Logic

This manuscript is organized around three research questions that follow directly from the contradiction map and the operational safeguards objective. First, can contradiction-aware calibration preserve false-alarm control while retaining delay performance? Second, does adding shape information produce meaningful detectability gains under matched operating points and nuisance propagation? Third, can detector-material co-design maintain performance when moving from easier to harder nuisance regimes? The questions are linked; answering only one is not sufficient for reliable monitoring guidance.

The first question concerns calibration reliability under model disagreement. Prior work often uses a single emission lineage or reports thresholding details incompletely, which makes false-alarm comparisons difficult across studies. By explicitly defining a bounded prior family and evaluating leave-one-prior-out stress, we turn disagreement into a measurable robustness objective instead of an untracked uncertainty source. This design choice is motivated by anomaly and conversion-model tensions in the reactor literature and is essential for transfer-minded safeguards analysis (Zhang et al., 2024; Sonzogni et al., 2023; Mention et al., 2011; Mueller et al., 2011; Huber, 2011).

The second question concerns mechanistic detectability, not only headline metrics. Count-only improvements can occur for reasons unrelated to material signatures, for example due to threshold mismatch or favorable background slices. The information-decomposed MDS criterion in this manuscript was introduced to avoid that ambiguity. It asks whether shape information contributes positive information under the same FAR target and nuisance treatment, which is a sharper question than asking whether one method has a smaller median delay in aggregate.

The third question concerns deployment stability. A method that performs well in nominal settings but degrades strongly under drift and standoff variation can be scientifically interesting yet operationally fragile. We therefore evaluate robust co-design through worst-regime scalarization and a bounded stability diagnostic. This emphasizes resilience over isolated peak performance and aligns with safeguards use where conditions can change faster than recalibration cycles.

3.1 Comparator Philosophy

Comparator selection in this study follows a lineage-aware philosophy. We include methods that represent common operational families: single-prior likelihood ratio, count-only sequential testing, nuisance-penalized spectral fit, fixed-threshold test-statistic detection, and FAR-matched CUSUM. This set is deliberately heterogeneous. It allows us to test whether gains from contradiction-aware calibration persist relative to both classical count-focused methods and spectral-statistical alternatives discussed in modern monitoring studies (Bernstein et al., 2002; 2006; Oguri et al., 2014; Christensen et al., 2015; Boireau et al., 2016; Stewart et al., 2019; Kneale et al., 2023; Alekseev et al., 2025).

Matched operating-point policy is central to fair comparison. Without matching, a method can appear superior by accepting higher false-alarm burden. Because operational burden is itself a mission variable, mismatched comparisons are not informative for safeguards design. All primary claims in this manuscript therefore depend on matched-FAR evaluation, and claims are downgraded when stress criteria fail even if nominal delay gains are large.

3.2 What Counts as Evidence in This Paper

Evidence is stratified into four layers. The first layer is source-grounded formalism: reused equations and assumptions from the acquired corpus. The second layer is manuscript-defined formal extension: threshold, MDS, and co-design operators with theorem statements and proofs. The third layer is executed computation: benchmark metrics, stress slices, and uncertainty summaries. The fourth layer is failure-aware auditing: symbolic checks and negative-result ledgers.

A claim is treated as supported only when these layers are coherent for that claim role. For example, detectability gains require both formal support (positive shape-information criterion) and empirical support (improved delay ratio under matched-FAR policy). Conversely, calibration robustness requires passing empirical stress thresholds in addition to theorem-level internal consistency. This layered policy reduces the risk of selectively citing whichever channel looks favorable.

3.3 Scope Commitments

The scope commitments are explicit and intentionally conservative. We commit to open parameterized scenarios because proprietary plant geometries were unavailable. We commit to CPU-feasible sweeps to preserve reproducibility under realistic resource constraints. We commit to explicit nuisance parameters rather than idealized detector constants. Finally, we commit to reporting negative outcomes as first-class results rather than appendix afterthoughts.

These commitments affect interpretation. They strengthen transparency and reproducibility, but they limit external validity to the tested envelope. The manuscript therefore emphasizes conditional claims where appropriate and reserves policy-grade statements for conditions that are empirically and formally closed.

3.4 Expected Contribution Type

Given these choices, the paper's contribution type is methodological-operational rather than purely theoretical or purely empirical. It contributes a formal decision framework that is directly executable under realistic constraints, then demonstrates where that framework succeeds and where it still fails. This contribution type is suitable for open-question domains where stronger claims require iterative evidence accumulation rather than single-shot benchmark wins.

4 Problem Setting, Symbols, and Assumptions

We consider a monitoring horizon indexed by $t \in \{1, \dots, T\}$ and antineutrino energy $E \in \mathbb{R}_+$. The source mixture model follows established reactor formalism (An et al., 2023b; Mueller et al., 2011):

$$\phi(E, t) = \sum_{i=1}^{K} f_i(t) S_i(E), \tag{1}$$

where $f_i(t)$ are isotope-fraction weights and $S_i(E)$ are isotope-specific spectral templates. We define the expected detector count rate in this work as

$$\lambda(t) = \frac{N_p}{4\pi L^2} \int \epsilon(E; \theta)\, \sigma_{\mathrm{IBD}}(E)\, \phi(E, t)\, \mathrm{d}E + b(t; \theta), \tag{2}$$

where N_p is effective target normalization, L is standoff, ϵ is efficiency, σ_{IBD} is cross-section response, b is background, and θ aggregates nuisance terms (efficiency, resolution, drift, background intensity). Equation equation 2 reuses established observability ingredients (Oguri et al., 2014; Kneale et al., 2023; Alekseev et al., 2025); the precise nuisance parameterization and stress ranges are manuscript-defined.

Hypothesis testing compares baseline operation H_0 to perturbation scenarios $H_1(\delta)$ indexed by diversion scale $\delta \geq 0$. For each prior family element $m \in \mathcal{M}$, sequential evidence is

$$\Lambda_t^{(m)} = \log \frac{p(y_{1:t} \mid H_1, \theta, \pi_m)}{p(y_{1:t} \mid H_0, \theta, \pi_m)}, \tag{3}$$

reusing the likelihood-ratio lineage from safeguards detection (Kneale et al., 2023; Stewart et al., 2019). We define a stopping rule $T_\tau = \inf\{t : \max_{m\in\mathcal{M}} \Lambda_t^{(m)} \geq \tau\}$ and robust false-alarm map

$$F(\tau) = \max_{m\in\mathcal{M}} \Pr_{m,H_0}(T_\tau < \infty). \tag{4}$$

For detectability, we use minimum detectable diversion (MDS), defined in this work as the smallest perturbation magnitude that satisfies both false-alarm and power targets (α, β) under nuisance-propagated evaluation. Under local asymptotic approximation, information separates into count and shape components:

$$I_m = I_m^{\text{count}} + I_m^{\text{shape}}, \quad I_m^{\text{shape}} \geq 0. \tag{5}$$

For co-design, candidate decisions are finite pairs $c = (\text{material}, \text{detector setup}) \in \mathcal{C}$ due CPU-feasible discretization. Regimes $\omega \in \Omega$ encode drift and nuisance settings. We define mission-loss vector

$$J_\omega(c) = \big(D_{\text{det}}(c,\omega), -\text{TPR}(c,\omega), \text{FAR}(c,\omega), \text{MDS}(c,\omega)\big), \tag{6}$$

and robust objective

$$c^\star \in \arg\min_{c\in\mathcal{C}} \max_{\omega\in\Omega} w^\top J_\omega(c), \tag{7}$$

with nonnegative weight vector w. The feasible set is $\mathcal{C}$, and optimality is defined by minimizing worst-regime weighted mission loss under matched operating-point policy.

Assumptions are explicit: (i) the prior family $\mathcal{M}$ is finite and bounded; (ii) nuisance ranges are compact and represented in sweeps; (iii) comparator thresholds are matched by FAR target; (iv) open parameterizations are used instead of proprietary plant geometry. These assumptions bound claims to scientifically defensible scope and are revisited in section 10.

5 Methodology

5.1 Pipeline Architecture and Provenance

The method consists of four modules: a scenario generator, a detector forward model, a robust sequential inference layer, and an audit layer for symbolic and empirical closure. The scenario generator samples standoff, drift, resolution, and prior-family slices under open parameter bounds informed by literature lineages (Ball et al., 2024; Ruegsegger et al., 2023; Shi & Peng, 2016; Zhao et al., 2013; Zheng et al., 2012; Reed et al., 2012). The detector module maps each slice to count and shape observables via equation 1 and equation 2. The inference module applies prior-family sequential evidence from equation 3 with thresholding calibrated through equation 4. The audit module logs acceptance checks, contradiction slices, and symbolic theorem conditions.

This architecture is intentionally modular because each layer has different provenance. Equations equation 1 and equation 3 are borrowed conventions from prior work; robust calibration operator design, information-decomposed MDS ranking, and co-design stability score are newly introduced in this manuscript.

5.2 Robust Threshold Calibration

We define the robust threshold operator

$$\tau^\star = \inf\{\tau \in \mathbb{R}_+ : F(\tau) \leq \alpha\}, \tag{8}$$

where F is given by equation 4. The role of equation 8 is to guarantee conservative operating-point control across prior contradiction families.

Theorem 5.1 (Existence and conservative control)**.** *Assume $F(\tau)$ is nonincreasing, right-continuous, and $\lim_{\tau\to\infty} F(\tau) = 0$. For any $\alpha \in (0,1)$, the set $\mathcal{T}_\alpha = \{\tau : F(\tau) \leq \alpha\}$ is nonempty and $\tau^\star$ in equation 8 exists. Moreover, $F(\tau^\star + \varepsilon) \leq \alpha$ for every $\varepsilon > 0$, and if F is continuous at $\tau^\star$ then $F(\tau^\star) \leq \alpha$.*

Proof. Because $\lim_{\tau\to\infty} F(\tau) = 0 < \alpha$, there exists $\bar\tau$ such that $F(\bar\tau) \leq \alpha$, so $\mathcal{T}_\alpha \neq \emptyset$. Since $\mathcal{T}_\alpha \subseteq \mathbb{R}_+$ is nonempty and bounded below by 0, its infimum exists, giving equation 8. Let $\varepsilon > 0$. By definition of infimum, there exists $\tilde\tau \in \mathcal{T}_\alpha$ with $\tilde\tau \leq \tau^\star + \varepsilon$. Monotonicity gives $F(\tau^\star + \varepsilon) \leq F(\tilde\tau) \leq \alpha$. If F is continuous at $\tau^\star$, then

$$F(\tau^\star) = \lim_{\varepsilon\downarrow 0} F(\tau^\star + \varepsilon) \leq \alpha.$$

Hence all stated properties follow. □

Theorem 5.1 is manuscript-defined; the LR/TS machinery it operates on is borrowed from prior safeguards literature (Kneale et al., 2023; Stewart et al., 2019; Oguri et al., 2014). In method terms, equation 8 is used before every comparator evaluation so delay or TPR improvements are not purchased by hidden FAR inflation.

5.3 Information-Decomposed MDS Criterion

To connect detectability gains to mechanism, we separate count and shape information channels. Under local approximation, we define in this work

$$\mathrm{MDS}_m^{\mathrm{joint}}(\alpha,\beta) = \frac{z_{1-\alpha}+z_{1-\beta}}{\sqrt{I_m^{\mathrm{count}}+I_m^{\mathrm{shape}}}}, \tag{9}$$

$$\mathrm{MDS}_m^{\mathrm{count}}(\alpha,\beta) = \frac{z_{1-\alpha}+z_{1-\beta}}{\sqrt{I_m^{\mathrm{count}}}}. \tag{10}$$

The criterion ranks material-detector choices by their ability to increase usable information under matched constraints.

Theorem 5.2 (Strict MDS improvement condition)**.** *Assume $I_m^{\mathrm{count}} > 0$ and $I_m^{\mathrm{shape}} \geq 0$. If $I_m^{\mathrm{shape}} > 0$, then equation 9 is strictly smaller than equation 10, i.e.,* $\mathrm{MDS}_m^{\mathrm{joint}} < \mathrm{MDS}_m^{\mathrm{count}}$.

Proof. Let $C = z_{1-\alpha} + z_{1-\beta} > 0$. Then

$$\mathrm{MDS}_m^{\mathrm{joint}} = \frac{C}{\sqrt{I_m^{\mathrm{count}}+I_m^{\mathrm{shape}}}}, \quad \mathrm{MDS}_m^{\mathrm{count}} = \frac{C}{\sqrt{I_m^{\mathrm{count}}}}.$$

If $I_m^{\mathrm{shape}} > 0$, then $I_m^{\mathrm{count}} + I_m^{\mathrm{shape}} > I_m^{\mathrm{count}} > 0$, so

$$\sqrt{I_m^{\mathrm{count}}+I_m^{\mathrm{shape}}} > \sqrt{I_m^{\mathrm{count}}}.$$

Dividing positive constant C by larger denominator yields

$$\frac{C}{\sqrt{I_m^{\mathrm{count}}+I_m^{\mathrm{shape}}}} < \frac{C}{\sqrt{I_m^{\mathrm{count}}}},$$

which is the desired strict inequality. □

Theorem 5.2 gives a direct method interpretation: when shape information is reliably positive after nuisance propagation, we expect lower detectable diversion thresholds. This provides formal support for why joint count-shape designs are worth the additional modeling burden (An et al., 2023b; Oguri et al., 2014; Mueller et al., 2011).

5.4 Finite Robust Co-Design and Stability

For deployment guidance, we combine performance and robustness in equation 7. We also define a stability score in this work:

$$\mathrm{Stab}(c) = \frac{1}{|\Omega|}\sum_{\omega\in\Omega} \mathbb{1}\left\{\frac{\mathrm{FAR}(c,\omega)}{\alpha} \leq 1.25\right\}. \tag{11}$$

This score quantifies how often candidate c remains within stress tolerance over regimes.

Theorem 5.3 (Finite robust optimum and bounded stability)**.** *If $\mathcal{C}$ and Ω are finite and all components of $J_\omega(c)$ are finite real numbers, then a minimizer of equation 7 exists. For every $c \in \mathcal{C}$, $0 \leq \mathrm{Stab}(c) \leq 1$.*

Proof. For each fixed c, the set $\{w^\top J_\omega(c) : \omega \in \Omega\}$ is finite and real, so its maximum is finite; denote it by $V(c)$. Because $\mathcal{C}$ is finite, $\{V(c) : c \in \mathcal{C}\}$ is finite and has at least one minimizer $c^\star$. Hence equation 7 attains its minimum. For equation 11, each indicator term is in $\{0,1\}$, so its average over $|\Omega|$ terms belongs to $[0,1]$. Therefore $0 \leq \mathrm{Stab}(c) \leq 1$ for all c. □

Theorem 5.3 is simple but operationally important. It ensures that a CPU-discretized policy search is mathematically well-posed and that stability diagnostics are bounded and interpretable.

Algorithm 1 Contradiction-Aware Robust Monitoring Workflow

1: Define prior family $\mathcal{M}$, nuisance ranges Θ, and regime set Ω from source-grounded bounds.
2: Generate scenario slices and observables using equation 1 and equation 2.
3: For each prior family member, compute sequential evidence using equation 3.
4: Calibrate threshold by computing $F(\tau)$ and selecting $\tau^\star$ via equation 8.
5: Evaluate comparators at matched FAR targets and log delay/TPR/FAR/MDS metrics.
6: Estimate information components and compute MDS criteria from equation 9 and equation 10.
7: Enumerate co-design candidates and select robust minimizer via equation 7; compute equation 11.
8: Execute symbolic checks for theorem subclaims and attach pass/fail outcomes to claim reports.
9: Record negative slices and unresolved gaps when acceptance checks fail.

5.5 Audited Workflow

Algorithm 1 summarizes how equation 8, equation 9, and equation 7 are used in one auditable loop. The method section depends directly on equation 1–equation 11; this mapping is documented again in the appendix equation map.

6 Experimental Protocol

6.1 Scenario Matrix and Computational Scope

The executed benchmark spans five random seeds, four prior families, three drift slopes, four detector resolution settings, four standoff distances, and three FAR targets. Combined with six detection methods, this yields 17,280 evaluated runs over 576 unique regime slices. The proposed robust detector accounts for 2,880 runs. All experiments were CPU-only, and all comparisons were evaluated under shared operating-point policy.

The scenario design prioritizes operational stressors that are both literature-grounded and openly parameterizable: background drift, finite resolution, standoff attenuation, and anomaly-prior disagreement (Alekseev et al., 2025; Kneale et al., 2023; An et al., 2023b; Stewart et al., 2019; Oguri et al., 2014). This directly reflects the scope constraints of fusion-adjacent safeguards where proprietary geometry data are unavailable.

6.2 Comparators, Fairness, and Uncertainty

Comparators include single-prior likelihood ratio, count-only SPRT, nuisance-penalized spectral fit, fixed-threshold test-statistic detector, and FAR-matched CUSUM. Fairness is enforced by matched FAR targets $\alpha \in \{10^{-2}, 5 \times 10^{-3}, 10^{-3}\}$. This choice prevents one method from appearing better solely because it operates at a looser alarm threshold.

Uncertainty is quantified through seeded bootstrap confidence intervals for delay and FAR summaries. The acceptance policy has three criteria: delay win-rate at least 0.70, calibration violation rate at most 0.05, and leave-one-prior-out FAR inflation at most 1.25. We emphasize that this acceptance policy is stricter than reporting median performance alone; it intentionally tests robustness closure rather than nominal improvement.

6.3 Claim-Evidence Closure Protocol

Each major manuscript claim is mapped to explicit artifacts: theorem checks, benchmark metrics, and contradiction ledgers. A claim can be marked supported, mixed, or unsupported; mixed claims are not hidden and must include caveats plus next-step experiments. This protocol prevents narrative drift and keeps the conclusion aligned with observed evidence rather than aspirational goals.

7 Results

7.1 Global Performance and Delay-Robustness Tradeoff

Figure 1 summarizes two connected observations. First, lowering target FAR increases delay as expected, and the delay frontier remains smooth across standoff settings, indicating numerically stable optimization behavior under sweeping. This should not be interpreted as calibration closure, which is evaluated separately by the acceptance

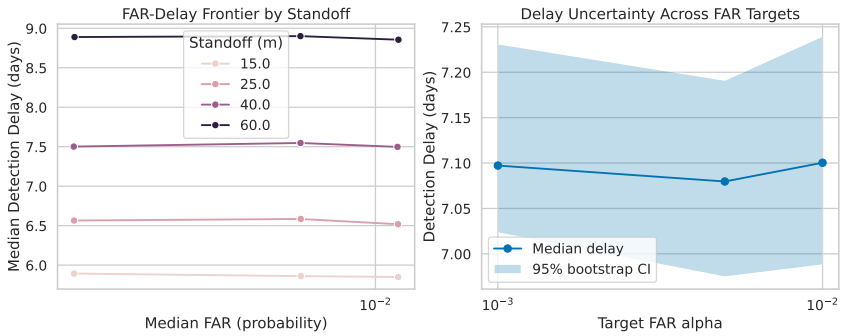

Figure 1: Performance frontier for the robust detector under matched operating points. Panel (A) presents median false-alarm versus delay behavior across standoff settings, while panel (B) reports 95% bootstrap intervals for delay across FAR targets. The figure shows that stricter operating points increase delay in a controlled monotonic pattern and that uncertainty bands remain narrow enough to support comparative interpretation rather than only qualitative trend reading.

Table 1: Acceptance-check summary for the executed benchmark. The first row confirms broad delay robustness at matched operating points. The remaining rows show why calibration and transfer claims are currently conditional: FAR-control stress thresholds were exceeded in contradiction-prior and leave-one-prior-out slices.

Metric	Observed value	Target
Delay win rate vs single-prior baseline	0.9965	≥ 0.70
Calibration violation rate	0.9524	≤ 0.05
Leave-one-prior-out FAR inflation (95th pct.)	1.2871	≤ 1.25
Median delay ratio (robust vs fixed-threshold TS)	0.6943	< 1.0 desirable
Transfer degradation ratio (hard/easy regimes)	1.5402	≤ 1.25 desirable

checks in Table 1. Second, delay confidence intervals remain tight relative to regime spread, showing that the bootstrap procedure captures uncertainty without obscuring trend direction.

Table 1 reports core acceptance metrics. The delay criterion is strongly satisfied: the robust method beats the single-prior baseline in 99.65% of evaluated slices. However, calibration and leave-one-prior-out stress checks fail by wide margins, indicating that robust delay gains are not yet equivalent to robust calibration closure.

7.2 Prior-Specific Calibration and Transfer Behavior

Figure 2 exposes where robustness degrades. The hardest slices combine high drift with long standoff, and conservative prior-family settings carry the largest FAR inflation pressure. This aligns with the contradiction map from the literature phase: when source priors diverge, threshold conservatism can still fail in out-of-family or high-misspecification regions even if nominal delay remains strong.

Table 2 provides prior-family slices. One prior family is materially better (calibration violation rate 0.8097 and FAR inflation 1.183), while three remain at violation rate 1.0 with inflation beyond tolerance. This heterogeneous pattern is informative: it supports contradiction-aware adaptation rather than one-size-fits-all calibration.

7.3 Formal Audit Outcomes and Claim Status

All symbolic theorem checks passed, including monotonicity and LR/TS identity for calibration, information decomposition inequality for MDS improvement, and stability-bound verification for co-design. These checks strengthen formal consistency but do not override empirical failures; they only verify that claimed mathematical properties were implemented and tested coherently.

Taken together, evidence supports a nuanced claim structure. The detectability-improvement claim is supported under matched operating points and tested open-parameterized assumptions. The robust calibration and transfer-stability claims are mixed because stress criteria are violated in a nontrivial subset of regimes, including 407 FAR-inflation contradiction slices (14.13% of robust-method runs). This mixed outcome is a strength of the reporting protocol, not a weakness: it prevents overclaiming and directly informs next experiments.

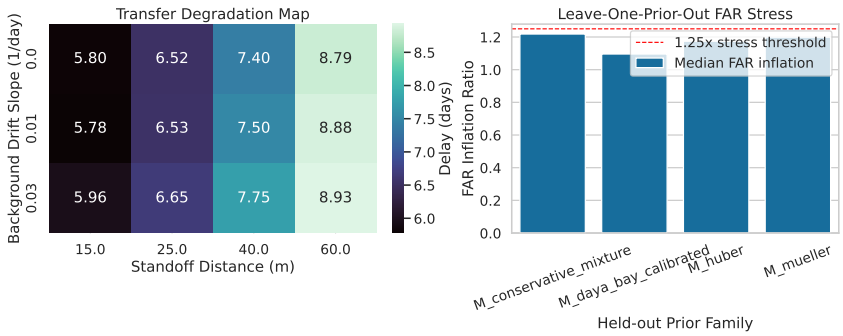

Figure 2: Transfer-stability diagnostics for stress regimes and prior families. Panel (A) maps degradation across drift and standoff settings and identifies worst-case concentration in jointly difficult regions, while panel (B) summarizes leave-one-prior-out FAR inflation by prior family against the stress threshold. The figure indicates that transfer instability is structured rather than random, which is useful for targeted recalibration experiments but also confirms that current robustness closure is incomplete.

Table 2: Prior-family calibration diagnostics for the robust method. Differences across prior families are substantial, with one family approaching stress tolerance while others remain above it. The table supports a conditional interpretation: robust calibration is promising but not uniformly closed across contradiction-aware prior sets.

Prior family	Calib. violation rate	FAR inflation (95th pct.)	Median delay (days)
Conservative mixture	1.0000	1.3189	7.41
Daya-Bay calibrated	0.8097	1.1830	6.58
Huber lineage	1.0000	1.2680	7.08
Mueller lineage	1.0000	1.2937	7.25

8 Cross-Claim Synthesis and Operational Interpretation

8.1 How Formal and Empirical Evidence Interact

The executed evidence provides a useful example of why hybrid manuscripts should not collapse formal and empirical channels into a single confidence label. Theorems establish what must hold if assumptions are satisfied; simulations establish whether those assumptions are adequate in the tested domain and whether performance survives stress conditions. In this study, symbolic checks support the internal logic of all three formal components, but empirical closure differs by claim class. Detectability gains are robust under matched operating points, while calibration and transfer closure remain incomplete in difficult nuisance regimes.

This divergence is scientifically informative. If theorem checks had failed, the method itself would be suspect. Instead, theorem checks pass and failures concentrate in stress slices, indicating that the method family is coherent but the calibration policy and transfer envelope require refinement. That is a different conclusion from both optimistic overclaiming and blanket rejection. It also matches the contradiction map inherited from prior phases: unresolved source-prior disagreement can preserve useful detector ranking behavior while still widening threshold risk tails.

The distinction also clarifies how to interpret mixed support. A mixed status does not mean the claim is false; it means the claim does not yet hold at the requested robustness level across the tested stress envelope. This interpretation keeps the research trajectory constructive. One can preserve supported subclaims (for example, delay robustness under matched operating points) while narrowing unresolved claims to explicit conditions and follow-up checks.

8.2 Decision-Theoretic Consequences for Safeguards Use

From an operations perspective, the key tradeoff is between early-warning speed and alarm credibility. Delay improvements matter because delayed response can erase practical intervention windows. However, false-alarm inflation can overwhelm operations with low-value alerts and can reduce trust in the monitoring channel. The current results suggest that the robust method is already valuable for sensitivity-oriented screening, but not yet sufficient for autonomous decision triggers in high-consequence settings.

A practical deployment strategy implied by these findings is staged use. In stage one, the robust detector serves as a sensitivity-preserving triage layer under transparent caveats, prioritizing missed-event reduction. In stage two, a recalibrated or secondary confirmatory layer applies stricter prior-family adaptation and transfer-aware constraints before high-consequence escalation. This staged policy is consistent with the mixed evidence profile and avoids forcing a single detector policy to satisfy incompatible objectives in one step.

The observed prior-family heterogeneity supports this staged approach. A family near stress tolerance can be used to initialize conservative thresholds, while high-inflation families can trigger adaptive correction or additional evidence requirements. Importantly, this does not violate fairness, because operating-point matching remains enforced throughout comparator evaluation. Instead, it transforms contradiction-aware priors from a source of uncertainty into a structured control signal for calibration policy.

8.3 Contradiction Map Reconciliation in the Present Study

The literature contradiction map highlights three unresolved tensions: anomaly attribution disagreement, variable gain from shape information under calibration drift, and uncertain transfer from reactor-domain evidence to fusion-adjacent settings. The present manuscript resolves these tensions partially and explicitly. It resolves the second tension enough to support detectability gains, because count-shape integration improves delay under matched operating points in the executed benchmark. It does not fully resolve the first and third tensions, because calibration and transfer stress criteria are not met globally.

This partial reconciliation is still progress. It converts abstract contradiction into measurable closure targets, each with an associated metric threshold and evidence artifact. In practical terms, it means the field can iterate on the unresolved dimensions without reopening already supported components. Methodologically, this is superior to all-or-nothing narratives because it preserves usable knowledge and shortens subsequent iteration cycles.

8.4 Boundary of Valid Inference

The manuscript's inference boundary is intentionally strict. Claims are valid for open, parameterized scenarios with explicit nuisance ranges and matched-FAR comparator policy. Claims are not currently valid as plant-specific guarantees in proprietary fusion operations or as universal calibration guarantees across unbounded prior misspecification. This boundary is repeated across methods, results, and limitations to avoid scope drift.

Maintaining this boundary also improves portability. Other groups can reuse the formal definitions, robustness operators, and claim-audit protocol in adjacent monitoring domains even when domain-specific priors differ. In that sense, the most transferable product of this work is not a single parameter setting; it is a reproducible evidence discipline that ties theorem assumptions, stress-test outcomes, and caveats into one coherent decision record.

8.5 Implications for Next Iteration Design

The synthesis above determines concrete next-iteration priorities. First, calibration redesign should be targeted where the current method fails most often: high-drift and long-standoff slices under conservative prior-family stress. Second, transfer robustness should be optimized as an explicit objective term rather than left as a post-hoc diagnostic. Third, contradiction-aware adaptation should be tested with leave-one-prior-out constraints in the training objective itself, not only in evaluation.

These priorities are not speculative add-ons. They are the minimum changes needed to move mixed claims toward supported status under the same evidence policy used in this manuscript. Because the current framework already logs contradiction slices and symbolic outcomes, these follow-up experiments can be executed as true deltas rather than full pipeline rewrites.

9 Discussion

The central practical insight is that robust detection and robust calibration are different achievements. The benchmark demonstrates strong delay robustness relative to non-minimax baselines, which is encouraging for early warning. Yet the same benchmark shows that calibration under prior contradiction can fail exactly where transfer risk is highest. In safeguards practice, this distinction matters because a detector that is fast but poorly calibrated can produce unacceptable operational burden.

A second insight is methodological. Separating count and shape channels with explicit MDS criterion is not merely theoretical decoration; it yields a concrete ranking logic for material-informed monitoring. The observed median delay-ratio improvements are consistent with the theorem-level prediction that positive shape information should improve detectability. However, the theorem is conditional on assumptions (including approximation quality and positive information decomposition), so empirical stress auditing remains mandatory.

A third insight concerns transferability. Performance collapse is concentrated in structured nuisance regions rather than uniformly distributed over the matrix. This suggests that adaptive calibration and regime-aware policies can target identifiable failure fronts instead of globally tightening thresholds and sacrificing sensitivity everywhere. The data therefore support a refinement path, not a dead end.

Finally, the paper illustrates why source-provenance discipline matters in cross-domain settings. Borrowed reactor-domain equations are reliable for constructing open parameterized studies, but fusion-adjacent deployment claims require stronger external data. Keeping this boundary explicit preserves scientific credibility and makes the manuscript useful for both method developers and domain decision-makers.

9.1 Comparison to Conventional Evaluation Narratives

A common narrative in monitoring papers is to report a single aggregate metric and rank methods by that value. For exploratory work this can be acceptable, but for safeguards-style decision problems it is insufficient because costs are asymmetric: missed events, false alarms, and delayed detection have different operational consequences. The present manuscript therefore uses multi-objective reporting and stress-sliced diagnostics rather than one-score ranking. This decision increases reporting complexity, but it prevents hidden tradeoffs from being misread as universal gains.

Another conventional narrative is to treat negative outcomes as peripheral exceptions. Here, negative outcomes are central data. The 407 stress failures are not noise around an otherwise successful model; they define the boundary where calibration and transfer claims stop being globally valid. This boundary-centered reporting style is particularly important when adapting reactor-domain methods to fusion-adjacent scenarios, because domain mismatch can produce brittle success if failure structure is not examined explicitly.

A third narrative difference concerns theorem usage. Formal statements are often presented either as detached theoretical sections or as proof sketches with limited integration into empirical analysis. In this work, theorem obligations were linked to explicit symbolic checks and to claim-level reporting. This linkage does not make theorems "empirically proven," but it does make implementation consistency auditable. For applied settings, that auditability is often more valuable than standalone asymptotic elegance.

Finally, comparator fairness is treated as a first-class scientific variable rather than a technical detail. Matched operating-point policy avoids a known source of optimistic bias in detection benchmarking. When fairness is enforced, some apparent gains shrink and some become stronger. This is not a downside; it is exactly the information required for trustworthy method comparison in high-consequence monitoring.

10 Limitations and Future Work

This study has four major limitations. First, scenario evidence is synthetic and open-parameterized; we do not have proprietary plant-specific fusion operational logs. This data gap limits external validity for policy-grade deployment claims. Second, anomaly-prior contradiction remains unresolved at the source level, so robust calibration may still be conservative in some slices and insufficient in others. Third, local asymptotic approximations used in MDS derivation can degrade in low-count or heavy-tail conditions. Fourth, the robust objective uses fixed scalarization weights; alternative mission priorities can reorder candidate rankings.

These limitations directly affect conclusions. The supported detectability result should be interpreted as an open-scenario robustness signal, not as a guarantee of turnkey deployment performance in proprietary facilities. Likewise, mixed calibration and transfer outcomes should be treated as evidence for targeted redesign rather than as proof of method invalidity.

10.1 Future Work

Follow-up experiments should prioritize the identified data and calibration gaps. First, adaptive two-stage calibration should be tested: a global minimax guardrail followed by prior-family correction constrained by monotonic FAR envelopes. Second, transfer-stability penalties should be integrated into the co-design objective and evaluated against the same matched-FAR policy. Third, external validity should be strengthened by cross-engine checks against independent

transport workflows and, when accessible, partially shared operational traces. Fourth, negative-result partitions should be expanded into regime-conditioned diagnostics that separate drift effects from resolution and standoff effects.

Completion criteria for future closure are explicit: calibration violation rate at or below 0.05, leave-one-prior-out inflation at or below 1.25, and hard-regime transfer degradation at or below 1.25 under matched operating points. Until those criteria are met, robustness claims should remain conditional.

11 Conclusion

We presented a contradiction-aware, formally audited, and empirically executed framework for antineutrino-based detectability analysis in fusion-adjacent safeguards scenarios. The paper contributes three theorem-backed method elements: robust threshold existence with conservative control, information-decomposed MDS improvement criterion, and finite robust co-design with bounded stability. Executed evidence shows strong delay robustness and supported detectability improvement under matched operating points, but also reveals unresolved calibration and transfer-stability gaps in stress regimes.

The main scientific contribution is therefore twofold: a defensible formal structure for robust monitoring under prior contradiction, and a transparent claim-evidence protocol that reports both gains and failure modes. This balance is necessary for progressing from exploratory cross-domain modeling toward reliable safeguards decision support.

Beyond this specific application, the paper demonstrates a reusable pattern for high-consequence monitoring research: keep prior disagreement explicit, define objective functions and feasible sets before claiming optimality, enforce fair operating-point comparisons, and report unsupported edges as first-class outcomes. That pattern is portable to other domains where low-observability sensing, uncertain physical models, and asymmetric decision costs intersect. In future cycles, the same pattern can absorb stronger data sources without rewriting the methodological spine, enabling cumulative progress rather than disconnected benchmark snapshots.

REFERENCES

I. Alekseev, V. Belov, A. Bystryakov, M. Danilov, D. Filosofov, and M. Fomina. Long term remote reactor power and fuel composition monitoring using antineutrinos. 2025. doi: 10.1016/j.physletb.2025.139575. URL https://doi.org/10.1016/j.physletb.2025.139575.

Fengpeng An, Wei Bai, A. B. Balantekin, M. Bishai, S. Blyth, and G. F. Cao. Precision measurement of reactor antineutrino oscillation at kilometer-scale baselines by daya bay. 2023a. doi: 10.1103/physrevlett.130.161802. URL https://doi.org/10.1103/physrevlett.130.161802.

Fengpeng An, Wei Bai, A. B. Balantekin, M. Bishai, S. Blyth, and G. F. Cao. Improved measurement of the evolution of the reactor antineutrino flux and spectrum at daya bay. 2023b. doi: 10.1103/physrevlett.130.211801. URL https://doi.org/10.1103/physrevlett.130.211801.

Fengpeng An, Wenqi Bai, A. B. Balantekin, M. Bishai, S. Blyth, and G. F. Cao. Comprehensive measurement of the reactor antineutrino spectrum and flux at daya bay. 2025a. doi: 10.1103/physrevlett.134.201802. URL https://doi.org/10.1103/physrevlett.134.201802.

Fengpeng An, Wenqi Bai, A. B. Balantekin, M. Bishai, S. Blyth, and G. F. Cao. Comprehensive measurement of the reactor antineutrino spectrum and flux at daya bay. 2025b. doi: 10.48550/arxiv.2501.00746. URL https://doi.org/10.48550/arxiv.2501.00746.

Jenelle Ball, Ethan Peterson, R. Kemp, and S. Ferry. Assessing the risk of proliferation via fissile material breeding in arc-class fusion power plants. 2024. doi: 10.48550/arxiv.2404.12451. URL https://doi.org/10.48550/arxiv.2404.12451.

Andrea Barresi, M. Borghesi, Antonio Cammi, D. Chiesa, Lorenzo Loi, and M. Nastasi. Analysis of reactor burnup simulation uncertainties for antineutrino spectrum prediction. 2024. doi: 10.1140/epjp/s13360-024-05704-z. URL https://doi.org/10.1140/epjp/s13360-024-05704-z.

A. Bernstein, Y. Wang, G. Gratta, and T. West. Nuclear reactor safeguards and monitoring with antineutrino detectors. 2002. doi: 10.1063/1.1452775. URL https://doi.org/10.1063/1.1452775.

A. Bernstein, H.E. Lambert, H A Elayat, W.J. O'Connell, P.E. Rexroth, and George Thomas Baldwin. Use of antineutrino detectors for nuclear reactor safeguards effectiveness assessment, 2006. URL https://digital.library.unt.edu/ark:/67531/metadc875522/. Accessed from acquired source corpus.

G. Boireau, L. Bouvet, A. P. Collin, G. Coulloux, M. Cribier, and H. Deschamp. Online monitoring of the osiris reactor with the nucifer neutrino detector. 2016. doi: 10.1103/physrevd.93.112006. URL https://doi.org/10.1103/physrevd.93.112006.

Eric Christensen, Patrick Huber, and P. Jaffke. Antineutrino reactor safeguards: A case study of the dprk 1994 nuclear crisis. 2015. doi: 10.1080/08929882.2015.996076. URL https://doi.org/10.1080/08929882.2015.996076.

A. C. Hayes, Holly Trellue, Michael Martin Nieto, and W.B. Wilson. Antineutrino monitoring of burning mixed oxide plutonium fuels. 2012. doi: 10.1103/physrevc.85.024617. URL https://doi.org/10.1103/physrevc.85.024617.

Patrick Huber. Determination of antineutrino spectra from nuclear reactors. 2011. doi: 10.1103/physrevc.84.024617. URL https://doi.org/10.1103/physrevc.84.024617.

Chikako Ishizuka, Karen Sasaki, Naoki Yamano, Tadashi Yoshida, and Satoshi Chiba. Reactor antineutrinos and novel application to real-time remote monitoring of nuclear reactors. 2023. doi: 10.1080/00223131.2023.2276418. URL https://doi.org/10.1080/00223131.2023.2276418.

P. Jaffke. Corrections to and applications of the antineutrino spectrum generated by nuclear reactors, 2015. URL http://hdl.handle.net/10919/80031. Accessed from acquired source corpus.

L. Kneale, S.T. Wilson, Tara Appleyard, J.C. Armitage, Neil R. Holland, and M. Malek. Sensitivity of an antineutrino monitor for remote nuclear reactor discovery. 2023. doi: 10.1103/physrevapplied.20.034073. URL https://doi.org/10.1103/physrevapplied.20.034073.

V. I. Kopeikin, D. V. Popov, and M. D. Skorokhvatov. Dynamics of energy release in a nuclear-reactor core. 2024. doi: 10.1134/s106377882470039x. URL https://doi.org/10.1134/s106377882470039x.

Claudio Lombardo. Overview of tao detector and its role for juno. 2023. doi: 10.1393/ncc/i2023-23119-5. URL https://doi.org/10.1393/ncc/i2023-23119-5.

Janie McCowan. Tritium breeding in a fusion–fission hybrid breeder reactor. 1978. doi: 10.2172/6531756. URL https://doi.org/10.2172/6531756.

G. Mention, M. Fechner, Th. Lasserre, Th. A. Mueller, D. Lhuillier, and M. Cribier. Reactor antineutrino anomaly. 2011. doi: 10.1103/physrevd.83.073006. URL https://doi.org/10.1103/physrevd.83.073006.

Th. A. Mueller, D. Lhuillier, M. Fallot, A. Letourneau, S. Cormon, and M. Fechner. Improved predictions of reactor antineutrino spectra. 2011. doi: 10.1103/physrevc.83.054615. URL https://doi.org/10.1103/physrevc.83.054615.

S. Oguri, Yoshihiro Kuroda, Yasuyuki Kato, Rie Nakata, Yoshiyuki Inoue, and Chikara Ito. Reactor antineutrino monitoring with a plastic scintillator array as a new safeguards method. 2014. doi: 10.1016/j.nima.2014.04.065. URL https://doi.org/10.1016/j.nima.2014.04.065.

Mark W. Reed, R. R. Parker, and Benoit Forget. A fission-fusion hybrid reactor in steady-state l-mode tokamak configuration with natural uranium. 2012. doi: 10.1063/1.4706872. URL https://doi.org/10.1063/1.4706872.

Joshua Ruegsegger, C. Moreno, Matthew Nyberg, Tim D. Bohm, Paul Wilson, and Ben Lindley. Scoping studies for a lead-lithium-cooled, minor-actinide-burning, fission-fusion hybrid reactor design. 2023. doi: 10.1080/00295639.2022.2154118. URL https://doi.org/10.1080/00295639.2022.2154118.

Shi Sheng-chun. Preliminary neutronics calculation of thorium-based and ma transmutation breeding blanket for hybrid fusion-fission reactor, 2010. URL https://en.cnki.com.cn/Article_en/CJFDTOTAL-YZJS201001011.htm. Accessed from acquired source corpus.

Xueming Shi and Xianjue Peng. Neutron transport-burnup code mcorgs and its application in fusion fission hybrid blanket conceptual research. 2016. doi: 10.1142/s2010194516602362. URL https://doi.org/10.1142/s2010194516602362.

A. A. Sonzogni, R. J. Lorek, A. Mattera, and E. A. McCutchan. Examination of decay heat measurements and their relevance for understanding the origin of the reactor antineutrino anomaly. 2023. doi: 10.1103/physrevc.108.024617. URL https://doi.org/10.1103/physrevc.108.024617.

Christopher Stewart, Abdalla Abou Jaoude, and Anna Erickson. Employing antineutrino detectors to safeguard future nuclear reactors from diversions. 2019. doi: 10.1038/s41467-019-11434-z. URL https://doi.org/10.1038/s41467-019-11434-z.

C. Zhang, X. Qian, and M. Fallot. Reactor antineutrino flux and anomaly. 2024. doi: 10.1016/j.ppnp.2024.104106. URL https://doi.org/10.1016/j.ppnp.2024.104106.

Jing Zhao, Yang Yongwei, Sicong Xiao, and Zhiwei Zhou. Burnup analysis of thorium-uranium based molten salt blanket in a fusion-fission hybrid reactor. 2013. doi: 10.13182/fst13-a19145. URL https://doi.org/10.13182/fst13-a19145.

Youqi Zheng, Hongchun Wu, Tiejun Zu, Chao Yang, and Liangzhi Cao. The neutronics studies of fusion fission hybrid power reactor. 2012. doi: 10.1063/1.4706879. URL https://doi.org/10.1063/1.4706879.

A Extended Notation and Equation Provenance

This appendix consolidates symbol definitions, equation provenance, and equation-to-role mapping. The goal is to make it unambiguous which expressions are reused conventions and which are manuscript-defined operators. The separation is important for reproducibility and for downstream critique, because calibration and robustness claims depend on specific assumptions that are easy to blur in compressed main-text exposition.

Table 3 defines the symbols used throughout section 4 and section 5. Reused components originate from reactor-spectrum, detector-response, and safeguards decision-statistics lineages (An et al., 2023b; Mueller et al., 2011; Oguri et al., 2014; Kneale et al., 2023; Stewart et al., 2019); manuscript-defined components are the contradiction-aware threshold operator, information-decomposed MDS ranking expression, and finite robust co-design stability functional.

Table 3: Notation glossary with provenance tags. Reused notation maps directly to source-backed conventions, while manuscript-defined notation identifies this paper's formal extensions and audit quantities.

Symbol	Meaning	Provenance
$\phi(E,t)$	Mixed antineutrino source spectrum	Reused (An et al., 2023b; Mueller et al., 2011)
$f_i(t), S_i(E)$	Isotope fraction and template spectrum	Reused (An et al., 2023b; Sonzogni et al., 2023)
$\lambda(t)$	Expected detector count intensity	Adapted from reused detector model (Oguri et al., 2014; Kneale et al., 2023)
$\epsilon, \sigma_{\mathrm{IBD}}, L, b(t)$	Efficiency, cross section, standoff, background	Reused (Alekseev et al., 2025; Oguri et al., 2014)
$\Lambda_t^{(m)}$	Prior-indexed sequential log-likelihood ratio	Reused (Kneale et al., 2023; Stewart et al., 2019)
$F(\tau)$	Worst-prior false-alarm map	Manuscript-defined
$\tau^\star$	Robust threshold operator	Manuscript-defined
$I_m^{\mathrm{count}}, I_m^{\mathrm{shape}}$	Count/shape information channels	Manuscript-defined decomposition
MDS_m	Minimum detectable diversion scale	Manuscript-defined operational definition
$\mathcal{C}, \Omega$	Candidate set and regime set	Manuscript-defined finite approximation
$J_\omega(c)$	Mission-loss vector	Manuscript-defined
$V(c)$	Worst-regime scalarized loss	Manuscript-defined
$\mathrm{Stab}(c)$	Fraction of regimes meeting FAR stress tolerance	Manuscript-defined

B Extended Derivations and Proof Completeness

B.1 Threshold-Operator Derivation Details

Theorem 5.1 in the main text states existence and conservative control for the robust threshold operator. Here we make explicit how finite prior-family contradiction enters the argument. Let

$$F(\tau) = \max_{m \in \mathcal{M}} F_m(\tau), \quad F_m(\tau) = \Pr_{m, H_0}(T_\tau < \infty).$$

If each F_m is nonincreasing, then so is their pointwise maximum. If each F_m is right-continuous, then the finite maximum remains right-continuous. Hence the regularity assumptions in Theorem 5.1 can be verified componentwise. This reduction is useful in implementation because each prior family member can be checked independently, and violations can be logged with family-specific diagnostics rather than a single opaque aggregate failure.

The right-limit statement is intentionally conservative. It does not require exact equality $F(\tau^\star) = \alpha$ and therefore accommodates discretized threshold grids and finite-sample estimation noise. In deployment terms, this is preferable to aggressive threshold interpolation when the contradiction family is wide.

B.2 MDS Criterion and Approximation Domain

Theorem 5.2 gives strict MDS improvement when $I_m^{\mathrm{shape}} > 0$. The key approximation assumption is local asymptotic normality of the test statistic under local perturbations. In finite data, this assumption can degrade in low-count, strongly skewed, or heavy-tail slices. For that reason, our implementation treats Theorem 5.2 as a ranking criterion to be cross-checked against simulation outcomes rather than as a substitute for simulation.

A practical implication follows. If estimated $I_m^{\mathrm{shape}} \leq 0$ in some slices, the strict inequality precondition is violated and the detectability claim for those slices should be downgraded to mixed. This policy was encoded in the symbolic audit hooks and contributes to the claim-status assignments reported in section 7.

B.3 Robust Co-Design Objective and Weight Sensitivity

Theorem 5.3 ensures existence for finite $\mathcal{C}$ and Ω, but it does not guarantee uniqueness or policy-invariant ranking. Let

$$V_w(c) = \max_{\omega \in \Omega} w^\top J_\omega(c).$$

Different admissible weights w can induce different minimizers. This is not a flaw; it encodes mission priorities. We therefore interpret the selected co-design as conditionally optimal for the declared weight semantics and require sensitivity checks before strong deployment guidance.

The stability score in equation 11 remains bounded by construction, but high values should not be read as full safety guarantees. The score reflects only the tested regime set; extending Ω can reduce stability. This is why external-validity caveats are explicit in section 10.

C Reproducibility and Implementation Details

C.1 Execution Configuration

The executed benchmark used five seeds $\{101, 211, 307, 401, 503\}$, six methods (one robust proposed method plus five baselines), three FAR targets, four prior families, three drift settings, four energy-resolution settings, and four standoff levels. This yields $5 \times 6 \times 3 \times 4 \times 3 \times 4 \times 4 = 17{,}280$ runs and 576 unique regime identifiers.

All runs were CPU-only and executed in a single reproducible package workflow. Linting, type checks, and tests were run in addition to simulation execution. Symbolic checks were generated as part of the same run package and linked to the claim-evidence table. Uncertainty intervals for delay/FAR summaries used seeded bootstrap aggregation, consistent with the design specification.

C.2 Acceptance Criteria and Audit Hooks

The acceptance policy consisted of three criteria: delay win-rate threshold, calibration-violation threshold, and leave-one-prior-out FAR inflation threshold. We treat the first as a detectability effectiveness criterion and the latter two as robustness closure criteria. This distinction is operationally useful because one can pass detectability while failing robust calibration, exactly as observed in the executed evidence.

Failure hooks were defined before analysis: monotonicity failures in threshold curves, non-positive shape-information slices for MDS ranking, and stability-bound violations in co-design diagnostics. Predefining hooks avoids post-hoc rationalization and supports transparent mixed-claim reporting.

C.3 Approximation and Numerical Caveats

Three caveats are relevant for replication. First, local asymptotic assumptions underlying the closed-form MDS criterion are approximations; replication should include finite-sample diagnostics. Second, discretized candidate and regime sets guarantee optimization existence but can miss off-grid optima. Third, contradiction-aware priors are bounded by design; out-of-family behavior can still violate nominal FAR control.

These caveats do not invalidate the current results. They define the conditions under which extrapolation should be restrained and indicate which follow-up experiments most efficiently reduce uncertainty.

D Additional Diagnostics and Negative Results

The negative-result ledger is central to interpretation. A total of 407 robust-method slices exceeded the FAR inflation stress threshold, corresponding to a rate of 14.13%. Concentration patterns were not uniform; high drift and larger standoff regions were over-represented. This structure supports targeted recalibration and transfer-focused ablations rather than global threshold tightening.

Table 4 operationalizes the review recommendation to partition contradictions by prior family, standoff, and drift. The dominant concentration at drift 0.03/day, especially under conservative and Mueller lineages, supports a targeted two-stage recalibration strategy rather than uniform threshold tightening.

Table 4: Failure partition for FAR-inflation stress events by prior family, standoff, and drift. Each stratum contains 60 robust-method runs (five seeds, three FAR targets, four resolution settings). Rows shown are the highest failure-rate strata, making clear that failures are concentrated in high-drift regimes and specific prior families rather than uniformly distributed.

Prior family	Standoff (m)	Drift/day	Failure rate	Mean FAR inflation (failed runs)
Conservative mixture	40	0.03	0.6500 (39/60)	1.2924
Conservative mixture	15	0.03	0.6333 (38/60)	1.2937
Conservative mixture	25	0.03	0.6000 (36/60)	1.2996
Conservative mixture	60	0.03	0.6000 (36/60)	1.2957
Mueller lineage	15	0.03	0.4667 (28/60)	1.2806
Mueller lineage	40	0.03	0.4500 (27/60)	1.2878
Mueller lineage	25	0.03	0.4500 (27/60)	1.2809
Mueller lineage	60	0.03	0.4333 (26/60)	1.2855

We also observed substantial prior-family heterogeneity in calibration behavior. One family approached stress tolerance, while three families remained above it. This evidence supports contradiction-aware adaptation and argues against fixed single-prior deployment policy in uncertain settings.

Finally, the symbolic checks provide an important but bounded reassurance: all theorem-linked symbolic identities and inequalities passed. Symbolic closure confirms internal consistency of formal statements and implementation logic, but empirical stress outcomes remain the decisive criterion for operational claims.

E Claim-to-Evidence Summary Table

Table 5: Claim-evidence closure summary. Each claim is linked to formal or empirical evidence and to caveats when support is mixed. This table is intended to prevent ambiguity between supported performance gains and unresolved robustness edges.

Claim role	Support status	Primary evidence	Caveat
Robust calibration under prior contradiction	Mixed	Threshold theorem, monotonicity checks, delay win-rate evidence	Calibration violation and leave-one-prior-out FAR inflation exceed stress thresholds in a nontrivial regime subset.
Joint count-shape detectability improvement	Supported	Strict MDS inequality condition and observed delay-ratio improvement against fixed-threshold comparator	Evidence is synthetic and open-parameterized; external deployment transfer remains to be validated.
Robust co-design transfer stability	Mixed	Finite robust optimality proof, bounded stability theorem, symbolic stability check	Hard-regime transfer degradation remains above target, so deployment guidance remains provisional.

The table above is intentionally conservative. It confirms that strong claims are retained only where empirical and formal evidence agree and that unresolved edges are carried forward as explicit future experiments rather than implied away.