Supplementary Information I

Mathematical Derivations, Proofs, and Audit Identities

SAGE-QEC: Structure-Aware Gated Error Correction for Surface-Code Memories

---

Prepared from the executed manuscript and its reported paired campaign

Asia Al-Hammadi and Mohammed Al-Absi · RAD Technology

Scope statement. This supplement expands the mathematical definitions, conditional guarantees, and accounting identities already stated in the manuscript. It does not introduce new empirical measurements. Numerical illustrations that appear here are explicitly derived from the reported tables.

1 Reading guide

This document is organized from the statistical endpoint to the decoder policy, then to the selective-risk interpretation and the reproducibility boundary. Each proposition is conditional on the stated assumptions. The distinction between an exact finite-sample identity, a probability approximation, and an empirical observation is maintained throughout.

2 Notation at a glance

Symbol	Meaning	Role in the audit
s	Detector record	The only shot-level input available to routing
ℓ	Logical observable outcome	Evaluator quantity used after inference
f_{PM}	Nominal PyMatching decoder	Default decision
f_{S}	SAGE-QEC decoder	Selective policy under audit
$u(s)$	Intervention indicator	Whether the default is replaced
N	Number of paired shots	Denominator of empirical rates

3 1. Statistical endpoint and normalization

3.1 1.1 Shot-level error indicator

Let a paired campaign contain detector records $s_i \in \{0,1\}^{\{n_d\}}$ and logical outcomes $\ell_i \in \{0,1\}^{\{n_L\}}$ for $i = 1, ..., N$. For decoder f, define the binary error indicator

$$E_{i(f)} = 1[f(s_i) \neq \ell_i].$$

The empirical logical error rate is the arithmetic mean

$$\mathrm{LER}\ (f) = \frac{1}{N}\sum_{i=1}^{N} E_{i(f)}.$$

Because every $E_{i(f)}$ is binary, $0 \leq \mathrm{LER}\ (f) \leq 1$. The quantity is a proportion, not a probability assertion about an unobserved device. The campaign-level interpretation is therefore conditional on the simulator and sampling protocol used to create the paired records.

3.2 1.2 Paired difference and relative reduction

For two decoders evaluated on identical records, define

$$\Delta = \mathrm{LER}\ (\mathrm{f_S}) - \mathrm{LER}\ (\mathrm{f_{PM}}).$$

A negative Δ favors SAGE-QEC. The relative reduction is

$$R = \frac{\mathrm{LER}\ (\mathrm{f_{PM}}) - \mathrm{LER}\ (\mathrm{f_S})}{\mathrm{LER}}(\mathrm{f_{PM}}) \times 100\%.$$

The distinction between absolute and relative reduction is essential. Using the reported aggregate values,

$$\Delta = 24.8320\% - 27.1285\% = -2.2965 \ \text{ percentage points},$$

while

$$R = \frac{27.1285 - 24.8320}{27.1285} \times 100\% = 8.47\%$$

, up to rounding in the displayed endpoints.

> **Interpretation rule.** "8.47% relative reduction" does not mean an 8.47 percentage-point reduction. The executed aggregate difference is approximately 2.2965 percentage points.

3.3 1.3 Why pairing matters

Let E_i^S and E_i^P be the two error indicators on the same shot. Then

$$\Delta = \frac{1}{N}\sum_{i=1}^{N}(E_i^S - E_i^P).$$

Pairing preserves the shot-level covariance between decisions. An unpaired comparison would discard exactly the information needed to identify rescues and harmful overrides.

4 2. Decoder map and graphlike baseline

4.1 2.1 Detector records and logical outputs

The decoder receives a binary detector record and returns a logical prediction. For a graphlike detector error model with parity-check matrix H, a candidate correction c is syndrome-consistent when

$$Hc = s \bmod 2.$$

A logical-observable matrix L maps a valid correction to a logical prediction Lc. The nominal matching decoder minimizes a weighted graph cost subject to syndrome consistency.

The graphlike representation is exact for the modeled graphlike mechanisms. It is not automatically exact for a physical mechanism that activates more than two detectors. Such a mechanism is naturally represented by a hyperedge; replacing it with independent binary edges can lose correlation structure.

4.2 2.2 Selective policy abstraction

SAGE-QEC is represented as a policy above the nominal decoder. Let $g(s)$ be the selected alternative candidate and $u(s) \in \{0,1\}$ be the intervention indicator. The policy output is

$$\hat{\ell}_{S(s)} = (1 - u(s))\hat{\ell}_{\mathrm{PM}}(s) + u(s)g(s) \bmod 2.$$

This expression is a routing abstraction. It does not claim that $g(s)$ is a complete physical correction chain.

4.3 2.3 Observable features

The policy uses features computed before the test logical label is revealed. For detector matrix $S \in \{0,1\}^{\{T \times n_d\}}$, examples include block occupancy

$$b_k = \sum_{(t,j)\in B_k} S_{t,j},$$

temporal adjacency

$$a_t = \sum_j S_{t,j} S_{t+1,j},$$

and temporal persistence

$$q = \sum_j S_{1:\frac{T}{2},j} S_{\frac{T}{2}+1:T,j}.$$

These features summarize density, temporal concentration, and cross-time persistence. They do not reconstruct the underlying error.

5 3. Proposition I — default preservation

5.1 Statement

Let $u(s) \in \{0,1\}$ and let $g(s)$ be any alternative candidate. If $u(s) = 0$, then the SAGE-QEC output equals the PyMatching output. Therefore, the disagreement set is contained in the intervention set $I = \{s : u(s) = 1\}$.

5.2 Proof

Start from the routing map:

$$\hat{\ell}_{S(s)} = (1 - u(s))\hat{\ell}_{\mathrm{PM}}(s) + u(s)g(s) \bmod 2.$$

When $u(s) = 0$, the first coefficient is one and the second coefficient is zero. Hence

$$\hat{\ell}_{S(s)} = \hat{\ell}_{\mathrm{PM}}(s).$$

Therefore a disagreement can occur only where $u(s) = 1$, which proves

$$\left\{s : \hat{\ell}_{S(s)} \neq \hat{\ell}_{\mathrm{PM}}(s)\right\} \subseteq I. \quad \square$$

5.3 Operational consequence

The proposition is a structural safety property, not a performance guarantee. It narrows the location of possible changes but does not determine whether an intervention is correct. Any observed improvement must arise from the conditional quality of the selected alternatives on I.

5.4 Corollary: zero-intervention limit

If $u(s) = 0$ for every shot, then $I = \varnothing$ and $\mathrm{f_S} = \mathrm{f_{PM}}$ on the campaign. Consequently, $\mathrm{LER}\ (\mathrm{f_S}) = \mathrm{LER}\ (\mathrm{f_{PM}})$ and $\Delta = 0$. This is the exact zero-coverage endpoint of the risk–coverage curve.

5.5 Corollary: disagreement bound

For any paired campaign, the number of disagreements D satisfies

$$D \leq \sum_i u(s_i).$$

Dividing by N gives $\frac{D}{N} \leq \mathrm{OR}$, where OR is the override rate. Thus intervention coverage is an upper bound on the fraction of decisions that can change.

6 4. Proposition II — exact rescue–harm identity

6.1 Four mutually exclusive cases

For each shot, exactly one of the following holds: shared success $\left(E_i^P = 0, E_i^S = 0\right)$, shared failure $(1,1)$, rescue $(1,0)$, or harmful override $(0,1)$. Let R count rescues and H count harmful overrides.

6.2 Statement

$$\sum_{i=1}^{N} \left(E_i^S - E_i^P\right) = H - R.$$

Consequently,

$$\mathrm{LER}\ (\mathrm{f_S}) - \mathrm{LER}\ (\mathrm{f_{PM}}) = \frac{H - R}{N}.$$

6.3 Proof by partition

On a shared success, the summand is $0 - 0 = 0$. On a shared failure, it is $1 - 1 = 0$. On a rescue, it is $0 - 1 = -1$. On a harmful override, it is $1 - 0 = +1$. Summing over the four disjoint cases yields $H - R$. Division by N yields the second identity. $\square$

6.4 Sign criterion

The identity gives an exact finite-sample condition:

$$\mathrm{LER}\ (\mathrm{f_S}) < \mathrm{LER}\ (\mathrm{f_{PM}}) \quad \text{iff} \quad R > H.$$

This criterion is stronger than a comparison of endpoint bars because it exposes the paired mechanism of the endpoint.

6.5 Derived aggregate accounting

From the twelve reported cells, the listed rescue counts sum to $\sum R_j = 2,940$ and the listed harmful overrides sum to $\sum H_j = 1,071$. Hence the listed-cell net is $R - H = 1,869$ shots. This total is a derived audit of the table and should not be confused with an independent estimate of the full 81,000-shot campaign unless the table counts are known to cover the same denominator without aggregation weighting.

Audit caution. Counts, rates, and aggregate LER values may have different aggregation bases. The identity is exact within a common paired sample; cross-cell summation is presented here only as a transparent derived quantity.

7 5. Proposition III — selective value under coverage

7.1 Definitions

Let $C = |I\frac{|}{N}$ be intervention coverage. Define conditional rescue and harm probabilities on the intervention set:

$$r = \Pr\big(E^P = 1, E^S = 0 \mid u = 1\big),$$
$$h = \Pr\big(E^P = 0, E^S = 1 \mid u = 1\big).$$

7.2 Statement

Under the paired-sampling approximation,

$$\mathrm{LER}\ (\mathrm{f_S}) - \mathrm{LER}\ (\mathrm{f_{PM}}) \approx C(h - r).$$

Therefore SAGE-QEC improves LER at the selected operating point when $r > h$.

7.3 Derivation

By Proposition II,

$$\frac{H - R}{N} = \frac{H}{N} - \frac{R}{N}.$$

Since $|I\frac{|}{N} = C$, write

$$\frac{H}{N} = \left(|I\frac{|}{N}\right)\left(\frac{H}{|}I|\right) \approx Ch$$

, and

$$\frac{R}{N} \approx Cr.$$

Substitution gives

$$\Delta \approx C(h - r).$$

The sign is negative exactly when $r > h$.

7.4 Constrained operation

An industrial policy may impose $C \le C_{\max}$. The relevant requirement is not merely that $r > h$ at unconstrained coverage. It is that the inequality remains true in the allowed operating region. This is why a single LER point is insufficient: the full risk–coverage curve is the appropriate deployment object.

7.5 Special cases

If $C = 0$, then $\Delta = 0$. If $C > 0$ and $r = h$, the intervention layer is neutral in the first-order accounting. If $C > 0$ and $r < h$, the layer increases LER. These statements are algebraic consequences of the decomposition, not claims about any particular hardware distribution.

8 6. Override rate, precision, and coverage

8.1 Override rate

For shot-level indicators u_i, define

$$\mathrm{OR} = \frac{1}{N}\sum_i u(s_i).$$

The rate measures how often the learned policy requests replacement of the default. It is not a correctness metric.

8.2 Override precision

Among interventions that change the final logical decision, define the override precision as the rescue count divided by the number of changed decisions:

$$\mathrm{OP} = \frac{R}{R + H}.$$

The numerator is the rescue count. The denominator counts interventions that actually alter the endpoint. If an intervention selects the same logical output as PyMatching, it is operationally inert and should not inflate precision.

8.3 Relationship to the identity

When the denominator is the set of changed decisions, precision describes the beneficial share of changes, while Δ depends on the difference between rescues and harms normalized by all shots. Therefore a high precision can coexist with small aggregate improvement if coverage is low; conversely, a moderate precision can still reduce LER if the number of rescues exceeds harms.

8.4 Derived ranges from the audit table

The manuscript reports override rates of approximately 64%–99% and override precision of approximately 59%–83% across the aggregate cells. The reported range itself is a deployment warning: the current system is not yet a low-intervention safety layer under distribution shift.

9 7. Label-separation lemma

9.1 Statement

Let X denote the detector record and Y the logical outcome. If the routing policy is measurable with respect to X and its parameters θ are frozen using training and validation data, then at test time

$$u \perp Y \mid X.$$

9.2 Proof

The policy is a deterministic or seeded function $u = \pi(X; \theta)$. Conditional on X, the value of u is fixed once θ is fixed. The evaluator label Y is revealed only after routing for metric calculation. Therefore the routing variable cannot depend on the subsequently revealed label, establishing conditional independence. □

9.3 What the lemma does not prove

The lemma does not prove calibration, candidate completeness, or physical validity of the selected correction. It only formalizes the information boundary required to prevent test-label leakage.

9.4 Practical audit checklist

A compliant execution should record that structure thresholds are computed from training data, the policy is selected on validation data, regime names are not supplied at inference, and logical outcomes are used only after inference. The manuscript states all four conditions for the executed campaign.

10 8. Policy rule and threshold geometry

10.1 Inference rule

Let $\mathrm{p_{alt}}$ be the alternative-candidate confidence, p_2 the second-highest candidate score, τ the confidence threshold, δ the top-two margin, and z an out-of-distribution statistic. The policy selects the alternative candidate when $\mathrm{p_{alt}} \geq \tau$, $\mathrm{p_{alt}} - p_2 \geq \delta$, and $z \leq \mathrm{z_{max}}$; otherwise it retains the PyMatching output.

The three gates encode distinct requirements. The confidence gate asks whether the proposed candidate is sufficiently likely. The margin gate asks whether the proposed candidate is separated from its nearest competitor. The OOD gate limits routing outside the calibrated feature region.

10.2 Monotonicity intuition

Holding all other quantities fixed, increasing τ or δ can only remove interventions from the feasible gate set. Decreasing z_{max} also tightens the OOD gate. This does not imply monotonic improvement in LER because removing an intervention may remove either a rescue or a harm.

10.3 Candidate-set limitation

The current candidate set includes nominal PyMatching, nearby calibrated PyMatching models, and structure-conditioned logical hypotheses. These are logical-output proxies rather than complete data-qubit correction chains. Consequently, the policy is not a full maximum-likelihood or hypergraph decoder.

11 9. Constrained-risk formulation

11.1 Optimization target

Let Π be the admissible class of routing policies. A deployment-aware objective is

$$\min_{\pi \in \Pi} \mathrm{LER}\ (\pi) \quad \text{subject to} \quad C(\pi) \leq \mathrm{C_{max}}.$$

The constraint must be fixed before blind testing if the resulting claim is to be confirmatory rather than post hoc.

11.2 Why V10 is informative

The manuscript reports a later V10 experiment with an intervention budget of approximately 10% and an OOD gate. V10 was statistically indistinguishable from nominal PyMatching in most cells and slightly worse in two D=7 cells. This negative control indicates that the present candidate set does not yet provide enough high-confidence alternatives at realistic selective coverage.

11.3 Derived decision rule

Suppose an operating point satisfies $C \leq C_{\mathrm{max}}$. Proposition III implies a sufficient local requirement $r > h$. If the budget forces a region where $r \leq h$, tightening the gate cannot recover improvement unless the candidate ranking or candidate generator changes the conditional composition of interventions.

12 10. Bootstrap interval and inferential scope

12.1 Resampling algorithm

For a cell with paired observations $\left(E_i^S, E_i^P\right)$, sample N indices with replacement. For replicate b, compute

$$\Delta^b = \frac{1}{N} \sum_i \left(E_{i_b}^S - E_{i_b}^P\right).$$

After 800 replicates, report the 2.5th and 97.5th percentiles.

12.2 What is quantified
The interval quantifies shot-level uncertainty under the executed campaign and its data-generating process. A reported interval below zero indicates that the paired difference is negative throughout the displayed percentile range.

12.3 What is not quantified
The interval does not establish superiority under an arbitrary hardware distribution. Repeated seeds and structured regimes mean that shot-level resampling is not the same as independent replication across conditions. A future confirmatory analysis should use hierarchical resampling or a preregistered mixed-effects model.

12.4 Multiplicity boundary
The manuscript presents the twelve distance-by-regime cells as an exploratory performance map and does not apply a post-hoc multiplicity correction. The aggregate result is the primary summary. The cells should therefore be interpreted descriptively rather than as twelve independent confirmatory claims.

13 11. Re-derivation of the reported aggregate endpoint
The reported aggregate endpoint is

$$\mathrm{LER}_{\mathrm{PM}} = 0.271285, \quad \mathrm{LER}_{\mathrm{S}} = 0.248320.$$

Subtracting gives

$$0.248320 - 0.271285 = -0.022965.$$

Multiplying by 100 gives -2.2965 percentage points.
For the relative reduction,

$$\frac{0.271285 - 0.248320}{0.271285} = 0.0847$$

, which is 8.47% after multiplication by 100 and rounding to two decimal places.

13.1 Cell-level cross-check
For D=3 Burst-3,

$$\frac{8.711 - 4.622}{8.711} \times 100 = 46.94\%$$

. For D=5 Crosstalk,

$$\frac{33.126 - 30.726}{33.126} \times 100 = 7.24\%$$

. For D=7 Mixed,

$$\frac{43.496 - 41.437}{43.496} \times 100 = 4.73\%$$

. These calculations reproduce the displayed relative reductions up to rounding.

13.2 Directional consistency
Every listed cell has $\mathrm{LER}_{\mathrm{S}} < \mathrm{LER}_{\mathrm{PM}}$. The direction is therefore consistent across the twelve reported distance-by-regime summaries, while the size of the reduction varies substantially by regime and distance.

14 12. Scaling and distribution-shift interpretation
The strongest reported reductions occur in Burst-3 at D=3 and D=5, with relative reductions of 46.94% and 45.00%. The smallest reduction is D=3 Drift at 2.89%. At D=7, reductions range from 4.73% to 19.77%.
This pattern is consistent with a decoder that benefits from structured local signatures but captures only part of the mismatch under drift and mixed mechanisms. It is not evidence of universal dominance. The formal reason is that the empirical endpoint is distribution-indexed: changing the regime changes the joint distribution of records, labels, and candidate quality.

14.1 Non-commutation of averaging and ratios
The manuscript reports an aggregate relative reduction computed from aggregate LER values. In general,

$$\frac{\sum_j w_j P_j - \sum_j w_j S_j}{\sum_j w_j P_j} \neq \sum_j w_j \frac{P_j - S_j}{P_j}.$$

Therefore a simple arithmetic average of the twelve cell-level relative reductions need not equal 8.47%. The aggregate formula should remain the primary calculation.

15 13. Reproducibility and information boundary
15.1 Experimental design
The campaign uses distances $D \in \{3, 5, 7\}$, nominal parameters $p \in \{0.002, 0.003, 0.005\}$, and seeds $\{11, 22, 33, 44, 55\}$. This creates 45 distance–parameter–seed conditions. Each condition contains 450 shots per blind regime, yielding 1,800 blind shots per condition and 81,000 paired shots overall.

15.2 Information separation

Training regimes are Independent, Measurement, and Burst-2. Blind regimes are Burst-3, Temporal Drift, Crosstalk, and Mixed Noise. The regime label is not given to the model at inference. The detector record is the test-time observation; the logical outcome is evaluator-only.

15.3 Stress generators

Burst-3 uses event probability 0.16 and a 35% selected-detector fraction across three early temporal slices. Drift uses per-shot flip probability $0.01 + 0.035U$. Crosstalk uses seed-detector probability 0.012 with immediate-neighbor propagation. Mixed Noise combines independent flips at 0.012 with neighbor propagation triggered at 0.009. These are controlled simulator stress tests, not hardware measurements.

16 14. Assumption ledger and limitation map

Claim type	Required assumption	Boundary
Default preservation	Policy uses the intervention indicator and default output	Does not guarantee beneficial overrides
Rescue–harm identity	Common paired denominator	Exact within that paired sample
Selective-value condition	Stable paired sampling approximation	Not a universal hardware theorem
Label separation	Frozen parameters and label-after-routing	Does not prove calibration
Bootstrap interval	Resampling represents campaign variation	Does not identify arbitrary device distribution

16.1 Industrial interpretation

The mathematical guarantees establish auditability and routing structure. They do not establish production readiness. The manuscript explicitly identifies missing physical correction candidates, incomplete correlated-noise calibration, absent hardware calibration data, and unmeasured deployment latency and memory.

16.2 Next-stage mathematical target

A stronger study should define a physically valid candidate set, calibrate the correlated-noise baseline on the same model, preregister a coverage constraint, and report the full risk–coverage curve. These additions would convert the current qualification-stage prototype into a more complete selective-decoding evaluation.

17 15. Consolidated conclusions

The central algebraic result is

$$\Delta = \frac{H - R}{N}.$$

It explains why SAGE-QEC can improve the endpoint only when rescues exceed harmful overrides on the same paired shots. The central routing result is that all differences are confined to the intervention set. The central information result is that frozen, detector-only routing can be separated from the evaluator label. The central inferential result is that paired bootstrap intervals describe the executed campaign, not arbitrary hardware.

The reported campaign provides a coherent simulator-level result: aggregate LER decreases from 27.1285% to 24.8320%, an absolute reduction of approximately 2.2965 percentage points and a relative reduction of 8.47%, across 81,000 paired blind shots. The result is distribution-dependent, and the high intervention rate together with the V10 negative control limits the strength of an industrial claim.

> **Final audit statement.** The supplement supports a strong qualification-stage interpretation: SAGE-QEC is an auditable selective layer that improved the reported simulator campaign. It does not support a claim of universal decoder superiority, hardware readiness, or complete physical correction.

End of Supplementary Information I · Mathematical Derivations and Audit Identities