WHEN SAFETY JUDGES AGREE TOO MUCH: CORRELATION BREAKS NAIVE FUSION AND CAPS ITS GAINS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deployed LLM systems increasingly run several safety guards and combine their verdicts, either by multiplying their evidence or by escalating from a cheap guard to expensive ones. We give a statistical account of how the guards' agreement, together with their individual strength, governs what such combinations can and cannot do. Across seven public guards from four families and three human-labelled benchmarks, guards are strongly correlated, and on WildGuardTest and BeaverTails all of them fail together on 12–13% of unsafe responses, against at most 0.03% under independence. We prove that this correlation silently breaks naive fusion: the product of per-guard likelihood ratios is no longer an e-value, correlation inflates its null mean and log-variance by closed-form factors, and its false-alarm rate tends to a positive floor as more equicorrelated guards are added, for any fixed nominal level; empirically, the common escalation cascade exceeds a nominal 5% false-alarm rate in 6 of 7 settings, reaching 26%. Correlation can also cap the gain of fusion, and within a Gaussian-copula model does so quantitatively: a single quantity, the information the remaining guards add to a cheap one, bounds both what any combination rule, including adaptive cascades of arbitrary cost, can gain and what calling only the cheap guard loses. On real scores we use the model for gap sizes and rankings, not as a certified ceiling. We complement this with a cascade construction whose false-alarm rate is controlled in finite samples for any calling rule fixed before calibration and any dependence, using about 100 target-domain safe examples exchangeable with future safe traffic. Across seven in-domain and shifted settings, a static subset chosen on source data, at about a tenth of the full ensemble's parameter memory, retains nearly all of its detection at a calibrated 5% false-alarm rate, though less at stricter levels. An adaptive dependency-aware cascade does not consistently improve on it, and, for these seven guards and under the same target calibration, the largest guard alone is never more than 0.001 below a calibrated fusion of all seven in detection, in point estimates.

1 INTRODUCTION

Safety guards, i.e. classifiers that flag harmful prompts or responses, are now a standard layer in LLM deployments, and practitioners rarely use just one. Open guard families come in several sizes (Inan et al., 2023; Zeng et al., 2024; Padhi et al., 2025; Han et al., 2024; Zhao et al., 2025), and a natural engineering response to their individual errors is to *combine* them: multiply their likelihood ratios (naive Bayes), average them, or run a cheap guard first and escalate to a larger one when it is unsure, as proposed for LLM judges (Jung et al., 2025). Three expectations usually come with this practice: that combining guards keeps the false-alarm rate that each guard was calibrated for, that more guards catch more harm, and that escalation saves compute without costing detection.

We argue that all three expectations are governed by a single, usually ignored property: *how much the guards agree*. Guards are trained on overlapping data with overlapping policies, often from the same base model, so their scores are strongly dependent. On seven public guards from four families, evaluated on three human-labelled benchmarks, Qwen3Guard sizes correlate at 0.83–0.96

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

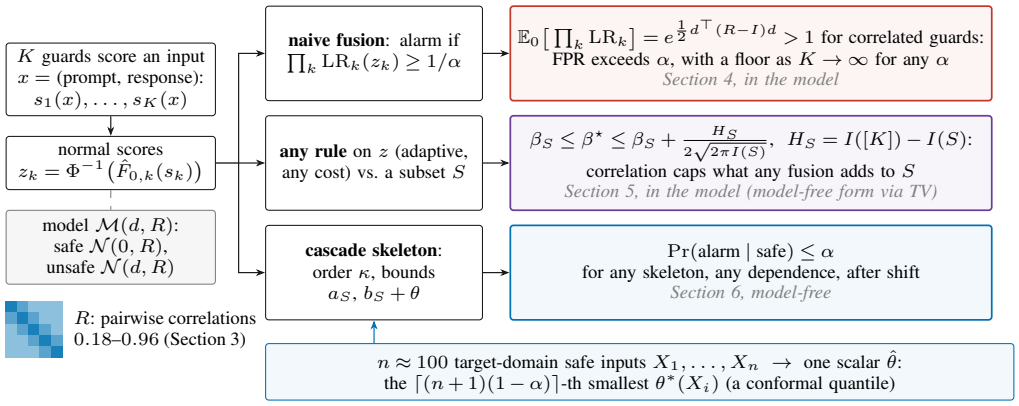


Figure 1: Setting and results. Guard scores are mapped to strongly correlated normal scores z. The Gaussian-copula model $\mathcal{M}(d, R)$ (grey) is used only to explain the failure of naive fusion and the cap H_S on any rule; validity does not use it: one scalar $\hat{\theta}$ calibrated on target-domain safe inputs controls the false-alarm rate of any cascade skeleton.

and cross-family pairs at 0.18–0.75; on WildGuardTest and BeaverTails, *every* guard misses on 12–13% of unsafe responses, against at most 0.03% if they erred independently (Section 3).

This dependence has three consequences (Figure 1).

Correlation breaks naive fusion (Section 4). Independence on safe inputs is what makes the product of per-guard likelihood ratios an e-value. Under a Gaussian-copula model we give the exact null law of its logarithm, whose mean and variance depart from the independent case by signal-weighted safe-class correlations, and show that for equicorrelated guards its false-alarm rate tends to a positive floor as guards are added, *for any fixed nominal level*. The practitioner's cascade fails for the same reason: at every step it counts again the part of the new guard's score that earlier guards already predicted. Empirically, it exceeds a nominal 5% false-alarm rate in 6 of 7 settings, reaching 26%.

Correlation caps the gain of fusion (Section 5). Reading classical Gaussian Neyman–Pearson detection as a statement about fusion, we show that one quantity, H_S, the information the other guards add to a subset S, bounds *from both sides* what is at stake: in the model, no combination rule, adaptive or not and at any cost, can detect more than $\beta_S + O(H_S)$, and calling only S loses at most that much; a model-free version ties the gap to how much the remaining guards still separate safe from unsafe *after conditioning on* S. The model's absolute ceiling is not reliable on real scores (on BeaverTails three single guards exceed it), so we use it only for the size of gaps and the ranking of subsets. On target traffic and for these seven guards, a calibrated LDA fusion of all seven is never more than 0.001 ahead of the largest guard, Qwen3Guard-8B, alone (Section 6.3).

Cheap recalibrated subsets can be competitive (Section 6). Calibrating one scalar on a small set of target-domain safe examples controls the false-alarm rate of any cascade in finite samples, for any calling rule fixed before calibration, any dependence and any misspecification, provided the calibration examples are exchangeable with future safe traffic; this restores validity after a shift, where even conformal fusion fitted on the source fails. Over seven settings and 20 seeds, a validation-selected static subset costing 8–12% of the full ensemble retains 96–100% of its detection at 100 calibration examples and $\alpha = 0.05$ (71–105% over all 32 configurations). Our dependency-aware adaptive cascade is cheaper than a dependency-blind one (29 of 32 configurations after Bonferroni correction) but, consistent with the ceiling, does not consistently beat the best static subset. Negative results are reported as such.

2 Setting and Data

Guards. We use $K = 7$ open guards from four families: Qwen3Guard 0.6B/4B/8B (Zhao et al., 2025), Llama-Guard-3 1B/8B (family: Inan et al., 2023), ShieldGemma-2B (Zeng et al., 2024) and

Granite-Guardian-3.1-2B (family: Padhi et al., 2025); WildGuard-7B (Han et al., 2024) is added as an eighth guard in a robustness check (Appendix). Each guard is run by forward inference only, following its model card template, and returns an unsafe score $s_k(x)$ for a (prompt, response) pair.

Data. Three human-labelled response-harm benchmarks: JBB (300 responses from JailbreakBench judge comparisons, 110 unsafe; Chao et al., 2024), WildGuardTest (1709, 284 unsafe; Han et al., 2024) and the BeaverTails-30k test split (3021, 1733 unsafe; Ji et al., 2023). Labels are the benchmarks' human labels.

Normal scores and cost. For each guard we map scores to $z_k = \Phi^{-1}\big((u_k + \frac{1}{2})/(n_{\mathrm{ref}} + 1)\big)$, where u_k is the number of the n_{ref} reference safe scores below s_k plus half the number tied with it; z_k is thus approximately standard normal on safe inputs and always finite. All combination rules act on z. The reference set is the safe part of the fitting data. Model parameters are plug-in estimates on the same fitting data: d is the difference of class means of z and R the correlation matrix of the class-centred scores, plus a $10^{-3}I$ ridge. Raw scores are heavily tied (log-odds from bf16 logits, with only 67–215 distinct values on the safe inputs of a benchmark), so the Gaussian-copula model of Sections 4–5 is applied to a discretised approximation; ties in calibration are broken at random, and the validity results of Section 6 do not depend on the model (Appendix B.1). We measure the cost of calling a guard by its parameter bytes in bf16 (1.5 GB for Qwen3Guard-0.6B, 56.06 GB for all seven); re-pricing with a non-embedding-parameter FLOPs proxy moves each method's cost by at most 2.0 points of the full cost and changes no comparison in Section 6.3 (Appendix C.4).

Error rates. We treat *safe* as the null hypothesis. The false-alarm rate (FPR) is the fraction of safe inputs flagged, and detection (det) the fraction of unsafe inputs flagged; the nominal level is $\alpha = 0.05$ unless stated otherwise.

3 Measuring dependence between safety judges

The scores of the seven guards are strongly dependent. Within each class, Qwen3Guard sizes correlate at 0.83–0.96, the two Llama-Guard-3 models at 0.60–0.80, and cross-family pairs at 0.18–0.75 (Figure 2a): a larger panel mostly repeats a shared signal.

The dependence is especially visible in failures. At the native operating point (each guard's own unsafe verdict, i.e. score above its default threshold), all judges miss the same unsafe item on 12–13% of unsafe examples on WGT and BT, versus at most 0.03% under independence; on JBB none of the 110 unsafe responses is missed by all seven (95% upper bound 3.3%), although most are missed by at least one (Figure 2b). Shared misses concentrate on slices where the benchmark's notion of harm is contestable, e.g. BeaverTails politics, all wrong on 63% (Table 13).

Consequently, fusion adds little discrimination. Over 20 random halvings (fit on one half, evaluate on the other, best single judge also chosen on the fitting half), the held-out AUROC gain over the best single judge is at most 0.0013 for stacking (logistic or gradient-boosted) and 0.004 for averaging standardised scores; in detection at a 5% false-alarm rate, no fitted rule gains more than 0.03 (Section 5). Adding WildGuard-7B as an eighth guard leaves this unchanged (gain at most 0.002; appendix).

Working model. Let $s_k(x)$ be the unsafe score of guard $k \in [K]$ and $z_k = \Phi^{-1}(F_{0,k}(s_k))$ its *normal score*, where $F_{0,k}$ is the distribution of s_k on safe inputs; thus $z_k \sim \mathcal{N}(0,1)$ under H_0 (safe) by construction. All results in this section are stated in the equal-covariance Gaussian-copula model

$$\mathcal{M}(d,R): \qquad H_0:\ z \sim \mathcal{N}(0,R), \qquad H_1:\ z \sim \mathcal{N}(d,R), \tag{1}$$

where $R \succ 0$ is a correlation matrix (so marginal variances are one under both hypotheses) and $d \in \mathbb{R}^K$ collects the per-guard separations (d_k^2 is the single-guard signal-to-noise ratio). For $S \subseteq [K]$ we write

$$I(S) = d_S^\top R_{SS}^{-1} d_S, \qquad I(\emptyset) = 0, \tag{2}$$

the squared Mahalanobis separation carried by the guards in S. The model is a *lens*, not a guarantee: we use it to explain mechanisms and to derive reference bounds, check each prediction against real scores, and never rely on it for the validity of the deployed procedure (Section 6).

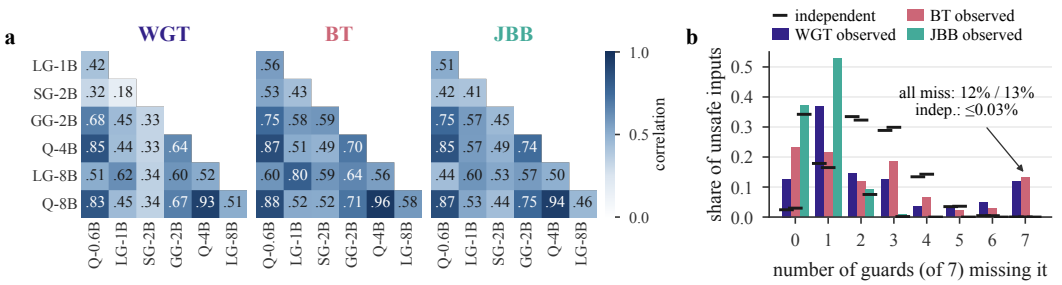


Figure 2: Safety judges agree too much. **(a)** Class-centred correlations of the normal scores (Q: Qwen3Guard, LG: Llama-Guard-3, SG: ShieldGemma, GG: Granite-Guardian; CIs in Table 15). **(b)** Number of guards missing each unsafe input at native thresholds (bars) vs. independent errors with the same miss rates (ticks).

4 Naive fusion silently breaks the false-alarm guarantee

The most common way to combine guards is to multiply their individual likelihood ratios (equivalently, naive Bayes on the guard outputs):

$$E_{\text{ind}}(z) = \prod_{k=1}^{K} \frac{f_{1,k}(z_k)}{f_{0,k}(z_k)}, \tag{3}$$

where $f_{j,k}$ is the marginal density of guard k under H_j. If the guards were independent under H_0, E_{ind} would be an e-value ($\mathbb{E}_0 E_{\text{ind}} = 1$), and thresholding it at $1/\alpha$ would keep the false alarm rate (FPR) below α by Markov's inequality. Guards are not independent, and the failure is exactly quantifiable.

Theorem 4.1 (Excess evidence of naive fusion)**.** *In $\mathcal{M}(d, R)$, $\log E_{\text{ind}} = d^\top z - \frac{1}{2}\|d\|^2$, and under H_0*

$$\log E_{\text{ind}} \sim \mathcal{N}\big(-\tfrac{1}{2}\|d\|^2,\ d^\top R\, d\big), \qquad \log \mathbb{E}_0 E_{\text{ind}} = \tfrac{1}{2}\, d^\top (R - I)\, d = \tfrac{1}{2} \sum_{k \neq l} d_k d_l R_{kl}. \tag{4}$$

Consequently the test "reject if $E_{\text{ind}} \geq 1/\alpha$" has

$$\text{FPR} = \bar{\Phi}\left(\frac{\log(1/\alpha) + \frac{1}{2}\|d\|^2}{\sqrt{d^\top R\, d}} \right), \tag{5}$$

which is at most α when $R = I$. Under dependence the product is an e-value iff $d^\top (R - I) d \leq 0$; with positively correlated guards whose separations d_k share a sign this fails, and the FPR can exceed α substantially (Corollary 4.2). For heteroscedastic Gaussian logits, $\mathbb{E}_0 E_{\text{ind}} = 1$ whenever the safe-class covariance is diagonal (Appendix A.2): with correctly specified marginal likelihood ratios, correlation on safe inputs is the only *source of failure.*

The ratio VIF $= d^\top R d / \|d\|^2$ of the actual to the assumed variance of the fused log-evidence exceeds one when guards agree on the signal ($d_k d_l > 0$) and are positively correlated: the product counts the same evidence several times.

Corollary 4.2 (Failure grows with K; asymptotic floor independent of α)**.** *Let $d = a\mathbf{1}$ with $a > 0$ and $R = (1 - \rho) I + \rho \mathbf{1}\mathbf{1}^\top$ with $\rho \in (0, 1)$. Then $\log \mathbb{E}_0 E_{\text{ind}} = \binom{K}{2} a^2 \rho$, and for every fixed α,*

$$\lim_{K \to \infty} \text{FPR}_K = \bar{\Phi}\Big(\frac{a}{2\sqrt{\rho}} \Big) > 0.$$

In the model, adding correlated guards drives the naive product to a false-alarm floor set only by per-guard strength and correlation, so every nominal level below it is eventually violated; this is a model-based warning, not a law for real guards (see the check below).

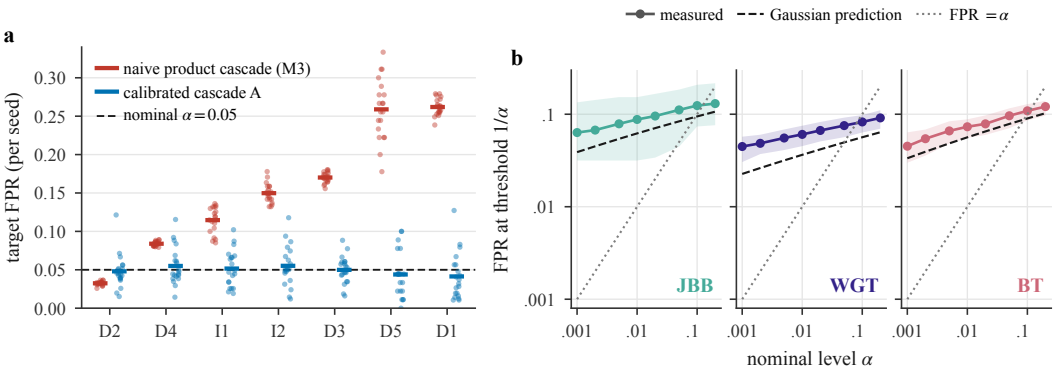


Figure 3: Naive fusion breaks the false-alarm guarantee. **(a)** Target FPR of the uncalibrated naive product cascade M3 and the calibrated cascade A; dots: 20 seeds, bars: mean ($n_{\text{cal}} = 100$, $\alpha = 0.05$; Table 4). **(b)** Held-out FPR of the product of all seven likelihood ratios at threshold $1/\alpha$, measured (mean, 2.5–97.5% range over 50 halvings) and predicted by Theorem 4.1 from (d, R_0) fitted on the other half.

Sequential version: the practitioner's cascade. Deployed systems call a cheap guard first and escalate if unsure, accumulating evidence as a running product; the same mechanism operates step by step.

Proposition 4.3 (Per-step excess of a naive cascade)**.** *In $\mathcal{M}(d, R)$, let a cascade have called the set S and consider guard $k \notin S$. Write $\gamma = R_{SS}^{-1} R_{Sk}$ and $v = 1 - R_{kS}\gamma$ (the H_0 regression of z_k on z_S). The naive increment $\tilde{r}_k = f_{1,k}(z_k)/f_{0,k}(z_k) = \exp(d_k z_k - \frac{1}{2}d_k^2)$ satisfies*

$$\mathbb{E}_0\big[\tilde{r}_k \mid z_S\big] = \exp\Big(d_k\, \gamma^\top z_S - \tfrac{1}{2} d_k^2 (1 - v)\Big), \tag{6}$$

whereas the correct conditional likelihood-ratio increment has conditional mean exactly one.

The log of this conditional mean is d_k times the *linear prediction $\gamma^\top z_S$ of the new guard from the guards already called*, offset by $-\frac{1}{2}d_k^2(1 - v)$: at every step the naive cascade counts again the part of the new score it could already have predicted, and the increment is anti-conservative whenever earlier guards lean far enough towards "unsafe". The *joint* likelihood ratio of the called guards, by contrast, is a nonnegative H_0-martingale for any adaptive calling rule (Lemma A.2).

Empirical check. Fitting d and the safe-class correlation R_0 on one half of the data and evaluating $\log E_{\text{ind}}$ on the held-out safe half (50 halvings), the closed form tracks its mean and variance to within 5% (variance 124.6 measured vs. 122.6 predicted on WildGuardTest). The FPR at the $1/\alpha$ threshold, however, exceeds the prediction by a nearly constant 0.01–0.03 over $\alpha \in [0.001, 0.2]$ (Figure 3b; at $\alpha = 0.05$, 0.075 vs. 0.050 on WGT), and at $\alpha = 0.001$ it is still 0.045–0.063, 45–63 times nominal, as the α-independent floor of Corollary 4.2 suggests. With the first two moments matched, this is an upper-tail mismatch at the test threshold, where the theorem *understates* the failure (Appendix B.2). The naive product cascade (M3) exceeds $\alpha = 0.05$ in 6 of 7 settings, with FPR up to 0.262 (Figure 3a).

5 Correlation Caps the Gain of Fusion

The same correlation bounds how much *any* combination of the guards can improve detection, and hence how close a single cheap guard already is to the best achievable.

5.1 Innovations

Lemma 5.1 (Innovation decomposition)**.** *In $\mathcal{M}(d, R)$, for $S \subseteq [K]$ and $k \notin S$, the log-likelihood-ratio increment from calling k after S is*

$$\log \frac{f_1(z_k \mid z_S)}{f_0(z_k \mid z_S)} = \eta_{k|S}\, u - \tfrac{1}{2}\eta_{k|S}^2, \qquad u = \frac{z_k - \gamma^\top z_S}{\sqrt{v}}, \quad \eta_{k|S} = \frac{d_k - \gamma^\top d_S}{\sqrt{v}}, \tag{7}$$

with γ, v as in Proposition 4.3, where $u \mid z_S \sim \mathcal{N}(0,1)$ under H_0 and $\mathcal{N}(\eta_{k|S}, 1)$ under H_1, regardless of the value of z_S. Moreover $\eta^2_{k|S} = I(S \cup \{k\}) - I(S)$ (a Schur complement).

$\eta^2_{k|S}$ is the *marginal information* of guard k given S. For two guards with separations a, b and correlation ρ, $\eta^2_{2|1} = (b - \rho a)^2/(1-\rho^2)$, which for $a = b$ equals $a^2(1-\rho)/(1+\rho)$: correlation is redundancy. On BeaverTails, Qwen3Guard-4B called after Qwen3Guard-0.6B contributes $\eta^2 = 0.05$, i.e. 1.4% of its standalone information $d^2 = 3.45$ (12.6% on WGT, 28.7% on JBB).

Summing these increments along the calling path, any adaptive cascade runs a Gaussian random walk whose step variances are the predictable information increments η^2_t (an "information clock"; Lemma A.2, Appendix A): correlation enters only through $I(\cdot)$. An information–cost converse in the style of Naghshvar & Javidi (2013) follows (Appendix A.6) but is loose on our guards, so the informative statement is about detection.

5.2 The redundancy sandwich

Theorem 5.2 (Redundancy sandwich)**.** *In $\mathcal{M}(d, R)$, fix $\alpha \in (0,1)$ and let $\beta^\star(\alpha)$ be the largest detection rate at* $\mathrm{FPR} \le \alpha$ *achievable by* any *decision rule that is a (possibly randomized, possibly adaptive) function of the K guard scores, at any cost. For any $S \subseteq [K]$ with $I(S) > 0$, let $H_S = I([K]) - I(S) \ge 0$ be the information the remaining guards add to S (the sum of innovations along any calling order that starts with S). Then*

$$\underbrace{\bar{\Phi}\big(z_{1-\alpha} - \sqrt{I(S)}\big)}_{\text{call only } S} \;\le\; \beta^\star(\alpha) \;=\; \bar{\Phi}\big(z_{1-\alpha} - \sqrt{I(S) + H_S}\big), \tag{8}$$

where $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$ and the left-hand side, $\beta_S(\alpha)$, is the best detection rate using the guards in S alone; the gap satisfies

$$\beta^\star(\alpha) - \beta_S(\alpha) \;\le\; \frac{\sqrt{I(S) + H_S} - \sqrt{I(S)}}{\sqrt{2\pi}} \;\le\; \frac{H_S}{2\sqrt{2\pi\, I(S)}}. \tag{9}$$

For a single guard $S = \{k\}$, $I(S) = d_k^2$, and we write $H_k = H_{\{k\}}$. With $d = a\mathbf{1}$ and equicorrelation ρ, $H_k = a^2(K-1)(1-\rho)/(1+(K-1)\rho)$, which vanishes linearly in $1-\rho$.

Reading the theorem. The Neyman–Pearson power formula for a Gaussian subset is classical (Kay, 1998; Coutino et al., 2018) and the sandwich is elementary; its value lies in the reading, with the ceiling covering adaptive and randomized rules of arbitrary cost. As a *ceiling*, H_S bounds what any fusion rule, stacking model or cascade can add to S; as a *floor*, what is lost by calling only S. When the remaining guards are strongly correlated with S *in the direction of the signal*, H_S is small and, in the model, no adaptive method beats a well-chosen static subset by more than (9). No particular cascade need attain the ceiling, and rules that read the text itself, or new guards, are outside it.

A model-free version. The sandwich does not need the Gaussian model; only its constants do. For arbitrary score laws, the gain of the best rule on all guards over the best test on Z_S is at most the expected total-variation distance between the safe and unsafe conditional laws of the remaining scores given Z_S, which is $2\Phi(\sqrt{H_S}/2) - 1$ in the model (Proposition A.5, Appendix A). Fusion can only help on inputs where, *given what the guards in S already say*, the remaining guards still look different on safe and unsafe content; co-failure (Section 3) is consistent with this being rare but does not measure it. The bound ignores α and is loose; the Gaussian constants are what make the gap small in numbers.

Empirical check (a diagnostic of the model gap, not a test of $\beta^\star$). Figure 4 and Appendix Table 1 use (d, R) estimated on each dataset, with R the pooled class-centred correlation. For Qwen3Guard-0.6B (2.7% of full cost), the model gap is 0.120 on WGT and 0.039 on BT. The measured gap between the full-ensemble LDA and the 0.6B guard is smaller: 0.056 $[0.007, 0.109]$ on WGT and 0.025 $[0.001, 0.060]$ on BT in sample (bootstrap intervals), and 0.048 and 0.023 when fitted on one half and measured on the other at a matched held-out FPR (50 splits). Other fusion

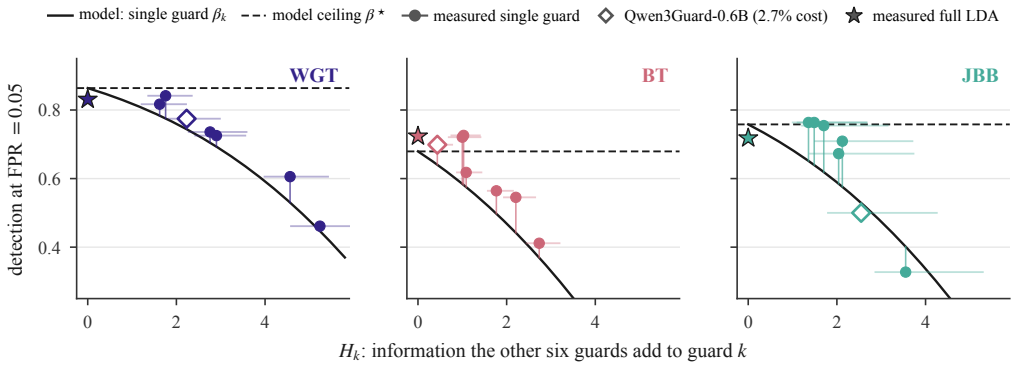


Figure 4: The redundancy sandwich on real guards ($K = 7$, $\alpha = 0.05$, in-sample). Solid: model detection of a single guard as a function of H_k (exact, since $d_k^2 = I([K]) - H_k$), reaching the model ceiling $\beta^\star$ (dashed) at $H_k = 0$. Points: measured detection of each guard (bars: bootstrap 95% interval of H_k); diamond: Qwen3Guard-0.6B; star: full-ensemble LDA.

rules fitted this way (logistic regression, boosted trees) gain at most 0.027 over the best single guard (Appendix B.3). The absolute ceiling is not reliable: on BT three single guards exceed it (0.699, 0.721, 0.726 vs. 0.679), as unsafe-class variances differ from one (0.83–1.30). We therefore use Theorem 5.2 for what it predicts robustly, the *size of the gap* and the ranking by H_S, not as a certified bound on real data. It also flags when the cheapest guard is *not* enough: on JBB, Qwen3Guard-0.6B ($H_k = 2.55$) detects only 0.500 and Qwen3Guard-4B ($H_k = 1.36$) 0.764, a ranking the cross-fitted check supports on 300 responses, though not the gap size.

6 Valid and Cheap Deployment

6.1 Model-free validity for any cascade

Deployment needs a guarantee that holds when the model is wrong and traffic has shifted. We separate a cascade's *skeleton*, fitted on any source data, from a *single scalar* calibrated on target-domain safe examples.

Construction. A skeleton consists of a calling rule $\kappa(S, \ell)$, lower ("safe") boundaries a_S, and upper ("unsafe") boundaries b_S, where ℓ is any running statistic computable online. For $\theta \in \mathbb{R}$, the policy π_θ at each step *first* checks $\ell \geq b_S + \theta$ (alarm and stop), then $\ell \leq a_S$ (safe and stop), and otherwise calls $\kappa(S, \ell)$; after all guards it alarms iff $\ell \geq b_{[K]} + \theta$. Because the safe-stop test and the calling rule do not depend on θ, π_θ alarms on x iff $\theta \leq \theta^*(x)$, the largest value of $\ell_t - b_{S_t}$ along the path of the never-alarming policy $\pi_{+\infty}$ (Lemma B.1, Appendix B.4).

Theorem 6.1 (Finite-sample false-alarm control for cascades)**.** *Let $X_1, \ldots, X_n$ be safe examples from the target domain, exchangeable with a safe test input X_{n+1}, and let the skeleton be independent of them. Let $\hat{\theta}$ be the $\lceil (n+1)(1-\alpha) \rceil$-th smallest of $\theta^*(X_1), \ldots, \theta^*(X_n)$ ($+\infty$ if that index exceeds n), with ties broken at random, and alarm iff $\theta^*(X_{n+1}) > \hat{\theta}$ (equivalently, deploy π_θ with θ just above $\hat{\theta}$). Then $\Pr(\text{alarm on } X_{n+1}) \leq \alpha$, for any skeleton, any dependence among guards, and any misspecification; without ties it is also $\geq \alpha - 1/(n+1)$.*

The proofs (a nesting argument followed by the standard split-conformal rank argument; Vovk et al., 2005; Lei et al., 2018) are in Appendix B.4. The guarantee needs only safe labels, and at least $n \geq 1/\alpha - 1$ of them (19 for $\alpha = 0.05$); below this the cascade never alarms. If calibration and future safe inputs are moreover i.i.d. with a continuous law of θ^*, the calibration-conditional FPR is distributed as $\mathrm{Beta}(m, n+1-m)$ with $m = \lfloor (n+1)\alpha \rfloor$ (mean $5/101 = 0.0495$, s.d. 0.0215 at $n = 100$, $\alpha = 0.05$), which sizes n for a PAC-style requirement (Appendix B.4). The choice of skeleton is thus purely an efficiency question, with limited headroom by Section 5.

6.2 VALIDITY UNDER DISTRIBUTION SHIFT

Source-calibrated fusion is not valid on a shifted target. The soft conformal e-value $e(x) = (n + 1)\, s(x)/\big(s(x)+\sum_{i\le n} s(X_i)\big)$, $s = \exp(\lambda \cdot \mathrm{LDA})$, fitted and calibrated on the source, has mean 4.86 on target safe inputs for WGT vanilla→adversarial and 21.1 for WGT→JBB (it should be at most 1); at $\alpha = 0.05$ the joint fusion reaches FPR 0.061 on vanilla→adversarial and the naive product 0.232 (Appendix Figure 6a). Recalibrating on target-domain safe examples brings the target FPR at $e \ge 20$ to at most 0.017 in every tested cross-dataset shift once $n_{\mathrm{cal}} \ge 50$ (conservative, by Markov's inequality). Detection recovers unevenly: to 91–108% of the in-domain reference in three shifts, but less for JBB→WGT and into JBB (Appendix B.5).

6.3 METHOD COMPARISON

Protocol. For each setting and seed we split the data into four disjoint parts. ***F*** (source) fits the normal-score reference, all LDA directions, the cascade skeletons and the static-subset scorers. ***V*** (source) alone selects hyperparameters and operating points: P-match is the cheapest configuration whose V-detection is within 0.02 of full LDA, P-cheap the most detecting one costing at most 10% of full; the static subset M6 is chosen among all 127 by the same rule. ***C*** (n_{cal} target safe examples) calibrates the single threshold of every calibrated method, and ***T*** (the remaining target examples) is used only for reporting. No target example enters F or V, as Theorem 6.1 requires. The primary configuration (D1, $n_{\mathrm{cal}} = 100$, $\alpha = 0.05$) and the expected outcome were fixed before test results were computed; the full data flow and all 32 configurations are in Appendix B.6. *Settings.* I1 (WGT), I2 (BT); shifts D1/D2 (WGT vanilla↔adversarial), D3/D4 (WGT↔BT), D5 (WGT→JBB). *Methods* (Appendix C.10): M0 one guard, M1 full LDA, M3 uncalibrated naive product cascade, M4c calibrated Trust-or-Escalate, M5 dependency-blind cascade, M6 static subset, A dependency-aware cascade.

Validity. A descriptive screening of the seed-mean target FPR against a pre-set stop line flags 4 of 220 method–configuration cells, each by at most 0.0024, and when the benchmark items are resampled, the 95% interval of the mean FPR contains α in 45 of the 49 method–setting cells at $n_{\mathrm{cal}} = 100$, $\alpha = 0.05$ and lies below it in the other four (Appendix B.6).

Efficiency. In D1 (Figure 5a; Appendix Table 2), the full LDA detects 0.594 at the full cost, the 0.6B guard 0.531 at 2.7% of that cost, M6 0.569 at 8.1%, A 0.570 at 19.9% and M5 0.606 at 32.6%. Across the seven settings M6 retains 96–100% of the full LDA's detection (item-resampling intervals, conditional on the realised splits and selections, have lower ends 0.91–0.98; Appendix C.3), and 71–105% over all 32 configurations, lowest at $\alpha = 0.01$. A is cheaper than M5 in all seven settings, by 2.5–19.9 points of the full cost (every item+split interval excludes zero), and detects 0.002–0.047 less (Figure 5c); A and M6 have no consistent winner in either cost or detection. Full LDA is the comparator by design, not the best detector: Qwen3Guard-8B alone (29% of the cost) detects between 0.0003 less (D2) and 0.091 more (D5) than the calibrated full LDA (Figure 5b). This concerns one fitted rule under the stated target-calibration protocol, not the model optimum $\beta^\star$, but it is consistent with the cap of Section 5.

Selecting subsets by H_S. Two pre-fixed rules use only (d, R) from F and the model gap $g(S) = \beta^\star - \beta_S$. ***H-match*** (cheapest S with $g(S) \le 0.02$) costs 11–39% without higher detection than M6, as the model gap overstates the real one. ***H-cheap*** (smallest $g(S)$ at cost $\le 10\%$) matches M6 (paired difference -0.013 to $+0.016$, not significant in six settings; $+0.056$ in D5), but picks the same pair in 139 of 140 setting–seed pairs, so H_S is shown only to rank a small candidate set as well as validation detection does (Appendix B.6).

Recommendation. Fix a budget (here in parameter memory), choose the affordable guard or subset on labelled source data, by validation detection (M6) or by ranking with H_S, which needs unsafe source labels through d and also flags when the cheapest guard is not enough, and recalibrate one scalar on target-domain safe examples (Theorem 6.1). Validity follows from the theorem; efficiency is an empirical property of this panel, not of fusion as such: a strong guard weakly correlated with the others along the signal would raise H_S for every subset without it, which ranking on source data would reveal.

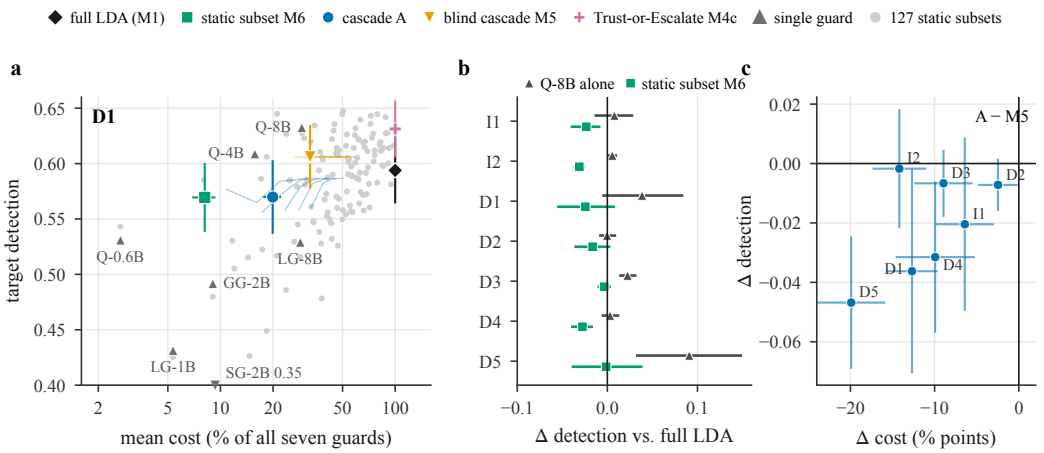


Figure 5: Method comparison on target test data ($n_{\text{cal}} = 100$, $\alpha = 0.05$, 20 seeds). **(a)** D1: selected operating points (mean ± 1 SE), single guards, the 127 static subsets and the grids of A and M5 (thin lines); other settings in Figure 8. **(b)** Detection of Qwen3Guard-8B alone and of M6 minus the calibrated full LDA (conditional item-resampling 95% intervals). **(c)** A minus M5 in cost and detection (item+split intervals).

7 Related work

Combining e-values and evidence under dependence. Products of e-values are valid under independence; under arbitrary dependence, (weighted) averaging is essentially the only admissible merge (Vovk & Wang, 2021; Wang, 2025), and intermediate notions of dependence admit other merges (Ming et al., 2026). Optimal e-statistics for composite nulls are characterised by reverse information projections (Grünwald et al., 2024; Lardy et al., 2024; Larsson et al., 2025). We propose no new merge; we quantify how far the widely used product departs from an e-value.

Sequential testing and sensor selection. Active and cost-aware sequential testing (Chernoff, 1959; Naghshvar & Javidi, 2013; Vershinin et al., 2026; 2025) assumes conditionally independent, repeatable observations and asymptotic regimes; our guards are correlated and each is called once. For Gaussian detection with correlated observations, Neyman–Pearson power is governed by the Mahalanobis quantity we call $I(S)$ (Kay, 1998; Poor, 1994); subset selection with this set function is studied by Coutino et al. (2018), with related results on sensor placement and greedy selection (Krause et al., 2008; Das & Kempe, 2011).

Cascades and correlated failures. Conformal cascades prune candidate sets with coverage guarantees (Fisch et al., 2021); Trust-or-Escalate escalates between LLM judges with a guarantee on human agreement, allowing abstention (Jung et al., 2025) (our baseline M4c). Our guarantee is on the false-alarm rate of a binary decision without abstention, by standard split-conformal arguments (Vovk et al., 2005; Lei et al., 2018). Guards are usually evaluated one at a time; concurrent work documents correlated failures across LLMs and a resulting accuracy ceiling (Chen, 2026); we study the consequences for fusion guarantees and cost.

8 Conclusion

The correlation among safety guards breaks the product of likelihood ratios and, through one quantity H_S, bounds what any combination can gain over a cheap subset; validity can be restored for any cascade by recalibrating one scalar on target-domain safe examples.

AI USE STATEMENT

In this work, we used generative AI tools only for (i) polishing the English grammar and wording of the manuscript and (ii) writing part of the code used in the experiments. We have not used generative AI tools for research ideation, the design and implementation of the methods, the theoretical analysis, or the structure and organization of the paper; these were carried out entirely by the authors. All AI-assisted content was checked in turn by multiple authors: polished text was checked to preserve the intended meaning and technical content, and AI-assisted code was reviewed and verified for correctness. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work evaluates publicly released safety classifiers on public, human-labelled benchmarks of harmful and benign content; no new harmful content was generated for this study and no model was trained. Our findings concern the reliability of false-alarm guarantees in safety filtering; we report failure modes of naive fusion so that deployments can avoid them, not to help circumvent guards.

REPRODUCIBILITY STATEMENT

Model assumptions are stated in Sections 4–5 and all proofs are in Appendix A. The guards, datasets, normal-score construction, data splits and cost proxy are described in Section 2 and the protocol paragraph of Section 6; all random seeds are fixed. Code and stored guard scores will be released as supplementary material.

REFERENCES

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Josef Chen. When does combining language models help? A co-failure ceiling on routing, voting, and mixture-of-agents across 67 frontier models. *arXiv preprint arXiv:2606.27288*, 2026.

Herman Chernoff. Sequential design of experiments. *Annals of Mathematical Statistics*, 30(3): 755–770, 1959.

Mario Coutino, Sundeep Prabhakar Chepuri, and Geert Leus. Submodular sparse sensing for Gaussian detection with correlated observations. *IEEE Transactions on Signal Processing*, 66(15): 4025–4039, 2018.

Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *International Conference on Machine Learning (ICML)*, 2011.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. Efficient conformal prediction via cascaded inference with expanded admission. In *International Conference on Learning Representations (ICLR)*, 2021.

Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 2024.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama Guard: LLM-based input-output safeguard for Human-AI conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. BeaverTails: Towards improved safety alignment of LLM via a human-preference dataset. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. In *International Conference on Learning Representations (ICLR)*, 2025.

Steven M. Kay. *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory*. Prentice Hall PTR, 1998.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(8):235–284, 2008.

Tyron Lardy, Peter Grünwald, and Peter Harremoës. Reverse information projections and optimal e-statistics. *IEEE Transactions on Information Theory*, 70(11):7616–7631, 2024.

Martin Larsson, Aaditya Ramdas, and Johannes Ruf. The numeraire e-variable and reverse information projection. *Annals of Statistics*, 53(3):1015–1043, 2025.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Jiahao Ming, Aaditya Ramdas, Yi Shen, Ruodu Wang, and Ian Waudby-Smith. Combining e-values using demi-supermartingales. *arXiv preprint arXiv:2603.10329*, 2026.

Mohammad Naghshvar and Tara Javidi. Active sequential hypothesis testing. *Annals of Statistics*, 41(6):2703–2738, 2013.

Inkit Padhi, Manish Nagireddy, Giandomenico Cornacchia, Subhajit Chaudhury, Tejaswini Pedapati, Pierre Dognin, Keerthiram Murugesan, Erik Miehling, Martín Santillán Cooper, Kieran Fraser, Giulio Zizzo, Muhammad Zaid Hameed, Mark Purcell, Michael Desmond, Qian Pan, Inge Vejsbjerg, Elizabeth M. Daly, Michael Hind, Werner Geyer, Ambrish Rawat, Kush R. Varshney, and Prasanna Sattigeri. Granite Guardian: Comprehensive LLM safeguarding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, 2025.

H. Vincent Poor. *An Introduction to Signal Detection and Estimation*. Springer, second edition, 1994.

George Vershinin, Asaf Cohen, and Omer Gurewitz. On cost-aware designs for sequential hypothesis testing. *arXiv preprint arXiv:2512.19067*, 2025.

George Vershinin, Asaf Cohen, and Omer Gurewitz. Active sequential hypothesis testing with non-homogeneous costs. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *Annals of Statistics*, 49(3):1736–1754, 2021.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Ruodu Wang. The only admissible way of merging arbitrary e-values. *Biometrika*, 112(2):asaf020, 2025.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. ShieldGemma: Generative AI content moderation based on Gemma. *arXiv preprint arXiv:2407.21772*, 2024.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, Baosong Yang, Chen Cheng, Jialong Tang, Jiandong Jiang, Jianwei Zhang, Jijie Xu, Ming Yan, Minmin Sun, Pei Zhang, Pengjun Xie, Qiaoyu Tang, Qin Zhu, Rong Zhang, Shibin Wu, Shuo Zhang, Tao He, Tianyi Tang, Tingyu Xia, Wei Liao, Weizhou Shen, Wenbiao Yin, Wenmeng Zhou, Wenyuan Yu, Xiaobin Wang, Xiaodong Deng, Xiaodong Xu, Xinyu Zhang, Yang Liu, Yeqiu Li, Yi Zhang, Yong Jiang, Yu Wan, and Yuxin Zhou. Qwen3Guard technical report. *arXiv preprint arXiv:2510.14276*, 2025.

A Proofs for Sections 4–5

Throughout, $\mathcal{M}(d, R)$ is (1); φ, Φ are the standard normal density and CDF, $\bar{\Phi} = 1 - \Phi$, and $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$. For $S \subseteq [K]$, $k \notin S$: $\gamma_{k|S} = R_{SS}^{-1} R_{Sk}$, $v_{k|S} = 1 - R_{kS}\gamma_{k|S} > 0$ (since $R \succ 0$); we drop the subscripts when k, S are fixed.

A.1 Theorem 4.1 and Corollary 4.2

Proof of Theorem 4.1. Under $\mathcal{M}(d, R)$ the marginals are $f_{0,k} = \mathcal{N}(0,1)$ and $f_{1,k} = \mathcal{N}(d_k, 1)$, so $\log f_{1,k}(z_k)/f_{0,k}(z_k) = d_k z_k - \frac{1}{2}d_k^2$ and $\log E_{\text{ind}} = d^\top z - \frac{1}{2}\|d\|^2$. Under H_0, $d^\top z \sim \mathcal{N}(0, d^\top R d)$, which gives the law in (4). The lognormal mean is $\exp(-\frac{1}{2}\|d\|^2 + \frac{1}{2}d^\top R d) = \exp(\frac{1}{2}d^\top(R-I)d)$, and $d^\top(R-I)d = \sum_{k\neq l} d_k d_l R_{kl}$ since $R_{kk} = 1$. For (5), $\Pr_0(\log E_{\text{ind}} \geq \log(1/\alpha)) = \Pr\big(\mathcal{N}(0,1) \geq (\log(1/\alpha) + \frac{1}{2}\|d\|^2)/\sqrt{d^\top R d}\big)$. When $R = I$, $\mathbb{E}_0 E_{\text{ind}} = 1$ and Markov's inequality gives $\text{FPR} \leq \alpha$. That the FPR can exceed α substantially follows from Corollary 4.2. The heteroscedastic statement is Proposition A.1. □

Proof of Corollary 4.2. $\sum_{k\neq l} d_k d_l R_{kl} = K(K-1)a^2\rho$, so $\log \mathbb{E}_0 E_{\text{ind}} = \binom{K}{2}a^2\rho$. Further $\|d\|^2 = Ka^2$ and $d^\top R d = a^2 K(1 + (K-1)\rho)$, so the argument of $\bar{\Phi}$ in (5) is $\big(\log(1/\alpha) + \frac{1}{2}Ka^2\big)/\big(a\sqrt{K(1+(K-1)\rho)}\big) \to a/(2\sqrt{\rho})$ as $K \to \infty$, for any fixed α. □

A.2 Heteroscedastic Gaussian logits

Let $l \in \mathbb{R}^K$ be guard logits with $l \sim \mathcal{N}(\mu_j, \Sigma_j)$ under H_j, $\Sigma_j \succ 0$, and $\sigma_{j,k}^2 = (\Sigma_j)_{kk}$. Define $A = \text{diag}(\sigma_{1,k}^{-2} - \sigma_{0,k}^{-2})$, $b_k = \mu_{1,k}\sigma_{1,k}^{-2} - \mu_{0,k}\sigma_{0,k}^{-2}$, and $c = \sum_k \big[\frac{1}{2}\log(\sigma_{0,k}^2/\sigma_{1,k}^2) - \mu_{1,k}^2/(2\sigma_{1,k}^2) + \mu_{0,k}^2/(2\sigma_{0,k}^2)\big]$.

Proposition A.1. *If $\Sigma_0^{-1} + A \succ 0$, then*

$$\mathbb{E}_0 E_{\text{ind}} = e^c \det(I + \Sigma_0 A)^{-1/2} \exp\Big(-\tfrac{1}{2}\mu_0^\top A\mu_0 + b^\top\mu_0 + \tfrac{1}{2}(b - A\mu_0)^\top(\Sigma_0^{-1} + A)^{-1}(b - A\mu_0)\Big),$$

otherwise $\mathbb{E}_0 E_{\text{ind}} = +\infty$. This depends on Σ_1 only through its diagonal, and equals 1 whenever Σ_0 is diagonal (for any Σ_1).

Proof. Expanding the Gaussian log-densities gives $\log E_{\text{ind}}(l) = -\frac{1}{2}l^\top A l + b^\top l + c$. With $w = l - \mu_0 \sim \mathcal{N}(0, \Sigma_0)$ this is $-\frac{1}{2}w^\top A w + g^\top w - \frac{1}{2}\mu_0^\top A\mu_0 + b^\top\mu_0 + c$ with $g = b - A\mu_0$. For $P = \Sigma_0^{-1} + A \succ 0$, completing the square gives $\mathbb{E}\exp(-\frac{1}{2}w^\top A w + g^\top w) = \det(\Sigma_0)^{-1/2}\det(P)^{-1/2}\exp(\frac{1}{2}g^\top P^{-1} g)$ and $\det(\Sigma_0)\det(P) = \det(I + \Sigma_0 A)$. If the symmetric matrix P is not positive definite, take a unit eigenvector e with eigenvalue $\lambda \leq 0$ and write $w = te + w_\perp$ with $w_\perp \perp e$; the joint integrand $\exp(-\frac{1}{2}w^\top P w + g^\top w)$ (Gaussian density included) contains the factor $\exp(-\frac{1}{2}\lambda t^2 + t\,(g^\top e - e^\top P w_\perp))$ with $e^\top P w_\perp = \lambda e^\top w_\perp = 0$, which is not integrable in t for $\lambda < 0$, nor for $\lambda = 0$ (it is then $\exp(t\,g^\top e)$). By Tonelli the expectation is $+\infty$, including the semidefinite case. Σ_1 enters only via $\sigma_{1,k}^2$ through A, b, c. If Σ_0 is diagonal, the H_0 density factorizes, and $\mathbb{E}_0 E_{\text{ind}} = \prod_k \int f_{1,k} = 1$. □

A.3 PROPOSITION 4.3

Proof. Under H_0, $z_k \mid z_S \sim \mathcal{N}(\gamma^\top z_S, v)$. Hence $\mathbb{E}_0[\exp(d_k z_k) \mid z_S] = \exp(d_k \gamma^\top z_S + \frac{1}{2} d_k^2 v)$ and $\mathbb{E}_0[\tilde{r}_k \mid z_S] = \exp(d_k \gamma^\top z_S - \frac{1}{2} d_k^2 (1 - v))$. □

Taking iterated expectations along any fixed order recovers $\mathbb{E}_0 \prod_k \tilde{r}_k = \exp(\frac{1}{2} d^\top (R - I) d)$, consistent with Theorem 4.1.

A.4 LEMMA 5.1

Proof. Under H_0, $z_k \mid z_S \sim \mathcal{N}(\gamma^\top z_S, v)$; under H_1, $z_k \mid z_S \sim \mathcal{N}(d_k + \gamma^\top (z_S - d_S), v)$, with the same γ, v because both hypotheses share R. The two conditional means differ by $\delta = d_k - \gamma^\top d_S$, which does not depend on z_S. The log-ratio of two normal densities with common variance v and means m, $m + \delta$ is $(\delta/\sqrt{v})u - \frac{1}{2}\delta^2/v$ with $u = (z_k - m)/\sqrt{v}$, which is the claim with $\eta = \delta/\sqrt{v}$. For the Schur-complement identity, write $R_{S \cup k}$ in block form; block inversion gives $d_{S\cup k}^\top R_{S\cup k}^{-1} d_{S\cup k} = d_S^\top R_{SS}^{-1} d_S + (d_k - R_{kS} R_{SS}^{-1} d_S)^2 / v$. □

Summing Lemma 5.1 along any chain $S = S_0 \subset S_1 \subset \cdots \subset [K]$ telescopes to $I([K]) - I(S) = H_S$, independently of the order.

A.5 THE INFORMATION CLOCK

Lemma A.2 (Information clock)**.** *Consider any policy that, at each step, chooses the next guard (or stops and decides) as a function of the scores seen so far and of independent randomness. Let S_t be the set called after t steps, $I_t = I(S_t)$, and ℓ_t the joint log-likelihood ratio of z_{S_t}. Then conditionally on the past, $\ell_t - \ell_{t-1} \sim \mathcal{N}(\mp\frac{1}{2}\eta_t^2, \eta_t^2)$ under H_0/H_1, where $\eta_t^2 = I_t - I_{t-1}$ is predictable. In particular $\ell_t + \frac{1}{2} I_t$ is an H_0-martingale and $\exp(\ell_t)$ is a nonnegative H_0-martingale.*

Proof. Let $\mathcal{F}_{t-1}$ contain the scores observed so far and the policy's randomness up to step t. The next guard k_t and hence S_t are $\mathcal{F}_{t-1}$-measurable. Since the policy's randomness is independent of z, the conditional law of z_{k_t} given $\mathcal{F}_{t-1}$ equals its law given $z_{S_{t-1}}$, and Lemma 5.1 applies with $\eta_t = \eta_{k_t | S_{t-1}}$. Because I is determined by S, $\eta_t^2 = I_t - I_{t-1}$. The two martingale statements follow from $\mathbb{E}[\eta u] = 0$ and $\mathbb{E}[\exp(\eta u - \frac{1}{2}\eta^2)] = 1$ for $u \sim \mathcal{N}(0, 1)$. □

A.6 INFORMATION CONVERSE AND THE COST COROLLARY

Proposition A.3 (Information converse)**.** *Any policy as in Lemma A.2 with false-alarm rate α' and detection rate β' satisfies*

$$\mathbb{E}_0[I(S_\tau)] \;\geq\; 2\,\mathrm{kl}(\alpha' \,\|\, \beta'), \qquad \mathbb{E}_1[I(S_\tau)] \;\geq\; 2\,\mathrm{kl}(\beta' \,\|\, \alpha'),$$

where kl *is the binary KL divergence and S_τ is the set called before stopping.*

In an idealised continuous-information analogue (the SPRT without overshoot) the bound is attained; with finitely many guards, each called once, it is generally loose. The proof is the standard KL/data-processing argument; the difference from Naghshvar & Javidi (2013) is that observations are correlated and each guard can be called at most once, and the innovation structure turns the KL into the closed form $\frac{1}{2} I(S_\tau)$. Taking the lower convex hull of $\{(I(S), c(S))\}_S$ yields a cost lower bound (Corollary A.4); on our guards it is loose, because a single guard already carries more information than $2\mathrm{kl}(\alpha \| \beta)$ at practical operating points. The informative statement is about detection, not cost.

Proof of Proposition A.3. Let O be the full record (randomness, called guards, observed scores, decision). Its density under H_j factorizes as $\prod_t \pi(k_t \mid \text{past})\, f_j(z_{k_t} \mid z_{S_{t-1}})$ times the decision kernel. The policy terms are the same under H_0 and H_1 and cancel, so $\log \frac{dP_1^O}{dP_0^O} = \ell_\tau$. Hence $\mathrm{KL}(P_0^O \| P_1^O) = -\mathbb{E}_0 \ell_\tau$. Since $\tau \leq K$ and $\ell_t + \frac{1}{2} I_t$ is an H_0-martingale (Lemma A.2), optional

stopping gives $\mathbb{E}_0\ell_\tau = -\frac{1}{2}\mathbb{E}_0 I_\tau$. The decision $D \in \{0,1\}$ is a function of O, so by data processing $\mathrm{KL}(P_0^O \| P_1^O) \geq \mathrm{kl}(\alpha' \| \beta')$. The H_1 statement is symmetric, using the H_1-martingale $\ell_t - \frac{1}{2}I_t$. □

Corollary A.4 (Cost converse). *Let $\mathcal{C}(x) = \min\{\mathbb{E}_\mu c(S) : \mu$ a distribution on $2^{[K]}$, $\mathbb{E}_\mu I(S) \geq x\}$ (the lower convex hull of $\{(I(S), c(S))\}$). For $\alpha < \beta$, any policy with $\mathrm{FPR} \leq \alpha$ and detection rate at least β has $\mathbb{E}_0[c(S_\tau)] \geq \mathcal{C}(2\mathrm{kl}(\alpha\|\beta))$ and $\mathbb{E}_1[c(S_\tau)] \geq \mathcal{C}(2\mathrm{kl}(\beta\|\alpha))$.*

Proof. On $a < b$, $\mathrm{kl}(a\|b)$ is decreasing in a and increasing in b, and $\mathrm{kl}(b\|a)$ is increasing in b and decreasing in a, so Proposition A.3 holds with (α, β) in place of (α', β'); $\mathcal{C}$ is nondecreasing. The H_0-law of S_τ is feasible for the program defining $\mathcal{C}$. □

A.7 Theorem 5.2

Proof. Ceiling. A decision rule reading the scores is a test $\phi(z, U) \in [0,1]$ with independent randomization U; averaging over U gives a test of z with the same size and power. In $\mathcal{M}(d, R)$ the likelihood ratio is $\exp(d^\top R^{-1} z - \frac{1}{2}I([K]))$; the statistic $T = d^\top R^{-1} z$ is $\mathcal{N}(0, I([K]))$ under H_0 and $\mathcal{N}(I([K]), I([K]))$ under H_1. By the Neyman–Pearson lemma the most powerful level-α test has power $\bar{\Phi}(z_{1-\alpha} - \sqrt{I([K])})$, so $\beta^\star(\alpha)$ equals this value. *Floor.* The same argument applied to z_S gives $\beta_S(\alpha) = \bar{\Phi}(z_{1-\alpha} - \sqrt{I(S)})$, attained at cost $c(S)$. Since $I([K]) = I(S) + H_S$, this is (8). *Gap.* $\beta^\star - \beta_S = \int_{\sqrt{I(S)}}^{\sqrt{I([K])}} \varphi(z_{1-\alpha} - t)\, dt \leq (\sqrt{I(S) + H_S} - \sqrt{I(S)})/\sqrt{2\pi}$, and $\sqrt{x+h} - \sqrt{x} \leq h/(2\sqrt{x})$ for $x > 0$, $h \geq 0$. *Equicorrelation.* $R^{-1}\mathbf{1} = \mathbf{1}/(1 + (K-1)\rho)$, so $I([K]) = a^2 K/(1 + (K-1)\rho)$ and $H_k = I([K]) - a^2 = a^2(K-1)(1-\rho)/(1 + (K-1)\rho)$. □

A.8 Model-free redundancy sandwich

Proposition A.5 (Model-free redundancy sandwich). *Let P_0, P_1 be* arbitrary *joint laws of the guard scores $Z = (Z_S, Z_{-S})$ with $P_{1,S} \ll P_{0,S}$ for the marginals of Z_S, and write $P_j(\cdot \mid Z_S)$ for the conditional law of Z_{-S} given Z_S under P_j (defined P_1-a.s.). Let $\beta_S^{\mathrm{NP}}(\alpha)$ be the Neyman–Pearson power of the best level-α test based on Z_S alone, and $\beta^\star(\alpha)$ the best detection rate at $\mathrm{FPR} \leq \alpha$ of any rule based on Z, as in Theorem 5.2. Then*

$$0 \;\leq\; \beta^\star(\alpha) - \beta_S^{\mathrm{NP}}(\alpha) \;\leq\; \mathbb{E}_{Z_S \sim P_1} \mathrm{TV}\big(P_1(\cdot \mid Z_S),\, P_0(\cdot \mid Z_S)\big) \;\leq\; \sqrt{\tfrac{1}{2}\,\mathrm{KL}_{\mathrm{cond}}}, \tag{10}$$

with $\mathrm{KL}_{\mathrm{cond}} = \mathbb{E}_{Z_S \sim P_1} \mathrm{KL}\big(P_1(\cdot \mid Z_S) \,\|\, P_0(\cdot \mid Z_S)\big)$. In $\mathcal{M}(d, R)$, $\mathrm{KL}_{\mathrm{cond}} = \frac{1}{2}H_S$.

Proof. Let $\phi(Z, U)$ be any level-α test (U independent randomization) and define $\psi(z_S) = \mathbb{E}_0[\phi(Z, U) \mid Z_S = z_S] \in [0,1]$, using a regular conditional distribution of (Z_{-S}, U) given Z_S under P_0. Then $\mathbb{E}_0\psi(Z_S) = \mathbb{E}_0\phi \leq \alpha$, so ψ is a level-α test based on Z_S and $\mathbb{E}_1\psi(Z_S) \leq \beta_S^{\mathrm{NP}}(\alpha)$. Moreover

$$\begin{aligned}\mathbb{E}_1\phi - \mathbb{E}_1\psi(Z_S) &= \mathbb{E}_{Z_S \sim P_1}\Big[\int \phi\, dP_1(\cdot \mid Z_S) - \int \phi\, dP_0(\cdot \mid Z_S)\Big] \\ &\leq \mathbb{E}_{Z_S \sim P_1} \mathrm{TV}\big(P_1(\cdot \mid Z_S), P_0(\cdot \mid Z_S)\big),\end{aligned}$$

because $0 \leq \phi \leq 1$ (U has the same law under both conditionals). Taking the supremum over ϕ gives the first inequality; the lower bound 0 holds because tests of Z_S are tests of Z. Pinsker's inequality and Jensen's inequality ($\sqrt{\cdot}$ is concave) give the last one. In $\mathcal{M}(d, R)$, given z_S the remaining scores are Gaussian with common covariance $V = R_{-S,-S} - R_{-S,S}R_{SS}^{-1}R_{S,-S}$ and mean difference $\Delta = d_{-S} - R_{-S,S}R_{SS}^{-1}d_S$ (not depending on z_S), so $\mathrm{KL}_{\mathrm{cond}} = \frac{1}{2}\Delta^\top V^{-1}\Delta = \frac{1}{2}(I([K]) - I(S)) = \frac{1}{2}H_S$ by the block-inverse identity; in this case the TV term equals $2\Phi(\sqrt{H_S}/2) - 1$. □

B Details deferred from the main text

B.1 Normal scores, ties, ridge and cost

The reference set is always the safe part of whatever split is used for fitting: all safe examples for the descriptive statistics of Section 3 and the in-sample sandwich table, the fitting half in the held-out

checks of Sections 4–5, and the fitting split F in Section 6; evaluation examples never enter it in held-out analyses. Model parameters are plug-in estimates on the same fitting data: d is the difference of class means of z, R the correlation matrix of the class-centred scores pooled over both classes, plus $10^{-3}I$ (a ridge; the regularised matrix, with diagonal 1.001, is used in place of R in all model formulas; rescaling it to unit diagonal changes $d^\top R^{-1} d$ by exactly 0.1% and the model ceiling $\beta^\star$ by at most 0.0004 on the three datasets), and LDA uses the pooled class-centred covariance with the same ridge. Ties are broken by adding independent $U(0, 10^{-9})$ jitter to the final calibration statistic of each method (the scalar score for fixed rules, θ^* itself for cascades), on calibration and test inputs alike, which implements the random tie-breaking of Theorem 6.1. Ties are pervasive: raw scores are log-odds computed from bf16 logits and take only 67–215 distinct values on the safe inputs of a benchmark, and on WildGuardTest and BeaverTails 96–99% of safe scores share their value with another safe input. The normal scores are therefore a coarse, discretised approximation to $\mathcal{N}(0, 1)$, and the Gaussian-copula model of Sections 4–5 is applied to this approximation; the validity results of Section 6 do not depend on it. We measure the cost of calling a guard by its parameter bytes in bf16 (1.5 GB for Qwen3Guard-0.6B, 56.06 GB for all seven), a parameter-memory proxy; we refer to it as cost throughout. As a compute proxy we also use the non-embedding parameter count (the input embedding is a lookup, and the guards read a single verdict logit), which ignores sequence length, attention and template length; re-pricing every selected operating point this way, with the selections unchanged, moves each method's cost by at most 2.0 points of the full cost and changes no comparison reported in Section 6.3 (Appendix C.4).

B.2 Failure of naive fusion: details of the empirical check

We estimate d and the *safe-class* correlation, written R_0 here (the null covariance of z, which is what the theorem needs; in the model $R_0 = R$) on one half of the safe and unsafe examples and measure the null distribution of $\log E_{\mathrm{ind}}$ on the other, disjoint half of the safe examples (50 random halvings), so the moment check is out of sample. On the $K = 7$ guards, the closed form tracks the mean and variance of $\log E_{\mathrm{ind}}$ on held-out safe inputs to within 5% (e.g. variance 124.6 measured vs. 122.6 predicted on WildGuardTest), but the measured FPR at $\alpha = 0.05$ is higher than predicted (JBB 0.111 vs. 0.084; WGT 0.075 vs. 0.050; BT 0.096 vs. 0.079; s.d. over halvings 0.034, 0.010, 0.012), and likewise at $\alpha = 0.01$ (0.088, 0.061, 0.073 vs. 0.062, 0.036, 0.056, i.e. 6–9 times the nominal level). Over $\alpha \in [0.001, 0.2]$ (Figure 3b) the excess of measured over predicted FPR is nearly constant, 0.01–0.03 (its 95% range over halvings excludes zero at every α on WGT, but not on JBB or BT), and both curves are flat in α: at $\alpha = 0.001$ the measured FPR is still 0.045–0.063, 45–63 times nominal, as the α-independent floor of Corollary 4.2 suggests; the product is conservative only at $\alpha = 0.2$ (and at 0.1 on WGT). With the first two moments matched, this is an upper-tail mismatch at the test threshold (we do not claim globally heavier tails); there, the theorem *understates* the failure. For the cascade, the naive product cascade (M3) exceeds $\alpha = 0.05$ in 6 of 7 settings, with FPR up to 0.262 (Figure 3a; Table 4 in the appendix).

B.3 Redundancy sandwich: details of the empirical check

The ceiling $\beta^\star$ is the NP optimum in the fitted equal-covariance model; what can be measured is the gap between a fitted fusion rule (full LDA) and single guards. Here R is estimated by the pooled class-centred correlation R_{pool}, as the equal-covariance model requires (the failure check of Section 4 instead uses the safe-class correlation R_0; the two differ on real scores). Table 1 and Figure 4 evaluate the theorem with (d, R) estimated on each dataset ($K = 7$, $\alpha = 0.05$). For Qwen3Guard-0.6B (2.7% of full cost), the model gap is 0.120 on WGT (bound 0.176) and 0.039 on BT (bound 0.042). The measured in-sample gap between the full-ensemble LDA and the single 0.6B guard is smaller in both cases: $0.831 - 0.775 = 0.056$ on WGT and $0.724 - 0.699 = 0.025$ on BT (bootstrap 95% intervals $[0.007, 0.109]$ and $[0.001, 0.060]$). These differences are not an artefact of fitting and evaluating on the same data: estimating (d, R) and the LDA direction on one half and measuring detection on the other (50 stratified splits; as in the definition of $\beta^\star$, every detection rate is taken at the held-out 95% safe quantile, i.e. at a matched empirical held-out FPR; this is a cross-fitted ROC comparison rather than an evaluation of a threshold fixed in advance) gives 0.048 $[0.002, 0.104]$ on WGT and 0.023 $[-0.002, 0.046]$ on BT, again below the corresponding model gaps (0.133 and 0.044), although on BT the upper percentile reaches the model gap (Appendix Table 14). Replacing the Pearson correlation by the Kendall-τ estimate $\sin(\pi\tau/2)$ moves the ceiling by at most 0.006. (For single guards $H_k = I([K]) - d_k^2$, so their ranking by H_k is that of d_k^2 and cannot

change with R; it is not evidence of robustness.) The fitted LDA is one fusion rule; its small held-out gain is consistent with the ceiling but does not certify it for every rule. We therefore repeated the cross-fit with other rules fitted on half A (logistic regression on z, gradient-boosted trees on raw scores) against the best single guard chosen by A-half AUROC (usually Qwen3Guard-8B), all at the B-half 95% safe quantile: the mean gain is at most 0.027 for LDA and logistic regression (WGT 0.008, BT 0.007, JBB 0.027), and boosted trees do worse than the single guard (Appendix Table 17). Setting each threshold on half A instead, the held-out FPR of LDA is 0.052–0.060 and its gain over Qwen3Guard-0.6B changes little (0.056 WGT, 0.026 BT, 0.237 JBB, vs. 0.048, 0.023, 0.170 at matched FPR). The absolute level of the model ceiling is not reliable on BT: three single guards already exceed it empirically (0.699, 0.721, 0.726 vs. 0.679), because BT's unsafe-class score variances differ from one (0.83–1.30), which the equal-covariance model ignores. We therefore use Theorem 5.2 for what it predicts robustly, namely the *size of the gap* and the ranking of guards by H_k, not as a certified upper bound on real data. The sandwich also says when the cheapest guard is *not* enough: on JBB, Qwen3Guard-0.6B has a large $H_k = 2.55$ and a model gap of 0.229, and indeed detects only 0.500, whereas Qwen3Guard-4B ($H_k = 1.36$) detects 0.764 (Table 3; in sample). Cross-fitted, the full LDA exceeds 0.6B by 0.170 $[-0.093, 0.396]$ but 4B by only -0.005 (Table 14): the ranking survives, the size of the gap is imprecise on 300 responses.

Table 1: Redundancy sandwich on real guards ($K = 7$, $\alpha = 0.05$, in-sample). H_k: information added by the other six guards to guard k. "Model gap" $= \beta^\star - \beta_k$ in $\mathcal{M}(d, R)$; "Bound": middle term of (9); "Emp.": measured single-guard detection. Full LDA: WGT 0.831, BT 0.724. Model ceiling $\beta^\star$: WGT 0.864, BT 0.679.

Guard	Cost (%)	WGT				BT			
		d_k^2	H_k	Gap/Bound	Emp.	d_k^2	H_k	Gap/Bound	Emp.
Qwen3Guard-0.6B	2.7	5.29	2.23	.120/.176	0.775	4.02	0.44	.039/.042	0.699
LlamaGuard3-1B	5.4	2.96	4.56	.334/.408	0.606	2.25	2.21	.237/.244	0.545
Granite-Guardian-3.1-2B	9.0	4.76	2.76	.159/.224	0.736	3.37	1.09	.104/.110	0.618
ShieldGemma-2B	9.3	2.29	5.24	.417/.491	0.461	1.72	2.73	.310/.319	0.411
Qwen3Guard-4B	15.7	5.77	1.76	.089/.136	0.842	3.45	1.00	.095/.100	0.721
LlamaGuard3-8B	28.6	4.62	2.90	.171/.237	0.725	2.69	1.77	.182/.188	0.564
Qwen3Guard-8B	29.2	5.90	1.62	.080/.125	0.817	3.43	1.03	.098/.104	0.726

B.4 Proofs and consequences of the calibration theorem

Lemma B.1 (Nesting). *For every input x, $\pi_\theta(x)$ raises an alarm iff $\theta \le \theta^*(x)$, where*

$$\theta^*(x) := \max_{t \le \tau_\infty(x)} \big(\ell_t(x) - b_{S_t}(x)\big)$$

and the path (S_t, ℓ_t) and stopping time τ_∞ are those of the never-alarming policy $\pi_{+\infty}$.

Proofs. Lemma B.1: until the first alarm, π_θ and $\pi_{+\infty}$ take identical actions, because the safe-stop test and the calling rule do not depend on θ; so π_θ alarms iff the path of $\pi_{+\infty}$ meets $\ell_t - b_{S_t} \ge \theta$ at some $t \le \tau_\infty$, a condition monotone in θ. Theorem 6.1: by the lemma the alarm event is $\{\theta^*(X_{n+1}) > \hat\theta\}$ with θ^* a fixed measurable function, and the claim is the standard split-conformal rank argument (Vovk et al., 2005; Lei et al., 2018). □

Consequences. (i) The guarantee needs only safe labels, and at least $n \ge 1/\alpha - 1$ of them (19 for $\alpha = 0.05$, 99 for $\alpha = 0.01$); below this the cascade never alarms ($m = \lfloor (n+1)\alpha \rfloor = 0$). (ii) If, in addition, calibration and future safe inputs are i.i.d. from a fixed distribution with a continuous law of θ^* (given the skeleton), then over the draw of the calibration set the calibration-conditional FPR $\Pr(\text{alarm on } X_{n+1} \mid X_{1:n})$ is distributed as $\mathrm{Beta}(m, n+1-m)$ for $m \ge 1$, since the alarm region is above the k-th order statistic with $k = \lceil (n+1)(1-\alpha) \rceil$ and $n+1-k = m$ (mean $m/(n+1)$, i.e. $5/101 = 0.0495$ with s.d. 0.0215 at $n = 100$, $\alpha = 0.05$), which sizes n for a PAC-style requirement; exchangeability alone gives the marginal bound of the theorem but not this law. (iii) As a separate rank-based reporting statistic (not the deployed alarm indicator, and not a transformation of the cascade score), replacing the indicator by $(n+1)\mathbf{1}\{\theta^*(x) > \hat\theta\}/(1 + \#\{i : \theta^*(X_i) > \hat\theta\})$ gives,

for the rank threshold $\hat{\theta}$ of the theorem and under its exchangeability and skeleton-independence assumptions, a conformal e-value (its expectation, taken jointly over the calibration set and the test input, i.e. marginally and not conditionally on the calibration set, is exactly 1 without ties, because $\#\{i : \theta^*(X_i) > \hat{\theta}\} = n-k$); this does not make other thresholded or post-selected cascade statistics e-values. (iv) Because Theorem 6.1 is indifferent to the skeleton, the choice of skeleton is purely an efficiency question, which Section 5 tells us has limited headroom.

B.5 Validity under shift: details

The finite-sample result in Theorem 6.1 matters because source and deployment traffic need not agree. In our shift experiments, an unmodified conformal fusion calibrated on the source is no longer an e-value on the target. This experiment uses the soft conformal e-value $e(x) = (n+1)\, s(x)/\big(s(x) + \sum_{i\le n} s(X_i)\big)$ with $s = \exp(\lambda \cdot \mathrm{LDA})$, which is an e-value when $x, X_1, \ldots, X_n$ are exchangeable safe inputs (*joint*: LDA on all seven normal scores fitted on 40% of the source, X_i a disjoint 30% of source safe examples, λ chosen on the fitting part; *product*: the product of the seven single-guard versions). Its mean on target safe inputs is 4.86 for WildGuardTest vanilla→adversarial (s.d. 2.69 over 50 source resplits; 666 target safe inputs; it should be at most 1), and 21.1 (s.d. 4.21; 190 target safe inputs) for WGT→JBB. In FPR terms the joint fusion reaches 0.061 at nominal $\alpha = 0.05$ on vanilla→adversarial (s.d. 0.030 over resplits, 666 target safe inputs) while the naive product reaches 0.232 (Figure 6). Recalibrating only on target-domain safe examples (recomputing e with X_i drawn from the target) brings the average target FPR at the threshold $e \ge 20$ to at most 0.017 in every tested cross-dataset shift once $n_{\mathrm{cal}} \ge 50$. This is well below 0.05 because thresholding an e-value at $1/\alpha$ is conservative (Markov's inequality); the rank-based thresholds of Section 6.3 are exact instead, so detection levels in the two experiments are not comparable. Detection recovers unevenly. With $n_{\mathrm{cal}} = 100$–200 it reaches 91–108% of the in-domain reference (the same e-value construction fitted and calibrated within the target) for WGT→BT, BT→WGT and JBB→BT, but only 73–77% for JBB→WGT. Into JBB it stays low (0.10 and 0.04 vs. 0.28); this experiment keeps at least half of the target safe examples for testing, so with 190 safe JBB responses only $n_{\mathrm{cal}} \le 50$ of our grid was run (the D5 protocol below uses $n_{\mathrm{cal}} = 100$ and a rank threshold) (Figure 6).

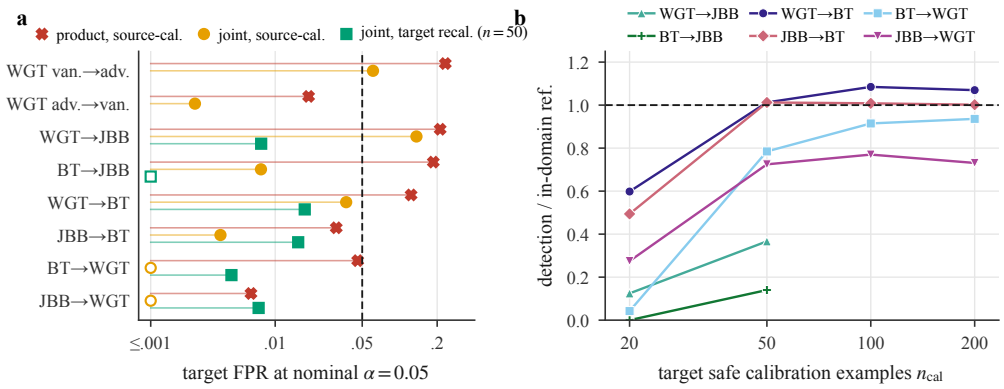


Figure 6: Validity under shift. **(a)** Mean target FPR (50 source resplits) of the soft conformal e-value thresholded at $e \ge 20$ (nominal $\alpha = 0.05$): product and joint fusion calibrated on the source, and joint fusion recalibrated on 50 target safe examples (not run for the WGT-internal shifts); hollow markers: FPR ≤ 0.001, drawn at the axis floor. **(b)** Detection at $e \ge 20$ after target recalibration, relative to the in-domain reference; shifts into JBB were run only up to $n_{\mathrm{cal}} = 50$.

B.6 Method comparison: protocol, screening and item-level results

Protocol and data flow. For each setting and seed we split the data into four disjoint parts. *F* (source; 70% of source examples in D1–D5, 40% in I1–I2) is used to fit the normal-score reference (safe part of F only), all LDA directions, the cascade skeletons and the static-subset scorers. *V* (source; the remaining 30%, or 20% in I1–I2) is the only data used to choose hyperparameters and operating points: P-match is the cheapest configuration whose V-detection is within 0.02 of full

LDA, P-cheap the most detecting configuration with cost at most 10% of full; the M6 subset among the 127 candidates is selected by the same rule on V, so M6 is a validation-selected subset that uses source labels only and could be deployed as is; the 96–100% retention below refers to this protocol. *C* (n_{cal} target safe examples) calibrates the single threshold of every calibrated method, including M4c, with the same budget; AB alone also uses C, estimating part of its skeleton on the first half C_1 and calibrating on the disjoint second half C_2 (so Theorem 6.1 applies to it with $n = |C_2|$). *T* (the remaining target examples) is used only for reporting. No target example enters F or V, so the skeleton and normal-score reference are independent of C and T; in the in-domain settings I1–I2, C and the safe part of T are a uniformly random partition of the safe examples outside F and V, hence exchangeable given F and V, as Theorem 6.1 requires. The primary configuration (D1, $n_{\text{cal}} = 100$, $\alpha = 0.05$) and the expected outcome were fixed before test results were computed; all other configurations are reported in the Appendix (seed-to-seed spread there reflects split, not dataset, variability).

Settings. In-distribution I1 (WGT) and I2 (BT); shifts D1/D2 (WGT vanilla↔adversarial), D3/D4 (WGT↔BT) and D5 (WGT→JBB). *Methods.* M0 (one guard), M1 (full LDA), M3 (uncalibrated naive product cascade), M4c (calibrated Trust-or-Escalate adaptation), M5 (dependency-blind cascade), M6 (static subset) and A (dependency-aware cascade); precise definitions, grids and operating-point rules of all methods, and a robust variant AB, are in Appendix C.10.

Table 2 gives the D1 operating point. All methods whose threshold is calibrated on C (M1, M6, M4c, M5, A, AB) were put through a descriptive screening rule (not a test of validity) over all 32 configurations (validity itself is Theorem 6.1's, and this test does not establish it): the mean target FPR over 20 seeds exceeds a pre-set stop line $\alpha + 3\sigma_\alpha/\sqrt{20}$, with σ_α the s.d. of the Beta law in Section 6.1, in 4 of 220 cells: AB at D1 ($n_{\text{cal}} = 200$, $\alpha = 0.01$; mean FPR 0.0171 vs. 0.0147), AB at D2 and at I1 (200, 0.05; 0.0609 and 0.0607 vs. 0.0603) and M5 at D2 (200, 0.01; 0.0155 vs. 0.0147). The stop line (Appendix Figure 7) is computed for $n = n_{\text{cal}}$ for every method; AB calibrates on $|C_2| = n_{\text{cal}}/2$ examples, and against the stop line for $n_{\text{cal}}/2$ (0.0166, 0.0644) the two AB cells at $\alpha = 0.05$ are below it, while D1 at $\alpha = 0.01$ still exceeds it by 0.0005 (at $m = 1$ the Beta law is right-skewed: a mean of 20 draws this high has probability 0.003, so one such cell among 220 is not unexpected) (220 cells: 32 configurations × 7 selected methods M1, M6:P-match, M4c:P-match, M5:P-match, A:P-match, A:P-cheap, AB:P-match, minus 4 configurations at $\alpha = 0.01$ where A has no P-cheap point; see Appendix). Seeds re-split one finite benchmark; resampling the benchmark items themselves (item bootstrap, below), the 95% interval of the mean FPR contains α in 45 of the 49 method–setting cells at $n_{\text{cal}} = 100$, $\alpha = 0.05$ and lies below it in the other four. Per seed, the fraction of seeds with target FPR above α is 0.46 on average over the 35 method–setting cells at $n_{\text{cal}} = 100$, $\alpha = 0.05$, against 0.44 predicted by the i.i.d. Beta law of Section 6.1 plus binomial test noise, and the largest per-seed FPR (0.083–0.138) is within the range expected for the maximum of 20 draws (0.10–0.17 at the 95th percentile) (Appendix Table 7) ; M3 exceeds the nominal level in 6 of 7 settings, reaching FPR 0.262 . In D1, the full LDA reaches detection 0.594 at the full cost, while the 0.6B single guard reaches 0.531 at 2.7% of that cost, and the static subset M6 reaches 0.569 at 8.1%; A reaches 0.570 at 19.9%, and M5 reaches 0.606 at 32.6% . (Retention figures such as "96–100%" are the ratio of mean T-detection of M6:P-match to that of M1 over the same 20 seeds, at $n_{\text{cal}} = 100$, $\alpha = 0.05$; over all 32 configurations the ratio is 0.71–1.05, lowest at $\alpha = 0.01$ where P-match is harder to meet. Full LDA is the comparator by design, not the best detector: in D1 Qwen3Guard-8B alone reaches 0.632 at 29% of the cost.) Across settings ("significant" below: paired t-test over the 20 seeds, $p < 0.05$, uncorrected; it reflects split variability only), A is significantly cheaper than M5 in all 32 configurations (29 of 32 also after a Bonferroni correction over the 32; the exceptions are the three D2 configurations), usually with a small detection decrease ; A and M6 have no consistent winner in either cost or detection (a descriptive statement, not a test of equivalence) (Figure 5; other settings in Figure 8).

Uncertainty over benchmark items. Seed standard errors reflect re-splitting one fixed benchmark: in D1 every seed tests on 697 of the same 797 target items. To also account for the sampling of the items, we resample the target items with replacement within each class (2000 replicates), reweight every seed's test metrics by the resampled counts and average over seeds, keeping the realised seed splits, selected operating points and calibrated thresholds fixed. The resulting intervals are *conditional* on the realised splits, the selected operating points and subsets, and the calibrated thresholds; they do not include selection or calibration uncertainty. We report the percentile interval

("item interval") and, as a sensitivity check, a normal variance-addition interval whose variance is the item variance plus the squared seed SE ("item+split"; a descriptive summary, not a guaranteed-conservative interval); all values are in Appendix Table 9. At $n_{\text{cal}} = 100$, $\alpha = 0.05$: (i) The retention ratio $\det(\text{M6})/\det(\text{M1})$ has item intervals with lower ends 0.91 (D1) to 0.98 (D3); M6 loses 0.001–0.031 detection to the full LDA, a loss distinguishable from zero in D4, I1 and I2 (item intervals; D4 and I2 also item+split). (ii) A is cheaper than M5 in all seven settings (Figure 5c), by 2.5–19.9 points of the full cost, with every item+split interval excluding zero; it detects 0.002–0.047 less, with item intervals excluding zero in six settings and item+split intervals in three (D1, D4, D5). (iii) No item+split interval for the detection difference between A and M6 excludes zero; in cost, A is cheaper in D2, D4 and I1 and more expensive in D1 and D5 (item+split), so the absence of a consistent winner is not an artefact of split noise. (iv) The single Qwen3Guard-8B (Figure 5b) detects 0.026–0.092 more than M6 in six settings (item intervals exclude zero; D2 $+0.016$ does not), at 17–21 more points of the full cost; M6 beats Qwen3Guard-0.6B alone by 0.020–0.095 in D1, D4, D5 and I1 and is within 0.008 of it elsewhere. (v) The fitted fusion of all seven guards does not beat a single guard: Qwen3Guard-8B alone (the largest guard; 29% of the cost) detects between 0.0003 less (D2) and 0.091 more (D5) than the calibrated full LDA (item intervals exclude zero in D3 and D5, in its favour; in I2 the lower end is $+0.0002$, borderline), and the single guard with the best source-validation AUROC, chosen per seed, is within its item+split interval of the full LDA in all seven settings. This concerns one fitted rule (LDA) under the stated target-calibration protocol, not the model optimum $\beta^\star$, but it is consistent with the model-based cap of Section 5.

Selecting subsets by H_S alone. M6 selects by validation detection on V. To test whether the model quantity can replace that search, we fixed two rules before computing their results (results of the other methods, including M6, were already known), using (d, R) estimated on F only and the model gap $g(S) = \beta^\star - \beta_S$ of Theorem 5.2: *H-match*, the cheapest S with $g(S) \le 0.02$, and *H-cheap*, the S with smallest $g(S)$ among subsets costing at most 10%. At $n_{\text{cal}} = 100$, $\alpha = 0.05$, H-match is not competitive: it costs 11–39% of the full ensemble (M6: 8–12%) without higher detection on average (D1 0.535 vs. 0.569, D5 0.617 vs. 0.625, I1 0.813 vs. 0.789). The reason is the one noted in Section 5: the model gap overstates the real gap (e.g. 0.12 vs. 0.05 for Qwen3Guard-0.6B on WGT), so an absolute threshold on it buys more guards than needed. H-cheap, which uses H_S only to *rank* affordable subsets, matches M6:P-match: the paired detection difference is between -0.013 and $+0.016$ and not significant in six settings, and $+0.056$ ($p = 0.04$) in D5, at equal or lower cost (Appendix Table 8). Resampling items gives the same picture with slightly different edges: the D5 gain has item interval $[0.026, 0.087]$, the small differences in D3, D4 (-0.013, -0.012) and I2 ($+0.009$) also exclude zero, and no item+split interval excludes zero. H-cheap beats the 0.6B guard alone by 0.054 in D1 and 0.151 in D5, the settings where that guard is weak. This is a limited test: under the 10% budget only five subsets are affordable, and H-cheap picks Qwen3Guard-0.6B with Llama-Guard-3-1B in 139 of the 140 setting–seed pairs (one D2 seed picks Granite-Guardian alone), so it shows that H_S ranks this small candidate set as well as validation detection does, not that it ranks diverse subsets robustly. Cost comparisons outside D1 are pooled at each target's prevalence.

				Cost (% full)		
Method	FPR	Detection	Det. SE	all	safe	unsafe
M1 (full LDA)	0.042	0.594	0.030	100.0	100.0	100.0
M0 (Qwen3Guard-0.6B)	0.048	0.531	0.031	2.7	2.7	2.7
M0 (Qwen3Guard-8B)	0.047	0.632	0.025	29.2	29.2	29.2
M6 (static subset)	0.052	0.569	0.031	8.1	8.1	8.1
M5 (dependency-blind cascade)	0.043	0.606	0.029	32.6	23.4	72.1
A (dependency-aware cascade)	0.041	0.570	0.033	19.9	14.2	44.4
M4c (Trust-or-Escalate)	0.047	0.631	0.025	100.0	100.0	100.0

Table 2: D1 target-test performance ($n_{\text{cal}} = 100$, $\alpha = 0.05$, 20 seeds). Cascade costs depend on the class: A spends 14.2% of the full cost on safe and 44.4% on unsafe inputs, because most safe inputs stop after the cheap guards; pooled cost therefore depends on prevalence (18.8% unsafe in the D1 target test set; $0.188 \cdot 44.4 + 0.812 \cdot 14.2 = 19.9$).

B.7 DETAILED LIMITATIONS

Our mechanisms are derived in a Gaussian-copula model. It matches the first two moments of the fused log-evidence to within 5%, but the observed false-alarm rate at the $1/\alpha$ threshold exceeds the Gaussian prediction (an upper-tail mismatch, so Theorem 4.1 understates the failure there) and on BeaverTails the unsafe-class variances differ from one, so the model's detection ceiling is below what three single guards achieve; we use the sandwich for the size of gaps and rankings, not as a certified bound. The sandwich check is also noisy on small benchmarks: on JBB (300 responses) the cross-fitted gap between the ensemble and the cheapest guard has a 95% range of $[-0.093, 0.396]$, so on JBB only the ranking (0.6B far from the ensemble, 4B at it) is supported, not a precise gap. Validity claims rest only on Theorem 6.1, which requires exchangeable target-domain safe examples; it does not protect against shifts after calibration, for which we suggest online auditing. Calibration needs only target safe examples, but choosing the subset or skeleton (by validation detection or by H_S) needs labelled safe and unsafe source data. The skeletons of the adaptive cascades are derived from the Gaussian model; the theorem keeps them valid when it is misspecified, but we have not tested skeletons built from nonparametric estimates, which might be more efficient where the model fits poorly (e.g. BeaverTails). Cost is parameter memory, with a per-token FLOPs check; we do not measure latency or account for prompt length, which differs across the guards' templates. Benchmarks differ in their definition of harm, which confounds cross-dataset shifts. We study guard scores only, not end-to-end agent behaviour, and ShieldGemma's scores depend on the policy text we supplied.

C ADDITIONAL RESULTS

Guard	Cost (%)	d_k^2	H_k	Gap/Bound	Emp.
Qwen3Guard-0.6B	2.7	2.95	2.55	.229/.250	0.500
LlamaGuard3-1B	5.4	3.46	2.04	.173/.194	0.673
Granite-Guardian-3.1-2B	9.0	3.79	1.70	.139/.158	0.755
ShieldGemma-2B	9.3	1.95	3.55	.357/.379	0.327
Qwen3Guard-4B	15.7	4.14	1.36	.106/.124	0.764
LlamaGuard3-8B	28.6	3.38	2.12	.182/.203	0.709
Qwen3Guard-8B	29.2	4.01	1.49	.118/.137	0.764

Table 3: Redundancy sandwich on JBB ($K = 7$, $\alpha = 0.05$, in-sample; 300 responses).

Setting	M3 FPR	A (P-match) FPR	Stop line	Exceeds?
D1	0.262 ± 0.011	0.041 ± 0.030	0.064	yes
I1	0.115 ± 0.017	0.052 ± 0.024	0.064	yes
D3	0.170 ± 0.007	0.050 ± 0.018	0.064	yes
D4	0.084 ± 0.003	0.055 ± 0.024	0.064	yes
D2	0.032 ± 0.003	0.048 ± 0.022	0.064	no
I2	0.150 ± 0.012	0.055 ± 0.026	0.064	yes
D5	0.259 ± 0.040	0.044 ± 0.030	0.064	yes

Table 4: FPR of the uncalibrated naive product cascade (M3) vs. the calibrated cascade A, nominal $\alpha = 0.05$, $n_{\text{cal}} = 100$, 20 seeds (mean $\pm$ SD; Figure 3a).

C.1 RESULTS ACROSS CONFIGURATIONS

Figure 7 summarises the false-alarm screening over all 32 configurations; Table 5 reports the paired comparisons. Pm and Pc denote P-match and P-cheap; paired differences are detection-rate differences followed by cost differences in GB. A star marks a paired t-test over the 20 seeds with $p < 0.05$, uncorrected for multiplicity; stars are exploratory and reflect split variability only.

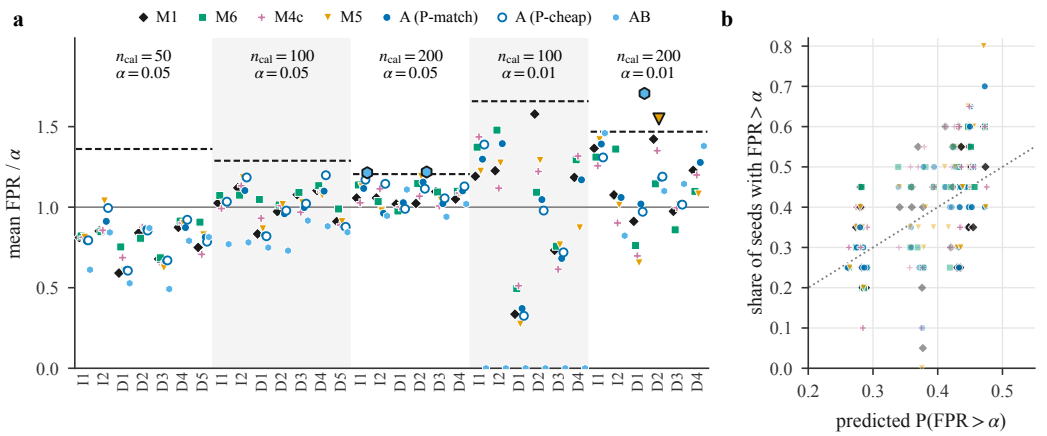


Figure 7: False-alarm screening over all configurations. **(a)** Mean target FPR over 20 seeds, divided by α, for the seven calibrated selections in each setting; dashed: the pre-set stop line $\alpha + 3\sigma_\alpha/\sqrt{20}$ (computed for $n = n_{\text{cal}}$), divided by α; outlined markers: the 4 of 220 cells above it (Section 6.3). AB never alarms at $n_{\text{cal}} = 100$, $\alpha = 0.01$ because it calibrates on $n_{\text{cal}}/2 < 1/\alpha - 1$ examples. **(b)** Share of seeds with target FPR $> \alpha$ against the prediction of the i.i.d. Beta law plus binomial test noise, for every method–setting–configuration cell (faded: $\alpha = 0.01$); dotted: equality.

C.2 Per-seed false-alarm rates and H_S-only selection

C.3 Item-level uncertainty of the main comparisons

Per-item decisions were regenerated for all seven settings at $n_{\text{cal}} = 100$, $\alpha = 0.05$ and checked against the stored results (every metric of all 4440 records per setting agrees to 10^{-12}). The item bootstrap resamples the union of target test items with replacement within each class (2000 replicates, seed 0); each seed's metrics are recomputed with the resampled counts restricted to that seed's test set, and averaged over the 20 seeds. Operating points and calibrated thresholds are held fixed, so the item interval does not include calibration or selection noise (the intervals are conditional on the realised seed splits and the fitted pipeline); the item+split column adds the seed variance (a normal variance-addition sensitivity interval, not a guaranteed-conservative one, $\pm 1.96\sqrt{\text{sd}^2_{\text{item}} + \text{SE}^2_{\text{seed}}}$). Cost differences are in points of the full cost; static-subset costs are fixed per item, so their item intervals are degenerate and are omitted.

C.4 Compute-based cost

Table 11 counts parameters from the released weight files (safetensors headers). As a FLOPs proxy we use the non-embedding parameters (total minus input embedding minus any stored output head), since each guard reads one verdict logit. ShieldGemma-2B loses most (its 256k-entry vocabulary makes 23% of its parameters an embedding table), and the 0.6B guard's share drops from 2.7% to 1.9%. Table 12 re-prices the selected operating points of Section 6.3 from the recorded calls, without re-selecting them. Token counts are not included: the guards' prompt templates differ in length (ShieldGemma's includes a policy text), which would reweight guards per input.

C.5 Cost–detection planes for the remaining settings

C.6 Eight-guard robustness

Adding WildGuard gives $K = 8$; all quantities are re-estimated with eight guards. This check uses separate diagnostics (held-out stacking AUROC, naive-Bayes log-odds inflation against a joint logistic model, product FPR at $e \geq 20$), not the null-moment statistics of Section 4. The stacking

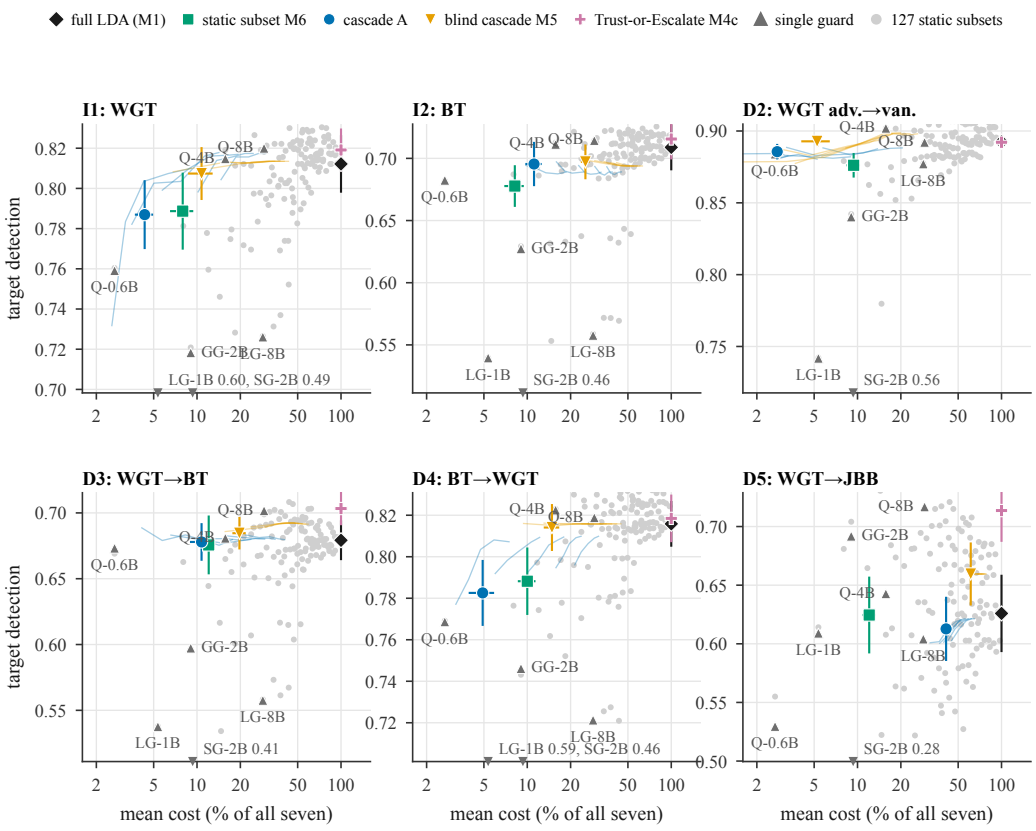


Figure 8: Target cost–detection planes for the six settings not shown in Figure 5a ($n_{\text{cal}} = 100$, $\alpha = 0.05$; selected operating points are means ± 1 SE over 20 seeds). Single guards below the plotted range are marked by a caret on the axis floor with their detection.

AUROC gain is at most 0.002; the mean absolute log-odds claimed by naive Bayes on held-out items is 5.2–6.3 times that of a jointly fitted logistic model. At threshold 20, the product rule has false-positive rates of 0.078–0.115 across the three datasets. The unsafe all-wrong rates are 0.116 on WGT and 0.125 on BT at the native operating point.

C.7 Baseline adaptations

AB splits its calibration set in half; its observed detection rate is no higher than A's in any setting (D1: -0.055); halving the calibration set changes both quantile resolution and threshold variability and need not lower power in general. The adapted M4 fixed-sequence procedure permits almost no early stops when ε is small, so its cost approaches a full call. Uncalibrated M4 misses the P-match requirement in 15 of 20 seeds. M4c has no feasible P-cheap operating point.

C.8 Common-failure slices

C.9 Uncertainty and cross-fitting

Cross-fitted sandwich estimates, bootstrap percentile intervals, and prevalence summaries are reported below. The paired tests in Table 5 are not corrected for multiple comparisons; the primary configuration (D1, $n_{\text{cal}} = 100$, $\alpha = 0.05$) was fixed in advance and all other configurations are reported in full.

C.10 METHOD SPECIFICATIONS

All objects below are fitted on the source split F; z are normal scores with reference the safe part of F; c_k is the cost of guard k (GB, $\sum_k c_k = 56.06$); d is as in Section 2, and R_F denotes the pooled class-centred correlation of Section 2 (R_{pool} in Appendix B.3) estimated on F (not the safe-class R_0 of Section 4).

Fixed-cost rules and calibration. M0 uses one guard's raw score; M1 the LDA score $w^\top z$ on all seven guards (w from the pooled class-centred covariance plus $10^{-3}I$); M2 the mean of z; M6 the LDA score on a subset S (all 127 subsets are fitted). Every calibrated rule, including the cascades below, uses the same threshold: the $\lceil (n+1)(1-\alpha) \rceil$-th smallest calibration statistic on C (after jitter), alarming when the jittered test statistic is strictly larger. Validation detection, used only for selection, is computed in the same way with the safe part of V in place of C.

Cascade A (dependency-aware). In $\mathcal{M}(d, R)$ with parameters (d, R_F) the cascade tracks the running log-odds $\ell = \text{logit}(p_0) + \sum_t (\eta_t u_t - \frac{1}{2}\eta_t^2)$, where for the t-th called guard k given the called set S, $u_t = (z_k - \gamma^\top z_S)/\sqrt{v}$, $\eta_t = (d_k - \gamma^\top d_S)/\sqrt{v}$, $\gamma = R_{SS}^{-1} R_{Sk}$, $v = 1 - R_{kS}\gamma$ (so $\eta_t^2 = I(S \cup \{k\}) - I(S)$), and p_0 is the unsafe prevalence in F. Which guard to call next and when to stop are obtained by dynamic programming over (called set, ℓ) on a 401-point grid of $\ell \in [-12, 12]$, with stopping loss $\min(\lambda_r p, 1-p)$ ($p = \text{expit}(\ell)$, λ_r the miss-to-false-alarm cost ratio), calling cost $\omega c_k / \sum_j c_j$ and 32-node Gauss–Hermite integration over the next innovation. This yields the calling rule $\kappa(S, \ell)$ (the minimiser; if it is undefined off-grid, the remaining guard with the largest η^2/c_k) and the continuation interval (a_S, b_S) of each called set, i.e. the skeleton of Section 6.1. The calibration statistic is $\theta^*(x) = \max_t(\ell_t - b_{S_t})$ along the never-alarm path, computed once per input; the calibrated threshold then defines both the decision and the realised cost. Grid: $\lambda_r \in \{0.5, 1, 2, 4, 8\}$, $\omega \in \{0.01, 0.03, 0.1, 0.3, 1\}$ (25 skeletons).

M5 (dependency-blind) and AB (robust). M5 is identical to A with $R = I$, i.e. it treats the guards as conditionally independent both in ℓ and in the dynamic program; the difference A−M5 therefore isolates the use of the dependence structure. AB uses $R^\star = \lambda R_F + (1-\lambda) R_C$ with $R_C = \frac{1}{2}\widehat{\text{corr}}(C_1) + \frac{1}{2}R_F$ estimated on the first half C_1 of the calibration set and λ on a 101-point grid minimising $d^\top (R^\star)^{-1} d$ (a least-favourable choice); it is calibrated on the second half C_2.

M3 (practitioner's naive cascade, uncalibrated). Guards are called in decreasing order of d_k^2/c_k; after each call $\ell \mathrel{+}= d_k z_k - \frac{1}{2}d_k^2$ (the naive product of Gaussian marginal likelihood ratios); stop with an alarm if $\ell \geq \log(1/\alpha)$, stop safe if $\ell \leq \log \alpha$; after all guards alarm iff $\ell \geq 0$.

M4 and M4c (Trust-or-Escalate adaptation). Guards are called in increasing order of cost. Each guard's raw score is Platt-scaled on F to p_k. For a tolerance $\varepsilon \in \{0.02, 0.05, 0.1, 0.2, 0.3\}$, each guard gets a confidence level λ_k by fixed-sequence testing on V: the smallest λ on a 0.001 grid (scanned downward from 1) before the one-sided 90% Clopper–Pearson upper bound on the error rate among V items with $\max(p_k, 1-p_k) \geq \lambda$ first exceeds ε. An input stops at the first guard whose confidence reaches λ_k (otherwise at the last guard). M4 alarms iff the stopping guard's $p_k \geq 0.5$ (no abstention, no calibration on C); M4c thresholds the stopping guard's p_k with the conformal threshold on C.

Operating points. For each seed and family (A, M5, AB, M4, M4c, M6), P-match is the configuration with the smallest mean V-cost among those whose V-detection is at least that of M1 minus 0.02 (if none qualifies, the highest V-detection, reported as infeasible); P-cheap is the configuration with the highest V-detection among those with mean V-cost at most 10% of the full cost (none for M4c). Costs are expected per-input costs; for cascades they depend on the class mix, and the per-class costs are stored with every result.

Setting	n	Alpha	A det	A cost (%)	A–M5 Pm	A–M6 Pm	A–M6 Pc	M6 det/cost	0.6B det	M1 det
D1	50	0.05	0.472	20.1%	-0.027 / -7.3*	+0.021 / +6.7*	+0.057 / +9.7*	0.450/8.1%	0.425	0.490
D1	100	0.05	0.570	19.9%	-0.036* / -7.1*	+0.000 / +6.6*	+0.014 / +9.5*	0.569/8.1%	0.531	0.594
D1	200	0.05	0.647	18.9%	-0.013 / -8.1*	+0.090* / +6.1*	+0.075* / +9.1*	0.556/8.1%	0.535	0.663
D1	100	0.01	0.187	22.2%	-0.033* / -6.7*	+0.019 / +5.3*	+0.041 / +9.6*	0.168/12.9%	0.152	0.203
D1	200	0.01	0.282	22.1%	-0.004 / -7.6*	+0.054* / +5.2*	+0.090* / +9.6*	0.228/12.9%	0.186	0.291
D2	50	0.05	0.858	3.0%	-0.004 / -2.0*	+0.016 / -3.6*	+0.016* / -2.7*	0.842/9.4%	0.854	0.874
D2	100	0.05	0.886	2.8%	-0.007* / -1.4*	+0.009 / -3.7*	+0.010* / -2.8*	0.876/9.4%	0.884	0.892
D2	200	0.05	0.895	2.7%	-0.003 / -1.1*	+0.014 / -3.7*	+0.010* / -2.8*	0.881/9.4%	0.896	0.895
D2	100	0.01	0.715	5.6%	-0.035* / -5.4*	+0.011 / -4.3*	+0.062 / -1.3*	0.704/13.3%	0.688	0.776
D2	200	0.01	0.747	4.9%	-0.036* / -5.4*	+0.008 / -4.7*	+0.029 / -1.7*	0.739/13.3%	0.728	0.795
D3	50	0.05	0.577	17.8%	-0.003 / -7.0*	+0.006 / +3.2*	+0.050 / +8.4*	0.571/12.1%	0.570	0.576
D3	100	0.05	0.678	10.8%	-0.007 / -5.0*	+0.002 / -0.7	+0.019 / +4.0*	0.676/12.1%	0.673	0.679
D3	200	0.05	0.694	9.5%	-0.003 / -5.2*	-0.005 / -1.4*	+0.014 / +3.4*	0.699/12.1%	0.694	0.691
D3	100	0.01	0.296	32.0%	-0.023* / -9.0*	+0.059 / +9.7*	+0.038 / +15.0*	0.237/14.8%	0.324	0.305
D3	200	0.01	0.370	28.9%	-0.002 / -11.1*	+0.115* / +7.9*	+0.101* / +13.6*	0.254/14.8%	0.368	0.359
D4	50	0.05	0.730	6.5%	-0.027* / -5.7*	+0.003 / -2.0*	+0.004 / -1.6*	0.727/10.0%	0.706	0.760

Table 5: Paired test-set differences across configurations (part 1/2).

Setting	n	Alpha	A det	A cost (%)	A–M5 Pm	A–M6 Pm	A–M6 Pc	M6 det/cost	0.6B det	M1 det
D4	100	0.05	0.783	4.9%	-0.032* / -5.6*	-0.006 / -2.9*	-0.001 / -2.1*	0.788/10.0%	0.769	0.816
D4	200	0.05	0.794	4.4%	-0.029* / -5.8*	+0.005 / -3.1*	+0.002 / -2.4*	0.789/10.0%	0.779	0.815
D4	100	0.01	0.511	10.2%	-0.019 / -9.8*	+0.032 / +2.8*	n/a	0.479/5.2%	0.482	0.562
D4	200	0.01	0.531	10.3%	-0.031* / -9.9*	+0.067* / +2.9*	n/a	0.464/5.2%	0.485	0.577
I1	50	0.05	0.717	5.7%	-0.041* / -4.2*	+0.004 / -1.3	+0.018 / +1.6*	0.712/8.0%	0.692	0.759
I1	100	0.05	0.787	4.3%	-0.020 / -3.6*	-0.002 / -2.0*	+0.025 / +0.7	0.789/8.0%	0.759	0.812
I1	200	0.05	0.800	4.2%	-0.024* / -3.5*	+0.003 / -2.1*	+0.022* / +0.4	0.797/8.0%	0.787	0.824
I1	100	0.01	0.546	9.4%	-0.033 / -6.4*	+0.039 / -0.9	+0.086* / +0.9	0.507/11.1%	0.506	0.565
I1	200	0.01	0.558	9.7%	-0.047* / -7.2*	+0.035 / -0.8	+0.082* / +1.0	0.523/11.1%	0.517	0.588
I2	50	0.05	0.647	14.2%	-0.007 / -8.3*	+0.033* / +3.4*	+0.039* / +4.0*	0.614/8.2%	0.615	0.658
I2	100	0.05	0.695	11.1%	-0.002 / -8.0*	+0.018 / +1.7	+0.020 / +2.3	0.678/8.2%	0.682	0.709
I2	200	0.05	0.700	11.1%	+0.007 / -9.3*	+0.005 / +1.6	+0.007 / +1.6*	0.695/8.2%	0.697	0.724
I2	100	0.01	0.401	29.5%	-0.002 / -12.3*	+0.030 / +13.4*	n/a	0.372/5.5%	0.359	0.391
I2	200	0.01	0.390	32.3%	+0.037* / -13.8*	-0.002 / +15.0*	n/a	0.392/5.5%	0.390	0.374
D5	50	0.05	0.514	39.8%	-0.067* / -7.6*	-0.029 / +15.5*	-0.042 / +20.3*	0.542/12.1%	0.437	0.519
D5	100	0.05	0.613	41.2%	-0.047* / -11.2*	-0.012 / +16.3*	-0.013 / +21.0*	0.625/12.1%	0.530	0.626

Table 6: Paired test-set differences across configurations (part 2/2).

	M1			M6:P-match			A:P-match		
Setting	obs.	theory	max	obs.	theory	max	obs.	theory	max
D1	0.25	0.43	0.122	0.50	0.43	0.138	0.30	0.43	0.127
D2	0.35	0.45	0.106	0.40	0.45	0.127	0.40	0.45	0.121
D3	0.55	0.44	0.104	0.45	0.44	0.104	0.50	0.43	0.088
D4	0.60	0.43	0.110	0.50	0.43	0.126	0.40	0.43	0.115
D5	0.25	0.43	0.122	0.45	0.43	0.111	0.25	0.43	0.100
I1	0.55	0.44	0.111	0.50	0.43	0.098	0.45	0.44	0.102
I2	0.60	0.45	0.108	0.60	0.45	0.096	0.60	0.45	0.118

	M5:P-match			M4c:P-match		
Setting	obs.	theory	max	obs.	theory	max
D1	0.30	0.43	0.118	0.55	0.43	0.094
D2	0.55	0.45	0.123	0.55	0.45	0.099
D3	0.55	0.43	0.083	0.35	0.44	0.107
D4	0.45	0.43	0.112	0.60	0.43	0.094
D5	0.30	0.44	0.100	0.35	0.43	0.122
I1	0.45	0.44	0.098	0.50	0.44	0.083
I2	0.65	0.45	0.137	0.65	0.45	0.094

Table 7: Per-seed target FPR at $n_{\text{cal}} = 100$, $\alpha = 0.05$ (20 seeds per cell). "obs.": fraction of seeds with FPR $> \alpha$; "theory": the same fraction under the i.i.d. Beta law of Section 6.1 plus binomial noise of the test set; "max": largest per-seed FPR (Section 6.3).

	H-match		M6:P-match		H-cheap		M6:P-cheap	
Setting	det	cost%	det	cost%	det	cost%	det	cost%
D1	0.535	30.3	0.569	8.1	0.585	8.0	0.557	4.0
D2	0.883	39.2	0.876	9.4	0.877	8.1	0.878	7.8
D3	0.680	38.5	0.676	12.1	0.663	8.0	0.659	5.7
D4	0.784	10.7	0.788	10.0	0.776	8.0	0.775	6.5
D5	0.617	38.5	0.625	12.1	0.681	8.0	0.625	5.7
I1	0.813	37.6	0.789	8.0	0.781	8.0	0.773	4.8
I2	0.693	14.6	0.678	8.2	0.687	8.0	0.670	6.5

Table 8: Selecting subsets by the model gap $g(S)$ alone ($n_{\mathrm{cal}} = 100$, $\alpha = 0.05$, 20 seeds; (d, R) from F). H-match: cheapest S with $g(S) \leq 0.02$; H-cheap: smallest $g(S)$ at cost $\leq 10\%$; compared with the validation-selected M6 at the same operating-point rule. Cost in % of the full ensemble, pooled at the target prevalence (Section 6.3).

Setting	Point	Item 95% interval	Seed SE	Item+split 95%
M6:P-match / M1, det ratio				
D1	0.959	[0.910, 1.015]	–	–
D2	0.982	[0.960, 1.004]	–	–
D3	0.995	[0.984, 1.006]	–	–
D4	0.966	[0.951, 0.980]	–	–
D5	0.998	[0.939, 1.066]	–	–
I1	0.971	[0.950, 0.990]	–	–
I2	0.956	[0.948, 0.964]	–	–
M6:P-match − M1, det				
D1	-0.024	[-0.056, 0.008]	0.022	[-0.078, 0.029]
D2	-0.016	[-0.037, 0.003]	0.007	[-0.040, 0.007]
D3	-0.004	[-0.011, 0.004]	0.014	[-0.032, 0.025]
D4	-0.028	[-0.040, -0.016]	0.008	[-0.048, -0.008]
D5	-0.001	[-0.040, 0.039]	0.030	[-0.072, 0.069]
I1	-0.023	[-0.041, -0.008]	0.014	[-0.055, 0.008]
I2	-0.031	[-0.037, -0.025]	0.014	[-0.059, -0.004]
A:P-match − M6:P-match, det				
D1	0.000	[-0.032, 0.033]	0.022	[-0.054, 0.055]
D2	0.009	[-0.000, 0.020]	0.009	[-0.010, 0.029]
D3	0.002	[-0.003, 0.007]	0.015	[-0.027, 0.031]
D4	-0.006	[-0.012, -0.000]	0.009	[-0.024, 0.012]
D5	-0.012	[-0.052, 0.028]	0.028	[-0.080, 0.056]
I1	-0.002	[-0.011, 0.007]	0.011	[-0.025, 0.021]
I2	0.018	[0.013, 0.023]	0.013	[-0.007, 0.043]
A:P-match − M6:P-match, cost (pts)				
D1	11.8	[10.3, 13.3]	1.6	[8.2, 15.3]
D2	-6.6	[-6.6, -6.6]	1.0	[-8.6, -4.7]
D3	-1.3	[-1.7, -0.9]	1.7	[-4.7, 2.1]
D4	-5.1	[-5.3, -4.9]	1.4	[-7.9, -2.3]
D5	29.2	[25.7, 32.8]	1.9	[24.0, 34.3]
I1	-3.6	[-3.8, -3.4]	1.2	[-6.0, -1.3]
I2	2.9	[2.3, 3.6]	2.0	[-1.0, 6.9]
A:P-match − M5:P-match, det				
D1	-0.036	[-0.057, -0.016]	0.014	[-0.070, -0.002]
D2	-0.007	[-0.015, -0.001]	0.002	[-0.016, 0.001]
D3	-0.007	[-0.010, -0.003]	0.005	[-0.018, 0.004]
D4	-0.032	[-0.047, -0.017]	0.010	[-0.057, -0.006]
D5	-0.047	[-0.066, -0.030]	0.006	[-0.069, -0.025]
I1	-0.020	[-0.035, -0.007]	0.013	[-0.049, 0.008]
I2	-0.002	[-0.008, 0.004]	0.010	[-0.021, 0.018]

Table 9: Item-bootstrap uncertainty of the main contrasts ($n_{\mathrm{cal}} = 100$, $\alpha = 0.05$, 20 seeds). "Point": mean over seeds of the per-seed difference (for the ratio: ratio of seed means). "Seed SE": s.d. over seeds$/\sqrt{20}$.

Setting	Point	Item 95% interval	Seed SE	Item+split 95%
A:P-match − M5:P-match, cost (pts)				
D1	-12.7	[-13.8, -11.7]	1.4	[-15.6, -9.8]
D2	-2.5	[-2.7, -2.3]	1.1	[-4.7, -0.2]
D3	-8.9	[-9.4, -8.6]	1.7	[-12.3, -5.6]
D4	-9.9	[-10.7, -9.2]	2.3	[-14.5, -5.4]
D5	-19.9	[-22.9, -17.2]	1.4	[-23.8, -16.0]
I1	-6.4	[-6.9, -5.9]	1.7	[-9.8, -3.1]
I2	-14.2	[-14.8, -13.6]	1.5	[-17.2, -11.1]
H-cheap − M6:P-match, det				
D1	0.016	[-0.024, 0.058]	0.020	[-0.041, 0.073]
D2	0.001	[-0.009, 0.010]	0.006	[-0.015, 0.016]
D3	-0.013	[-0.020, -0.007]	0.019	[-0.051, 0.025]
D4	-0.012	[-0.018, -0.006]	0.007	[-0.027, 0.004]
D5	0.056	[0.026, 0.087]	0.026	[-0.003, 0.116]
I1	-0.007	[-0.028, 0.012]	0.013	[-0.039, 0.025]
I2	0.009	[0.006, 0.013]	0.011	[-0.013, 0.032]
Qwen3Guard-8B − M6:P-match, det				
D1	0.063	[0.032, 0.097]	0.019	[0.013, 0.113]
D2	0.016	[-0.004, 0.035]	0.007	[-0.008, 0.040]
D3	0.026	[0.019, 0.033]	0.013	[-0.001, 0.053]
D4	0.031	[0.014, 0.048]	0.009	[0.006, 0.055]
D5	0.092	[0.041, 0.141]	0.028	[0.018, 0.167]
I1	0.031	[0.012, 0.051]	0.013	[-0.000, 0.063]
I2	0.037	[0.029, 0.045]	0.011	[0.014, 0.060]
M6:P-match − Qwen3Guard-0.6B, det				
D1	0.039	[0.019, 0.058]	0.014	[0.004, 0.073]
D2	-0.008	[-0.021, 0.005]	0.010	[-0.030, 0.015]
D3	0.002	[-0.002, 0.008]	0.015	[-0.027, 0.032]
D4	0.020	[0.013, 0.026]	0.007	[0.004, 0.035]
D5	0.095	[0.054, 0.135]	0.029	[0.025, 0.165]
I1	0.030	[0.019, 0.041]	0.012	[0.004, 0.055]
I2	-0.005	[-0.008, -0.001]	0.008	[-0.020, 0.011]
Qwen3Guard-8B − M1, det				
D1	0.039	[-0.006, 0.084]	0.024	[-0.026, 0.103]
D2	-0.000	[-0.009, 0.010]	0.003	[-0.011, 0.011]
D3	0.022	[0.013, 0.032]	0.013	[-0.004, 0.049]
D4	0.003	[-0.006, 0.013]	0.005	[-0.011, 0.017]
D5	0.091	[0.032, 0.150]	0.023	[0.017, 0.165]
I1	0.008	[-0.014, 0.029]	0.010	[-0.021, 0.037]
I2	0.006	[0.000, 0.011]	0.006	[-0.008, 0.019]
Best single guard (V-AUROC) − M1, det				
D1	-0.007	[-0.038, 0.027]	0.027	[-0.069, 0.055]
D2	0.000	[-0.009, 0.010]	0.004	[-0.012, 0.012]
D3	0.015	[0.006, 0.024]	0.015	[-0.015, 0.045]
D4	-0.015	[-0.023, -0.007]	0.008	[-0.032, 0.002]
D5	0.064	[0.010, 0.119]	0.027	[-0.011, 0.140]
I1	-0.009	[-0.027, 0.007]	0.011	[-0.036, 0.018]
I2	0.004	[-0.001, 0.009]	0.007	[-0.010, 0.018]

Table 10: Table 9, continued.

Guard	Params (B)	Input emb. (B)	Non-emb. (B)	Bytes %	FLOPs %
Qwen3Guard-0.6B	0.75	0.16	0.44	2.7	1.9
Llama-Guard-3-1B	1.50	0.26	0.97	5.4	4.2
ShieldGemma-2B	2.61	0.59	2.02	9.3	8.6
Granite-Guardian-3.1-2B	2.53	0.10	2.43	9.0	10.4
Qwen3Guard-4B	4.41	0.39	3.63	15.7	15.5
Llama-Guard-3-8B	8.03	0.53	6.98	28.6	29.8
Qwen3Guard-8B	8.19	0.62	6.95	29.2	29.6

Table 11: Parameter counts (billions) and each guard's share of the seven-guard total under the byte and the FLOPs price.

	M6:P-match		H-cheap		A:P-match		M5:P-match		Q8B	
Setting	bytes	FLOPs	bytes	FLOPs	bytes	FLOPs	bytes	FLOPs	bytes	FLOPs
D1	8.1	7.4	8.0	6.0	19.9	18.9	32.6	31.8	29.2	29.6
D2	9.4	8.2	8.1	6.3	2.8	2.0	5.2	4.4	29.2	29.6
D3	12.1	11.3	8.0	6.0	10.8	9.6	19.7	18.7	29.2	29.6
D4	10.0	9.4	8.0	6.0	4.9	4.1	14.8	14.0	29.2	29.6
D5	12.1	11.3	8.0	6.0	41.2	40.3	61.1	60.6	29.2	29.6
I1	8.0	6.9	8.0	6.0	4.3	3.4	10.8	9.9	29.2	29.6
I2	8.2	7.3	8.0	6.0	11.1	10.4	25.3	24.4	29.2	29.6

Table 12: Pooled cost (% of the full ensemble) of selected operating points under the byte price used in the paper and the non-embedding-parameter FLOPs proxy ($n_{\text{cal}} = 100$, $\alpha = 0.05$, 20 seeds; operating points selected under the byte price). Q8B: Qwen3Guard-8B alone.

Dataset	Slice	n	All wrong	Independence
JBB	Unsafe	110	0.000	0.0000
JBB	Safe	190	0.032	0.0000
WGT	Unsafe	284	0.120	0.0003
WGT	Safe	1425	0.001	0.0000
WGT	Unsafe, agreement 2 of 3	128	0.211	0.0051
BT	Unsafe	1733	0.133	0.0002
BT	Safe	1288	0.005	0.0000
BT	Unsafe, politics	95	0.632	0.1935

Table 13: Native-operating-point all-wrong rates for seven guards; independence values use marginal error rates within each slice.

Dataset	Guard	Model gap	LDA−single (95% range)	H_k mean
WGT	0.6B	0.133	0.048 [0.002, 0.104]	2.354
WGT	4B	0.100	-0.013 [-0.049, 0.027]	1.892
WGT	Q8B	0.093	0.006 [-0.027, 0.041]	1.779
BT	0.6B	0.044	0.023 [-0.002, 0.046]	0.487
BT	4B	0.090	0.004 [-0.015, 0.038]	0.937
BT	Q8B	0.091	-0.003 [-0.023, 0.019]	0.949
JBB	0.6B	0.230	0.170 [-0.093, 0.396]	2.522
JBB	4B	0.124	-0.005 [-0.273, 0.178]	1.533
JBB	Q8B	0.133	-0.060 [-0.327, 0.133]	1.622

Table 14: Cross-fitted redundancy sandwich (50 stratified half splits, seeds 0–49). (d, R) and the LDA direction are estimated on half A; detection rates are measured on half B, with thresholds at the 95% quantile of B's safe scores. Mean and 2.5/97.5 percentiles over splits.

Dataset	Quantity	Point estimate [2.5, 97.5]
WGT	Qwen 0.6B–4B R	0.847 [0.831, 0.862]
WGT	Qwen 0.6B–8B R	0.832 [0.812, 0.846]
WGT	Qwen 4B–8B R	0.925 [0.915, 0.933]
WGT	cross-fam. min: LG1B–Shield R	0.176 [0.132, 0.220]
WGT	cross-fam. max: 0.6B–Granite R	0.679 [0.654, 0.704]
WGT	Unsafe all wrong	0.120 [0.085, 0.158]
WGT	Unsafe all wrong (indep.)	$2.7e-04$ [$8.8e-05$, $7.0e-04$]
WGT	LDA−single	0.056 [0.007, 0.109]
WGT	$d^2_{0.6B}$	5.293 [4.706, 5.938]
WGT	$H_{0.6B}$	2.231 [1.811, 3.001]
WGT	d^2_{4B}	5.767 [5.093, 6.642]
WGT	H_{4B}	1.757 [1.345, 2.367]
WGT	d^2_{Q8B}	5.901 [5.267, 6.775]
WGT	H_{Q8B}	1.623 [1.196, 2.238]
BT	Qwen 0.6B–4B R	0.875 [0.855, 0.887]
BT	Qwen 0.6B–8B R	0.877 [0.856, 0.889]
BT	Qwen 4B–8B R	0.961 [0.951, 0.964]
BT	cross-fam. min: LG1B–Shield R	0.432 [0.398, 0.468]
BT	cross-fam. max: 0.6B–Granite R	0.747 [0.726, 0.764]
BT	Unsafe all wrong	0.133 [0.117, 0.149]
BT	Unsafe all wrong (indep.)	$2.3e-04$ [$1.5e-04$, $3.5e-04$]
BT	LDA−single	0.025 [$5.6e-04$, 0.060]
BT	$d^2_{0.6B}$	4.018 [3.625, 4.567]
BT	$H_{0.6B}$	0.437 [0.255, 0.792]
BT	d^2_{4B}	3.455 [3.068, 4.051]
BT	H_{4B}	1.000 [0.674, 1.430]
BT	d^2_{Q8B}	3.425 [3.016, 4.029]
BT	H_{Q8B}	1.029 [0.734, 1.416]
JBB	Qwen 0.6B–4B R	0.845 [0.799, 0.876]
JBB	Qwen 0.6B–8B R	0.870 [0.825, 0.894]
JBB	Qwen 4B–8B R	0.936 [0.910, 0.950]
JBB	cross-fam. min: LG1B–Shield R	0.410 [0.310, 0.495]
JBB	cross-fam. max: 0.6B–Granite R	0.754 [0.688, 0.804]
JBB	Unsafe all wrong	0.000 [0.000, 0.000]
JBB	Unsafe all wrong (indep.)	0.000 [0.000, 0.000]
JBB	LDA−single	0.218 [-0.055, 0.555]
JBB	$d^2_{0.6B}$	2.951 [2.326, 4.185]
JBB	$H_{0.6B}$	2.547 [1.779, 4.274]
JBB	d^2_{4B}	4.138 [3.393, 5.355]
JBB	H_{4B}	1.360 [0.983, 2.688]
JBB	d^2_{Q8B}	4.010 [3.160, 5.393]
JBB	H_{Q8B}	1.488 [1.094, 2.692]

Table 15: Stratified bootstrap (1000 replicates, seed 0): point estimate and 95% percentile interval. All-wrong rates at the native operating point (score > 0). The bootstrap interval is degenerate when no event is observed; for JBB (0 of 110) the exact Clopper–Pearson 95% upper bound is 0.033.

Dataset	n	Unsafe proportion
WGT	1709	0.166
BT	3021	0.574
JBB	300	0.367
WGT adv. test (D1)	697	0.188
Method	D1 test precision (WGT adversarial)	
M1	0.793	
A P-match	0.788	
A P-cheap	0.790	
M5 P-match	0.786	
M6 P-match	0.733	
M4c P-match	0.774	

Table 16: Dataset sizes and unsafe proportions; precision on the D1 test set (WGT adversarial minus calibration), averaged over 20 seeds.

Dataset	Rule	det (a)	det diff (a) [95%]	det (b)	FPR (b)
WGT	LDA	0.821	0.008 [-0.046, 0.056]	0.836	0.055
WGT	Logistic	0.817	0.004 [-0.048, 0.052]	0.835	0.055
WGT	HGB	0.772	-0.040 [-0.102, 0.007]	0.858	0.091
WGT	Best single guard	0.813	0.000 [0.000, 0.000]	0.822	0.054
BT	LDA	0.718	0.007 [-0.023, 0.045]	0.723	0.052
BT	Logistic	0.717	0.006 [-0.022, 0.043]	0.724	0.053
BT	HGB	0.693	-0.019 [-0.057, 0.017]	0.817	0.133
BT	Best single guard	0.711	0.000 [0.000, 0.000]	0.715	0.051
JBB	LDA	0.697	0.027 [-0.188, 0.255]	0.740	0.060
JBB	Logistic	0.696	0.026 [-0.182, 0.309]	0.739	0.059
JBB	HGB	0.645	-0.025 [-0.329, 0.273]	0.940	0.136
JBB	Best single guard	0.670	0.000 [0.000, 0.000]	0.737	0.058

Table 17: Cross-fitted fusion rules (50 stratified halvings). (a): threshold at the B-half 95% safe quantile (matched empirical FPR); difference to the best single guard chosen on half A, with 2.5/97.5 percentiles. (b): threshold at the A-half 95% safe quantile, evaluated on B. For boosted trees the A-half threshold is set on in-sample fitted scores, so its (b) FPR is inflated and (b) is not comparable.