Online Resource 1: Factorial Benchmarking of Exact and Finite-Shot Gradient Resolvability in a Four-Qubit Hybrid Quantum Neural Network

Brandon Shen (ORCID: 0009-0002-3545-2106)[1*]

[1]Independent Researcher, Belmont, California, United States.

Corresponding author: brandonshen123@gmail.com

Supplementary Information for *Discover Computing*

1 Scope and frozen provenance

This supplement reports the corrected centered H1–H4 family and the post-primary analyses frozen before manuscript revision. The machine-readable source of truth is `results/final_submission_v1/manifest.json`. The numerical-results freeze, subm ission-numerical-results-freeze-v1, preserves the frozen scientific outputs. The accompanying OSF project is available at https://doi.org/10.17605/OSF.IO/THV6G. The distinct final manuscript/repository release, `sndc-submission-v1`, contains the finalized manuscript and supplement sources, code, documentation, verification scripts, and submission-facing repository state. Historical outputs under `results/superseded/`, `results/superseded_pooled/`, and other explicitly superseded directories are audit-only and are not current primary results.

2 Factor-coding correction and estimands

The adopted factors are

$$E_c = E - \tfrac{1}{2}, \qquad L_c = L - \tfrac{1}{2}, \qquad R_c = R - \tfrac{1}{2},$$

with values $\{-1/2, +1/2\}$. Depth is $D_z = (D - 3.2)/1.7204650534085253$, where 3.2 and 1.7204650534085253 are respectively the mean and population standard deviation of $D \in \{1, 2, 3, 4, 6\}$. Because $E_c L_c R_c$ remains in each applicable model, centered

lower-order interactions average equally over the third binary factor. Direct-0/1 lower-order coefficients instead describe simple interactions at the third factor's reference level. The parameterizations yield identical fitted values, likelihood, residuals, and random effects; only the interpretation of lower-order coefficients changes. The correction was an implementation–interpretation correction rather than result-dependent model selection. H1 and H2 use E_cL_c, H3 uses E_cR_c averaged over L, and H4 uses $L_cR_cD_z$.

3 Implemented forward pass, dimensions, and SNR definition

For every block ℓ, $x^{(\ell)}, z^{(\ell)}, \theta^{(\ell)} \in \mathbb{R}^4$, $W_\ell \in \mathbb{R}^{4\times4}$, and $b_\ell \in \mathbb{R}^4$. The code sets $x^{(1)} = z^{(0)} = 0$. Each block resets to $|0000\rangle$, applies $R_y(\theta_q^{(\ell)} + x_q^{(\ell)})$ for $q = 0, \ldots, 3$, and then applies its four ordered CNOTs. For $\ell < D$, a common Z-basis measurement returns all four $z_j^{(\ell)} = \langle Z_j \rangle$ and $x^{(\ell+1)} = W_\ell z^{(\ell)} + b_\ell + \gamma z^{(\ell-1)}$, with $\gamma = R$. The terminal block is evaluated with the selected objective. There is no clipping, wrapping, normalization, nonlinearity, or rescaling. The $4D$ quantum-rotation derivatives enter H1–H4; the $20(D-1)$ classical W, b derivatives do not.

Quantum angles are independent Uniform$(0, 2\pi)$ draws. Each W_ℓ is Glorot-uniform on $[-\sqrt{6/8}, \sqrt{6/8}]$ and every b_ℓ is zero. SHA-256-derived quantum, classical, shot, and bootstrap streams are separated. All configurations at fixed initialization and D reuse identical arrays and indices; distinct D values use depth-specific seeds and do not share shallow prefixes.

At fixed parameters, the population target is $\mathrm{SNR}^*_{\mathrm{est},k} = |\mu_k|/\sigma_k$, where μ_k and σ_k^2 are the repeated-shot mean and variance of the complete gradient estimator. The analyzed plug-in quantity is

$$\widehat{\mathrm{SNR}}_{\mathrm{est},k} = |\bar{g}_k|/s_k, \quad s_k^2 = \frac{1}{29}\sum_{r=1}^{30}(\hat{g}_{kr} - \bar{g}_k)^2.$$

Duplicate replicate keys and invalid rows are rejected before aggregation; the frozen validation reported neither. When $s_k^2 = 0$, the ratio is undefined and excluded without an epsilon or floor. The include-zero sensitivity assigns zero only when $\bar{g}_k = s_k = 0$. H2–H4 concern this finite-replicate plug-in ratio.

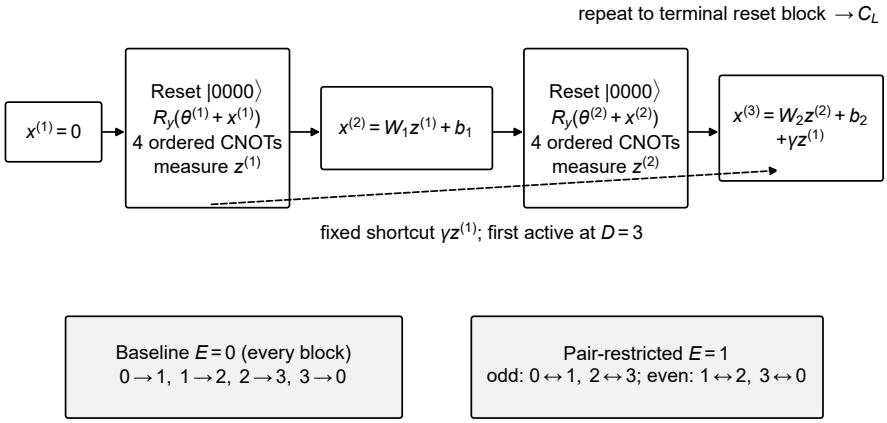

Fig. 1 Code-verified reset-per-block forward pass. Solid arrows transmit four-dimensional classical features; the dashed fixed-gain shortcut is active only for $R = 1$ and first changes the third block input. Every quantum box starts with a reset register. Baseline and pair-restricted CNOT schedules are shown without implying continuous quantum-circuit depth across blocks

4 Corrected primary family and bootstrap

For H2–H4, outer resampling selected complete initialization clusters and retained all configurations, depths, budgets, and matched parameters; within selected scientific cells, finite-shot replicates were resampled and SNR was recomputed before refitting. H1 resampled complete initialization clusters. Percentile intervals are reported without treating either Wald or bootstrap inference as universally preferable.

The frozen finite-shot production dataset used 30 independent gradient replicates per scientific cell and 50 independent top-level initialization clusters. The replicate pilot began at 30, evaluated candidate larger counts on prespecified representative cells, and did not satisfy its signed-mean and plug-in-SNR precision/stability criterion uniformly; the problem was most evident at shallow block counts. The already collected production dataset therefore retained 30, with no production recollection, making the sample sizes pilot-informed rather than pilot-calibrated.

The initialization-count pilot evaluated candidates in increments of 25 using the then-implemented H2–H4 mixed-model coefficients on the $\operatorname{asinh}(\mathrm{SNR}_{\mathrm{est}})$ response scale. At $N_{\mathrm{init}} = 50$, the recorded 90th-percentile Wald 95% CI half-widths were 0.1026212747780318 for $E{:}L$, 0.10253728880520716 for $E{:}R$, and 0.07250672321768473 for $L{:}R{:}\mathrm{depth}_z$, each below the configured 0.20 rule. That threshold was operational precision guidance on the transformed response scale, not a smallest effect of interest, equivalence margin, or scientific-importance claim. The pilot preceded the later centered-factor interpretation correction and did not evaluate H1.

The final H2–H4 pool preserves the earlier 443 successful draws unchanged (8 validated draws plus three independently seeded 145-draw shards) and appends non-overlapping deterministic continuation draws to reach exactly 1,000 unique valid fits under the revised stopping instruction. A further 178 valid fits completed before

Hyp.	Estimand	Estimate	Wald 95% CI	Holm p	Bootstrap 95% CI	Evidence
H1	$E_c L_c$, exact	0.004043	[0.001924, 0.006162]	0.000739	[0.000473, 0.007535]	Both
H2	$E_c L_c$, end-to-end	0.014338	[0.004255, 0.024422]	0.015963	[−0.016240, 0.045992]	Wald only
H3	$E_c R_c$, end-to-end	-0.011615	[−0.021697, −0.001534]	0.047875	[−0.031375, 0.009563]	Wald only
H4	$L_c R_c D_z$	-0.010179	[−0.020688, 0.000331]	0.057658	[−0.025964, 0.005863]	Neither

Table 1 Corrected primary family. Wald intervals and Holm-adjusted p-values follow the primary model-based decision rule. Bootstrap intervals use initialization-cluster resampling for H1 and nested initialization-cluster/within-cell replicate resampling for H2–H4. "Both" means that the Wald and bootstrap intervals exclude zero; "Wald only" means that the Wald/Holm rule rejects while the bootstrap interval includes zero; and "Neither" means that the Wald/Holm rule does not reject and the bootstrap interval includes zero. H1 used 2,000 completed fits with zero failures; H2–H4 used exactly 1,000 unique valid fits. Bootstrap iterations are distinct from the 50 independent initialization clusters.

workers stopped; they are preserved in the superseded audit location and excluded from every reported interval. The 443-fit result remains an audit checkpoint, not the final result. Machine-readable provenance files record attempted, included-successful, excluded-successful, failed, and rejected counts, global iteration IDs, shard seeds, convergence status, and coefficient vectors. These counts refer to completed bootstrap fits, not independent initialization clusters. Percentile endpoints retain finite Monte Carlo uncertainty.

Completed fits	H2 E_cL_c interval	H3 E_cR_c interval	H4 $L_cR_cD_z$ interval
40	[-0.011931, 0.035640]	[-0.028213, 0.000850]	[-0.024829, 0.006450]
100	[-0.019772, 0.041956]	[-0.033480, 0.009079]	[-0.027224, 0.005826]
200	[-0.017723, 0.044496]	[-0.032630, 0.006696]	[-0.024829, 0.005711]
400	[-0.016692, 0.043008]	[-0.030956, 0.010311]	[-0.026885, 0.006433]
443	[-0.016638, 0.043563]	[-0.030943, 0.010259]	[-0.026975, 0.006403]
1,000	[-0.016240, 0.045992]	[-0.031375, 0.009563]	[-0.025964, 0.005863]

Table 2 H2–H4 nested-bootstrap percentile endpoint checkpoints. Every interval includes zero. The 443 row is the preserved historical checkpoint; 1,000 is the final included fit count. Rows are cumulative completed refits of resampled data, not independent datasets; endpoint movement reflects finite Monte Carlo uncertainty. Order-statistic/binomial-rank endpoint diagnostics in the machine-readable summary quantify Monte Carlo uncertainty and are not scientific confidence intervals.

5 Independent-seed H1 and depth weighting

The independent-seed centered H1 estimate was 0.007726 (Wald 95% CI $[0.005592, 0.009860]$); its 2,000-fit, zero-failure bootstrap interval was $[0.003402, 0.011923]$. The direction was retained but magnitude remained uncertain.

In the original dataset, moderation tests were mixed-model $\chi^2(4) = 22.844$, $p = 0.000136$, and cluster-robust $\chi^2(4) = 6.008$, $p = 0.1985$. Independent-seed tests were $\chi^2(4) = 7.563$, $p = 0.1090$, and $\chi^2(4) = 1.810$, $p = 0.7706$. The original-dataset depth conclusion is covariance-sensitive; cross-seed magnitude differences have no clear depth localization under the frozen rule.

D	Original estimate	Original 95% CI	Independent-seed estimate	Independent-seed 95% CI
1	-0.005251	[-0.013708, 0.003206]	0.001862	[-0.006653, 0.010378]
2	-0.003579	[-0.009560, 0.002401]	0.003120	[-0.002901, 0.009141]
3	0.012588	[0.007706, 0.017471]	0.006250	[0.001334, 0.011166]
4	0.004664	[0.000435, 0.008892]	0.007885	[0.003627, 0.012143]
6	0.003446	[-0.000007, 0.006898]	0.010870	[0.007393, 0.014346]
Equal depth	0.002374	[-0.000164, 0.004911]	0.005997	[0.003443, 0.008552]
Observation/parameter weighted	0.004043	[0.001929, 0.006157]	0.007726	[0.005597, 0.009854]

Table 3 Post-primary H1 depth contrasts and weighting summaries. Observation and matched-parameter weights are identical: 0.0625, 0.125, 0.1875, 0.25, and 0.375 at $D = 1, 2, 3, 4, 6$. Intervals are unadjusted Wald 95% intervals.

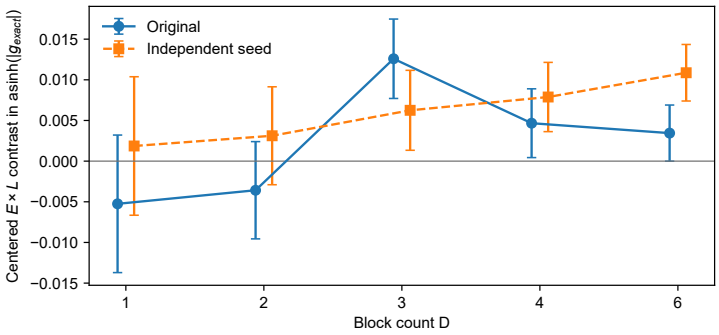

Fig. 2 Post-primary centered H1 contrasts by block count. Bars are unadjusted Wald 95% intervals; zero is the reference. Dataset identity is encoded by shape and line style, not color alone. No monotonic depth law is assumed

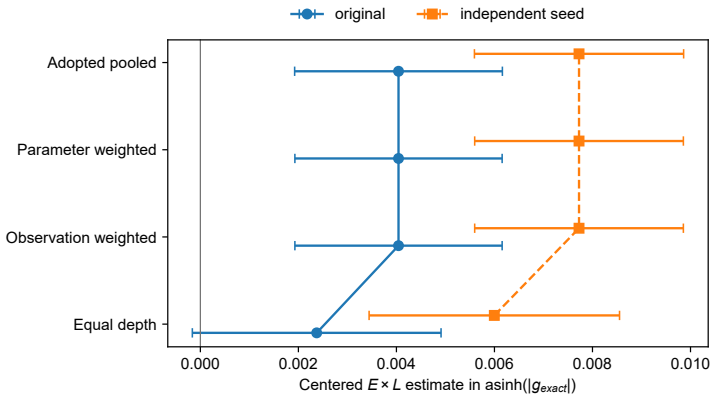

Fig. 3 Post-primary equal-depth, observation/matched-parameter, and adopted pooled H1 summaries. Bars are Wald 95% intervals. Weightings answer different questions and were not selected by significance

6 H3 post-primary explanatory analyses

Depth-specific L-averaged estimates at $D = 1, 2, 3, 4, 6$ were 0.006642, 0.000696, -0.026492, -0.011553, and -0.011255. Joint moderation did not reject with model-based ($p = 0.545$) or cluster-robust ($p = 0.721$) covariance. All 50 leave-one-initialization-out fits completed; all retained the negative averaged sign, 41 had raw $p < 0.05$, and estimates ranged from -0.014433 to -0.008562. These diagnostics do not justify deletion. The frozen classification is a model-dependent directional signal, objective-specific under end-to-end mode.

Contrast or sensitivity	Estimate	Wald 95% CI	Bootstrap 95% CI
End-to-end, $L = 0$	-0.000958	[-0.015251, 0.013336]	[-0.029117, 0.024560]
End-to-end, $L = 1$	-0.022273	[-0.036493, -0.008052]	[-0.049490, 0.007197]
End-to-end, L-average	-0.011615	[-0.021697, -0.001534]	[-0.031375, 0.009563]
Active depths $D \geq 3$, L-average	-0.014788	[-0.024480, -0.005095]	Not planned
Conditional mode, L-average	0.006978	[-0.004745, 0.018701]	Not available

Table 4 Post-primary H3 explanatory and sensitivity results. The 1,000-fit bootstrap intervals contain zero for both objective-specific contrasts and their average. Active-depth and conditional-mode rows are not substitutions for the primary model.

7 Conditional exact-gradient indices

The frozen descriptive indices are

$$J_{EL|R=0} = \frac{G_{EL}G_0}{G_E G_L}, \qquad J_{EL|R=1} = \frac{G_{ELR}G_R}{G_{ER}G_{LR}}.$$

Unique exact-gradient parameter rows are pooled before the configuration-level RMS construction, so deeper depths receive more observation/parameter weight. The indices are conditional on R, not transformations of centered coefficients.

Dataset	Estimand	Point	Median	Percentile 95% CI
Original	$J_{EL\|R=0}$	1.241760	1.238674	[1.082210, 1.414864]
Original	$J_{EL\|R=1}$	1.126633	1.124329	[0.991331, 1.285152]
Independent seed	$J_{EL\|R=0}$	1.163219	1.160949	[1.007100, 1.352596]
Independent seed	$J_{EL\|R=1}$	1.242137	1.238157	[1.049874, 1.449258]

Table 5 Conditional descriptive exact-gradient indices. Each initialization-cluster bootstrap completed 2,000 iterations with zero failures. Values above one are above the multiplicative no-interaction reference, not percentage improvements in optimization.

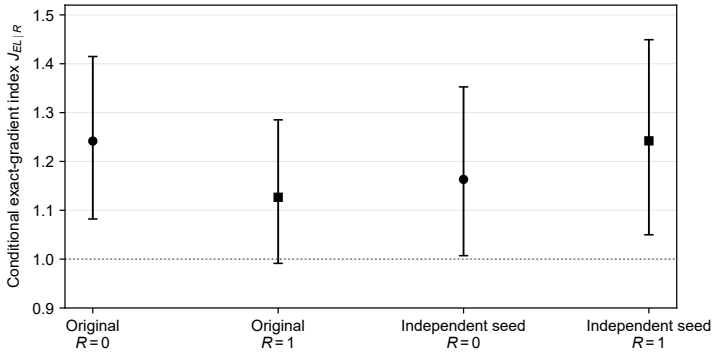

Fig. 4 All four frozen conditional $J_{EL|R}$ estimates with percentile 95% bootstrap intervals and reference line $J = 1$. Marker shape identifies R and line style identifies dataset. These are descriptive secondary ratios

8 Prepared-state metrics at initialization

The frozen table contains 4,000 unique terminal-block states after exact-state duplication across shot budgets was removed. Fidelity and normalized-energy bounds, row uniqueness, energy spectral bounds, and fidelity–infidelity complementarity all passed. Original configuration–depth mean fidelity ranged approximately 0.0338–0.0817 and normalized energy 0.4631–0.5529; independent-seed ranges were 0.0322–0.0962 and 0.4561–0.5537. The ranges overlap strongly. Since L changes evaluation rather than preparation, paired $L = 0/L = 1$ prepared-state summaries are identical. No significance test or causal gradient–task relationship is inferred.

Dataset	Cell-mean fidelity range	Cell-mean normalized-energy range
Original	0.0338–0.0817	0.4631–0.5529
Independent seed	0.0322–0.0962	0.4561–0.5537

Table 6 Descriptive ranges across configuration–depth cell means for terminal-block prepared states at initialization. The overlapping ranges do not represent optimization trajectories or a test of gradient–task causality.

9 Measurement implementation and resource accounting

Global infidelity is sampled directly from exact statevector overlap probability as one abstract overlap setting; no inverse target-state circuit, preparation gates, or physical cost is represented. TFIM energy uses a shared Z-basis multinomial setting for three ZZ terms and a shared X-basis multinomial setting for four X terms. End-to-end mode samples $D-1$ forward-feature Z jobs and re-encodes noisy features; conditional mode holds those features exact. Modes are not pooled.

All 320 resource rows conserve requested B exactly, allocations are deterministic and nonnegative, and no production job has zero shots. All 1,833 zero-variance cells join to complete resource records. At $D = 6$, conditional/global uses 48 jobs and end-to-end/local 61, a 27.08% increase. The experiment matches total implemented simulator shots, not circuits, jobs, shots per circuit, physical gates, calibration, readout, hardware noise, or wall-clock time.

10 Eligibility and zero-variance resource join

Exactly zero replicate variance was restricted to $L = 0$: 509 of 102,400 end-to-end and 1,324 of 102,400 conditional cells, with none under $L = 1$. The primary finite-shot estimand therefore conditions on positive estimated variance. The resource join demonstrates that these exclusions do not reflect missing accounting records. They remain an objective-linked selection limitation.

11 Implementation validation and model diagnostics

The black-box finite-difference audit tested 16 initialization–configuration–depth cases, configurations 1, 2, 3, and 5 at $D = 1, 6$, and all 224 associated quantum parameters. Central-difference steps were 10^{-3}, 10^{-4}, and 10^{-5}. At 10^{-5}, maximum relative error was 1.27×10^{-5} overall and 9.77×10^{-7} for the 214 gradients with magnitude at least 10^{-5}, against a 10^{-6} clean-subset tolerance. Four near-zero gradients exceeded the relative tolerance but agreed to six significant figures in absolute value; there were zero sign reversals, order-of-magnitude discrepancies, or failures to improve from 10^{-3} to 10^{-4}.

The adopted end-to-end H2–H4 model used 101,891 observations in 50 initialization clusters and 3,200 nested initialization–depth–parameter levels. L-BFGS converged without fallback; one `ConvergenceWarning` indicated a possible boundary solution,

but the 10^{-8} variance-component rule did not flag singularity. A forced BFGS refit also converged and agreed on target coefficients to 8–9 significant figures. All 50 leave-one-initialization-out refits converged. Residual SD ranged from 0.659 at $D = 1$ to 0.276 at $D = 6$ and from 0.301 at $B = 250$ to 0.513 at $B = 2000$; 1.14% of standardized residuals exceeded magnitude three. Convergence does not validate homoscedastic or normal covariance assumptions.

Matched $R = 0$ and $R = 1$ gradients are exactly equal at $D = 1, 2$ because the shortcut has not yet reached a real input; the structural check passed exactly. Active-depth, conditional-mode, categorical-depth, and influence analyses remain post-primary explanations rather than changes to H1–H4.

12 Reader-facing outcome index

Frozen configuration–depth–budget summaries contain the median and IQR of $\widehat{\mathrm{SNR}}_{\mathrm{est}}$, the fraction with $\widehat{\mathrm{SNR}}_{\mathrm{est}} \geq 1$, exact-gradient SNR, signed and absolute bias, sign agreement, and exact-gradient landscape variance in `results/production_confirmatory/configuration_summaries.csv`. Prepared-state fidelity, energy, entanglement entropy, and reduced-state purity are summarized here and in `results/task_metrics/`. Classical gradients are excluded from H1–H4, and no substantive classical-gradient conclusion is claimed. Finite-difference validation, optimizer diagnostics, estimator-mode guards, influence analysis, and resource accounting are summarized above and traceable through `verification/`.

13 Historical correction and audit record

Historical direct-0/1 coefficient tables and older pooled estimator-mode outputs are retained in explicitly superseded directories for auditability and are not current primary factorial conclusions. In particular, the former H3 value near -0.000958 is the end-to-end $E \times R$ simple interaction at $L = 0$; it now appears as a post-primary explanatory contrast, not the centered H3 estimand. Historical `n=400` H1 bootstrap text is superseded by 2,000 completed fits, and the former H4 Holm $p = 0.115$ is superseded by the corrected family value 0.057658.

14 Reproducibility index

All frozen inputs, numerical outputs, figures, and verification reports are indexed by `results/final_submission_v1/manifest.json`. The numerical-results freeze submission-numerical-results-freeze-v1 preserves the frozen scientific outputs. The accompanying OSF project is available at https://doi.org/10.17605/OSF.IO/THV6G. The separate final manuscript/repository release `sndc-submission-v1` contains the finalized manuscript sources, supplement sources, documentation, verification scripts, and submission-facing repository state. `MANUSCRIPT_COMMIT.txt` records the exact substantive commit against which those materials were finalized. The principal materials are grouped under:

- `results/primary_corrected/` and `results/independent_seed_h1/` for corrected primary and independent-seed outputs;
- `results/h1_depth_weighting/`, `results/h3_centered_robustness/`, and `results/jel_conditional/` for post-primary analyses;
- `results/task_metrics/` and `results/resource_accounting/` for descriptive task and measurement-resource outputs; and
- `verification/` for plans, checksums, audit records, and machine-readable result reports.

15 Remaining limitations

Limitations include the four-qubit, single-task, reset-per-block setting; one implementation per intervention and one initialization family; 50 top-level clusters per dataset; exact-state and abstract shot simulation without hardware noise, readout, calibration, or optimization trajectories; initialization-only prepared-state metrics; L-linked zero-variance selection; parameter-count-induced depth weighting; covariance-sensitive depth conclusions; H2/H3 disagreement across inference procedures; H3 estimator-mode disagreement; weighting-dependent conditional J; absence of a locally archived full protocol source; finite Monte Carlo uncertainty in the 1,000-fit percentile endpoints; and no matched physical-resource claim.

Appendix A Protocol and implementation interpretation

A.1 Factor-coding provenance

The accessible protocol materials fixed the factorial interventions, hypotheses, transformations, model families, and Holm procedure but did not specify direct 0/1 versus centered $\{-1/2, +1/2\}$ coding. The original implementation used direct 0/1 predictors and incorrectly interpreted lower-order pairwise coefficients as averages over the third factor. Re-expression with centered columns leaves fitted values, likelihood, residuals, and random effects unchanged while correcting the estimands; historical direct-0/1 coefficients remain archived as reference-level simple interactions.

A.2 Reset-per-block versus register-persisting semantics

The protocol prose permits reset-per-block and register-persisting readings. This study adopts reset per block: each quantum block begins from $|0000\rangle$, is measured independently, and passes only classically re-encoded features forward. Accordingly, D counts repeated quantum evaluations rather than layers of one continuously evolving register. All code, gradients, shot accounting, and results use this interpretation. A register-persisting implementation would change state preparation, entanglement propagation, derivative assembly, and measurement allocation and would therefore constitute a materially different experiment not adjudicated here.