Supplementary Information II

Numerical Tables, Figures, and Campaign Audit

SAGE-QEC paired blind evaluation under structured noise

All values reproduced from, or deterministically derived from, the executed manuscript

Data provenance. This file does not add simulated observations. Figures were regenerated from the twelve reported distance-by-regime cells and the intervention-audit table. Percentages are displayed in the same units as the manuscript.

1 Contents and conventions

The numerical supplement moves from campaign design to aggregate endpoints, cell-level performance, intervention accounting, uncertainty intervals, and deployment trade-offs. “PM” denotes PyMatching and “S” denotes SAGE-QEC. Relative reductions are calculated as $\frac{\mathrm{LER}_{\mathrm{PM}}-\mathrm{LER}_{\mathrm{S}}}{\mathrm{LER}_{\mathrm{PM}}}$ and are not percentage-point differences.

2 1. Experimental campaign design

Design element	Declared setting	Role in the claim
Code family	Rotated surface-code memory	Fixes the physical decoding task
Code distances	3, 5, 7	Tests scaling across geometry
Nominal parameters	0.002, 0.003, 0.005	Tests several operating points
Independent seeds	11, 22, 33, 44, 55	Tests stochastic repeatability
Training regimes	Independent, Measurement, Burst-2	Information available during fitting
Blind regimes	Burst-3, Drift, Crosstalk, Mixed	Distribution-shift evaluation
Primary endpoint	Paired logical error rate	Main scientific comparison
Secondary endpoints	Override rate, precision, latency	Deployment-risk characterization

Each of the 45 distance–parameter–seed conditions contains 450 shots per blind regime. With four blind regimes, this gives 1,800 paired shots per condition and 81,000 paired blind test shots in total.

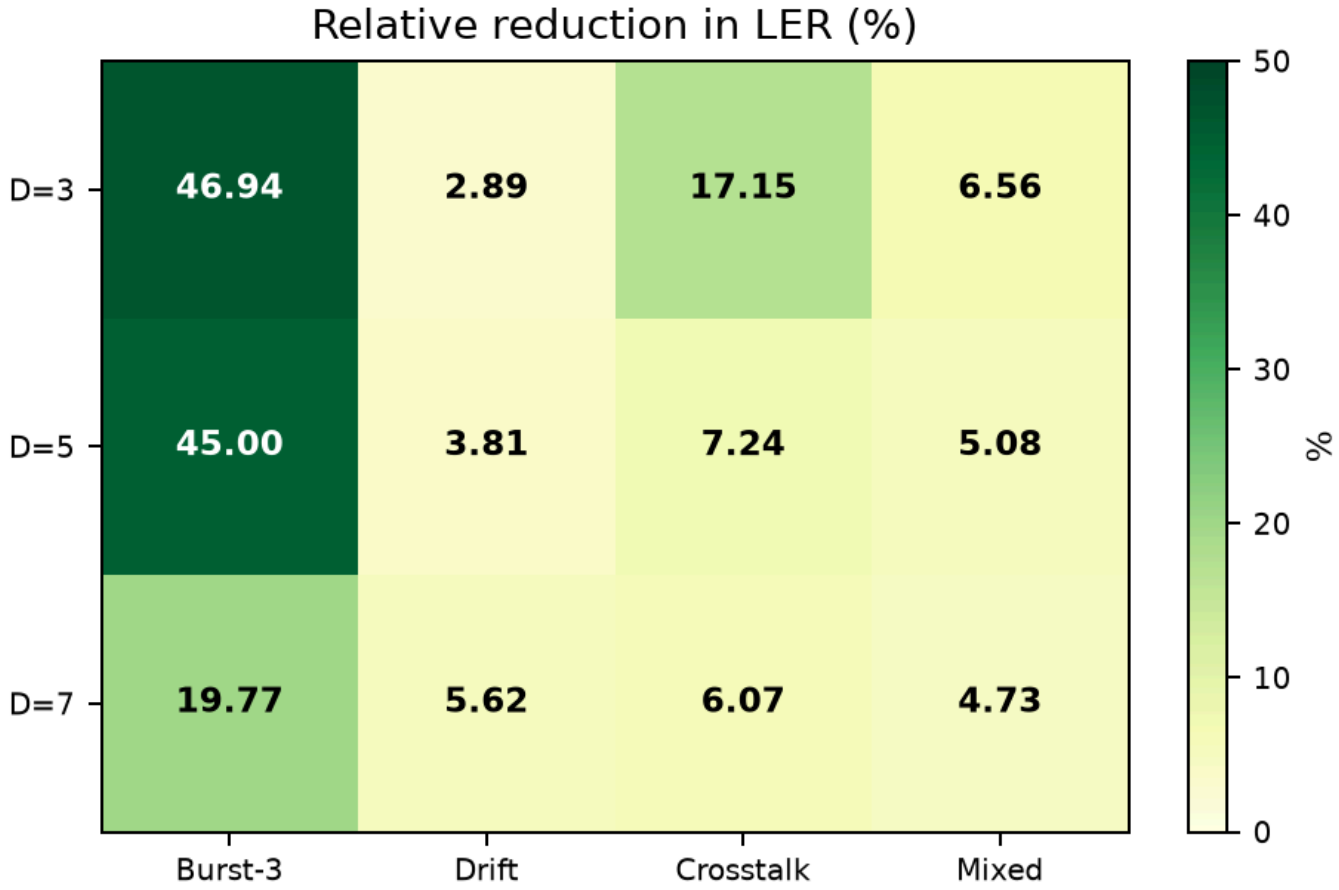

Figure 1: Figure 1. The twelve reported relative reductions, arranged by distance and withheld regime.

3 2. Aggregate endpoint

Decoder / metric	Aggregate value
PyMatching LER	27.1285%
SAGE-QEC LER	24.8320%
Absolute reduction	2.2965 percentage points
Relative reduction	8.47%

The absolute reduction is the difference between the two aggregate rates. The relative reduction divides that difference by the PyMatching rate. The distinction matters for accurate reporting.

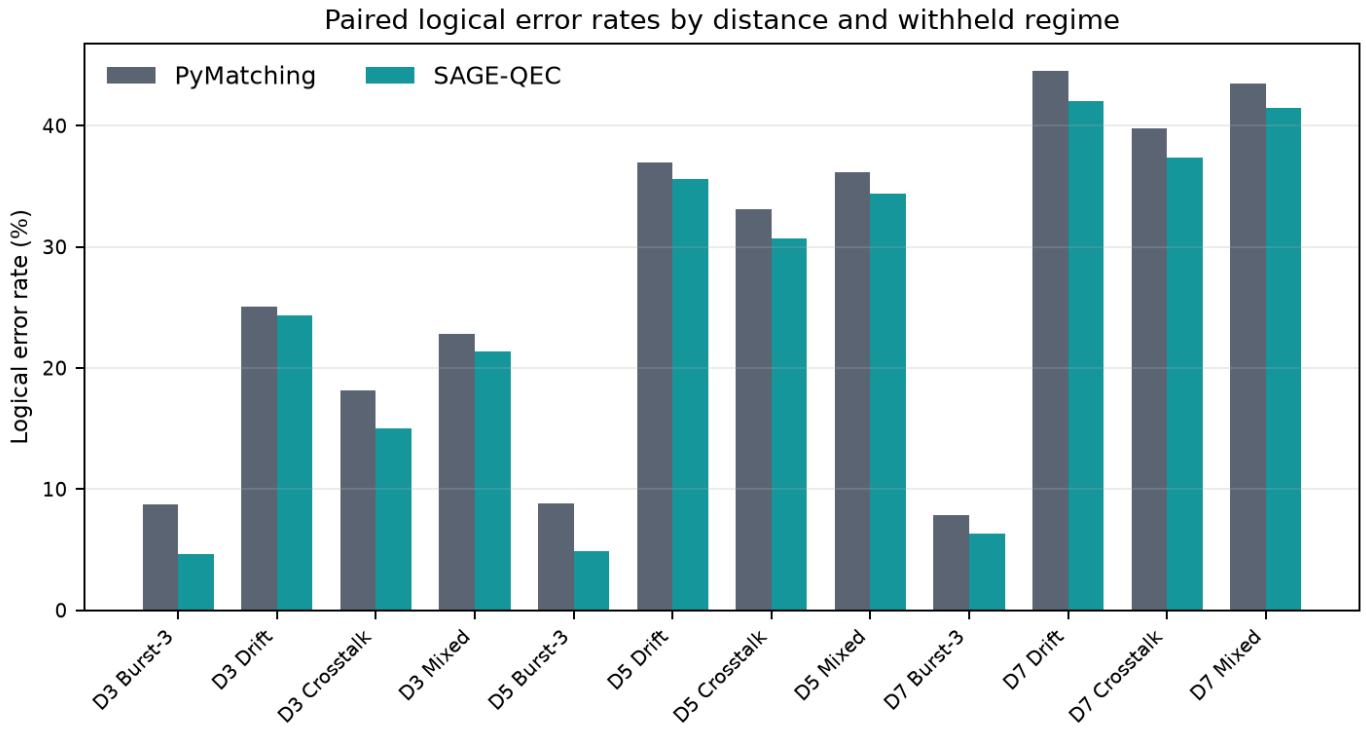

Figure 2: Figure 2. Logical error rates by distance and withheld regime.

4 3. Cell-level performance table

D	Regime	PM LER	SAGE LER	Reduction
3	Burst-3	8.711%	4.622%	46.94%
3	Drift	25.067%	24.341%	2.89%
3	Crosstalk	18.133%	15.022%	17.15%
3	Mixed	22.830%	21.333%	6.56%
5	Burst-3	8.800%	4.844%	45.00%
5	Drift	36.978%	35.570%	3.81%
5	Crosstalk	33.126%	30.726%	7.24%
5	Mixed	36.193%	34.356%	5.08%
7	Burst-3	7.867%	6.311%	19.77%
7	Drift	44.548%	42.044%	5.62%
7	Crosstalk	39.793%	37.378%	6.07%
7	Mixed	43.496%	41.437%	4.73%

All twelve cells favor SAGE-QEC in the executed campaign. The largest displayed relative reductions occur in Burst-3 at D=3 and D=5.

5 4. Paired intervention audit

Cell	Rescues	Harmful	Override rate	Override precision
D3 Burst-3	353	77	98.8%	82.1%
D3 Drift	79	30	97.3%	72.5%
D3 Crosstalk	262	52	98.4%	83.4%
D3 Mixed	136	35	98.0%	79.5%
D5 Burst-3	437	170	99.2%	72.0%
D5 Drift	161	66	97.2%	70.9%
D5 Crosstalk	258	96	98.3%	72.9%
D5 Mixed	196	72	97.6%	73.1%
D7 Burst-3	331	226	72.6%	59.4%
D7 Drift	241	72	64.1%	77.0%
D7 Crosstalk	256	93	69.2%	73.4%
D7 Mixed	221	82	65.5%	72.9%

A rescue is a baseline failure corrected by SAGE-QEC. A harmful override is a baseline success changed into an error. The endpoint is therefore a net result, not a count of successful overrides alone.

6 5. Rescue–harm visualization

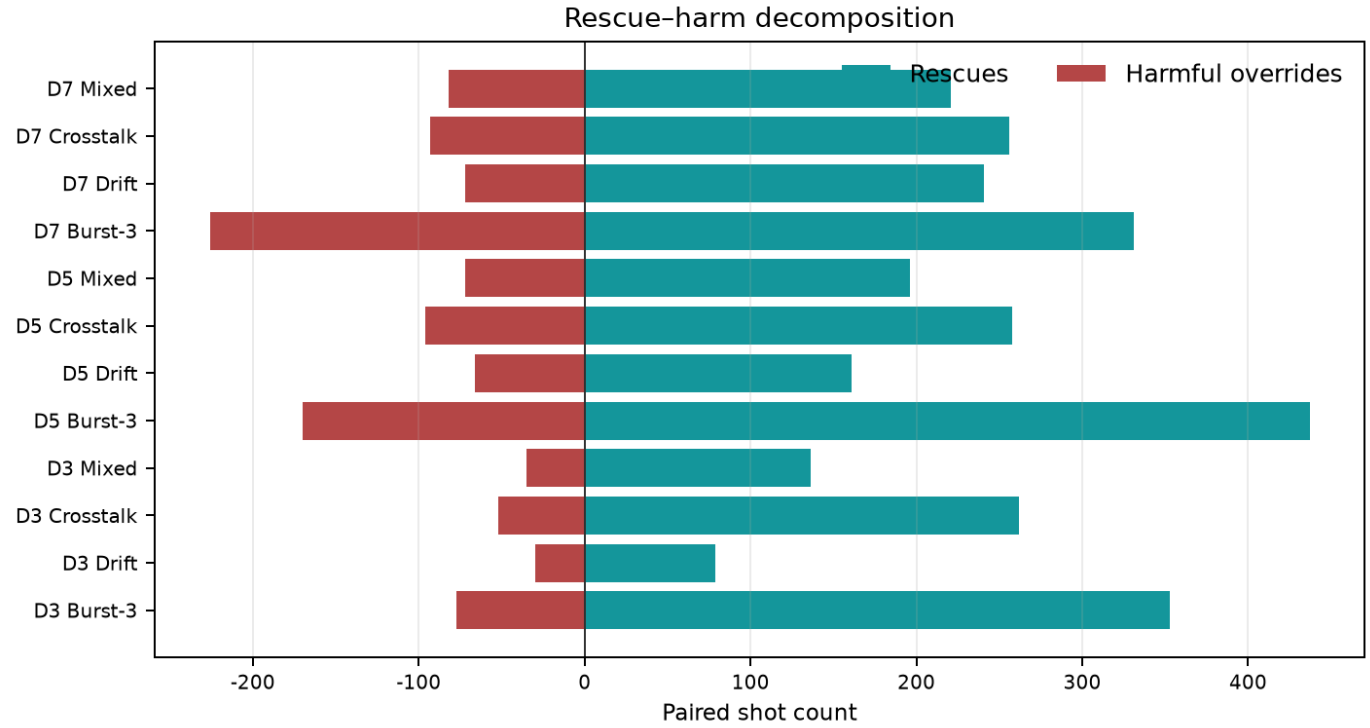

Figure 3: Figure 3. Paired rescue and harmful-override counts. Positive bars indicate rescues; negative bars indicate harmful overrides. The plotted counts are the exact integers reported in the intervention audit. D=7 Burst-3 has a notably less favorable precision profile than D=3 Burst-3, despite a negative net LER difference.

Derived identity. Within a common paired denominator, $\Delta N = H - R$ and $\Delta\ \mathrm{LER} = \Delta\frac{N}{N}$. The chart should be read together with the endpoint table rather than as a standalone performance metric.

7 6. Bootstrap uncertainty

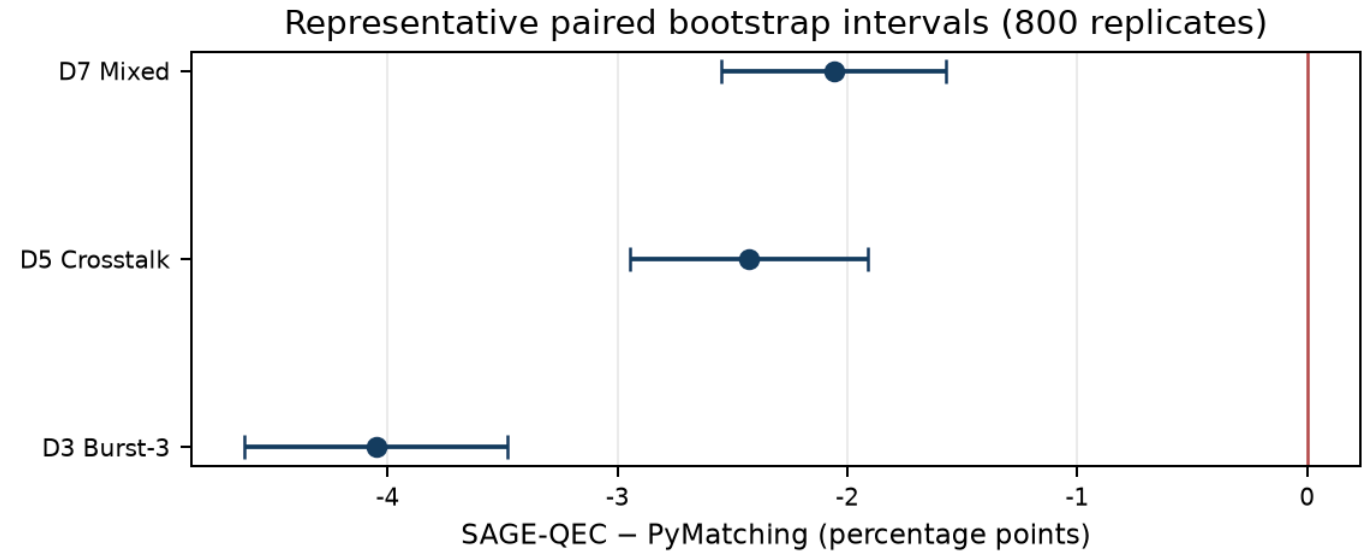

Figure 4: Figure 4. Representative paired bootstrap intervals for SAGE-QEC minus PyMatching.
The intervals shown are the manuscript's representative 2.5th–97.5th percentile intervals from 800 paired bootstrap replicates: D=3 Burst-3, D=5 Crosstalk, and D=7 Mixed. Every displayed interval lies below zero.

Cell	Lower bound (pp)	Upper bound (pp)
D3 Burst-3	−4.623	−3.481
D5 Crosstalk	−2.948	−1.910
D7 Mixed	−2.549	−1.570

8 7. Coverage–precision trade-off

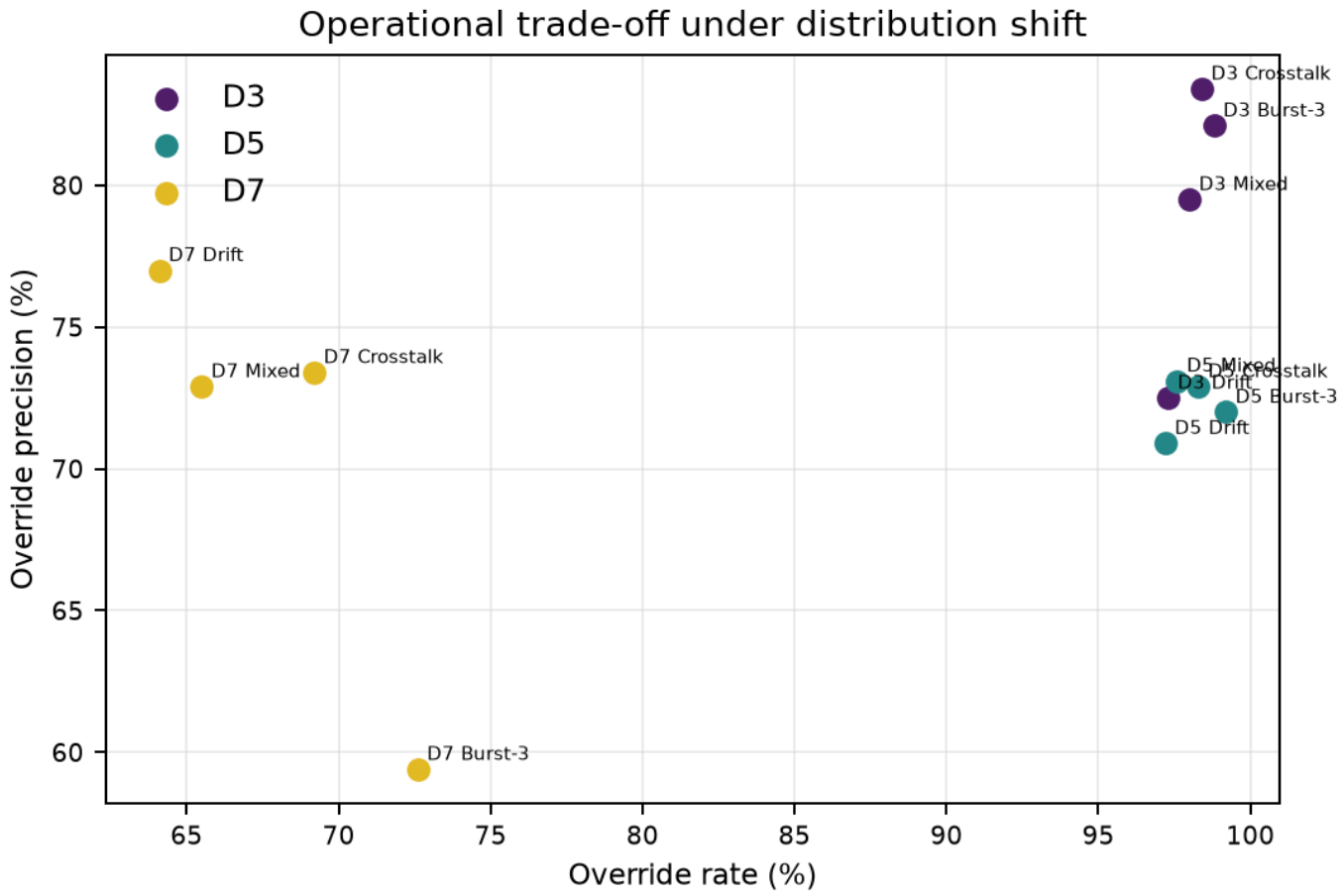

Figure 5: Figure 5. Operational trade-off under distribution shift.

The executed campaign reports override rates of approximately 64%–99% and precision of approximately 59%–83%. High coverage can be acceptable in a simulator qualification study, but it is not equivalent to a low-intervention deployment guarantee.

The later V10 negative control imposed an intervention budget of approximately 10% together with an OOD gate. It was statistically indistinguishable from nominal PyMatching in most cells and slightly worse in two D=7 cells. This finding motivates a future physically complete candidate generator.

9 8. Figure index and reporting notes

No.	Figure	Interpretive use
1	Relative-reduction heatmap	Compact view of all twelve cell-level gains
2	Logical error grouped bars	Direct endpoint comparison
3	Rescue–harm decomposition	Paired mechanism behind net LER
4	Bootstrap intervals	Uncertainty for representative cells
5	Coverage–precision scatter	Operational deployment trade-off

9.1 Reproducibility notes

The graphics use only the values transcribed from the manuscript's aggregate result, cell table, intervention audit, and representative intervals. No random data, inferred shot labels, or unreported confidence intervals were added. The plotting script is retained with the deliverables as a machine-readable audit trail.

9.2 Aggregation note

The 12 cell summaries are distance-by-regime aggregates across nominal parameters and seeds. The manuscript separately reports 45 simulator conditions and 81,000 paired shots. The figures preserve the manuscript's aggregation level and should not be interpreted as 12 independent experiments.

10 9. Audit-ready data dictionary

Field	Definition	Unit / provenance
Distance	Rotated surface-code distance	Dimensionless; reported
Regime	Withheld stress mechanism	Categorical; reported
PM LER	Logical error rate of PyMatching	Percent; reported

SAGE LER	Logical error rate of SAGE-QEC	Percent; reported
Reduction	(PM–SAGE)/PM	Percent; derived/reported
Rescues	PM wrong, SAGE correct	Count; reported
Harmful	PM correct, SAGE wrong	Count; reported
Override rate	Intervention frequency	Percent; reported
Override precision	Beneficial changed decisions / changed decisions	Percent; reported

This dictionary is designed to make the supplement independently auditable without changing the scientific scope of the source manuscript.

11 10. Consolidated numerical conclusions

Question	Answer supported by the executed campaign
Did SAGE-QEC reduce aggregate LER?	Yes: 27.1285% to 24.8320%
How large was the absolute reduction?	Approximately 2.2965 percentage points
How large was the relative reduction?	8.47%
Did all twelve cells favor SAGE-QEC?	Yes, in the reported aggregate map
Was intervention low under shift?	No; approximately 64%–99%
Was hardware readiness established?	No; no hardware calibration or deployment latency measurement
Was universal superiority established?	No; the claim is simulator- and distribution-dependent

Bottom line. The numerical evidence is internally coherent for a qualification-stage simulator study. Its strongest claim is reproducible paired improvement under the declared stress regimes. Its most important caveat is the high intervention rate and the incompleteness of the candidate correction set.

End of Supplementary Information II · Numerical Tables and Figures