UNDER REVIEW

RESCIENCE C

Replication / Quantum Computation and Statistical Physics

[Re] Partial Replication of the Willow Fano-Factor Results in "The Rotation Gap Is Not An Error"

Ivan Nestorov[1, ID]
[1]VolMax Studio Lab, Belgrade, Serbia

Edited by
(Editor)

Received
–

Published
–

DOI
–

Abstract

We present a pre-registered partial replication of the empirical quantum error-correction detector statistics reported in Section II F and Table III of Stenberg (2026, arXiv:2604.11963v1) [1]. Analyzing raw detector bitstream records from the 105-qubit Google Willow archive (Zenodo DOI: 10.5281/zenodo.13273331 [2], 420 experiments across code distances $d \in \{3, 5, 7\}$), we recompute all four primary and secondary claims under strict cryptographic and algorithmic verification gates. We confirm that: (1) the aggregate empirical Fano factor reproduces to $F = 2.4157 \approx 2.42 \pm 0.36$ (Claim W_1); (2) distance-stratified Fano factors reproduce to $d = 3 \rightarrow 2.2942$, $d = 5 \rightarrow 2.5935$, and $d = 7 \rightarrow 2.7978$ (Claim W_2); (3) the one-way ANOVA across distances yields $F = 59.13, p = 2.46 \times 10^{-23}$ (Claim W_3); and (4) the one-sample t-statistic against Poisson expectation ($F = 1$) reproduces to $t = +80.15$ (Claim W_4). We explicitly document our three-layer mixed implementation provenance: the primary detector decoding path is independently implemented (`read_detection_counts_fast`), while the statistical estimator on the analytical path deliberately transcribes the author's published operations under the preregistered design, with reference author decoding operations retained off-path for 10/10 exact bitwise equivalence verification. Finally, we document two epistemic limitations: a structural zero-power defect in the preregistered QEC round-sensitivity test resulting from Google's balanced experimental design ($N_r = 28$), and an exploratory finding demonstrating that equal-distance weighting shifts the recomputed mean Fano factor from 2.4157 to 2.5618 ($|\Delta| = 0.1461$, with the corresponding source-reported aggregate being 2.42) due to the predominant representation of $d = 3$ experiments (64.3%) in the public repository.

1 Introduction

In quantum error correction (QEC) with surface codes, syndrome measurements are performed over repeated rounds of stabilizer checks. The detection events (changes in syndrome eigenvalues indicating possible Pauli faults) are expected to follow approximately Poissonian or Bernoulli-like statistics under independent Markovian noise. Recently, Stenberg (2026) [1] analyzed open telemetry from Google's 105-qubit Willow architecture [2] and reported super-Poissonian detector event variance, quantified via the Fano factor:

$$F = \frac{\sigma^2}{\mu} > 1, \tag{1}$$

reporting an aggregate mean of $F = 2.42 \pm 0.36$ across $N = 420$ surface-code experiments, with monotonic scaling across code distances $d \in \{3, 5, 7\}$ and statistical rejection of Poisson variance ($t = +80.0$).

Correspondence should be addressed to Ivan Nestorov (volmax.core@gmail.com)
The authors have declared that no competing interests exist.
Code is available at https://github.com/VolMax-Studio/rotation-gap-fano-s1 – DOI 10.5281/zenodo.22309778..
Data is available at https://doi.org/10.5281/zenodo.13273331 – DOI 10.5281/zenodo.13273331.

The purpose of this study is to perform a pre-registered partial replication and deterministic audit of these empirical Willow detector statistics (Claims W_1–W_4).

1.1 Scope and Replication Boundaries

This replication specifically targets the four empirical claims stated in Section II F and Table III of Stenberg (2026) [1]:

- **Claim W_1 (Aggregate Fano Factor):** Overall empirical mean $F = 2.42 \pm 0.36$ across all admitted Google Willow experiments ($N = 420$).
- **Claim W_2 (Distance Monotonicity):** Stratified distance means $d = 3 \to 2.29$, $d = 5 \to 2.59$, and $d = 7 \to 2.80$.
- **Claim W_3 (Distance Variation ANOVA):** One-way ANOVA across code distances yielding $F \approx 59.1$ ($p \approx 0$).
- **Claim W_4 (Poisson Rejection t-test):** One-sample t-test against $F = 1$ yielding $t = +80.0$.

Explicit Non-Scope: This study does *not* evaluate logical error rate scaling, rotation gap hypotheses, ternary state conjectures, IBM Quantum hardware runs, or theoretical physics assertions regarding the origin of non-Markovian noise. It is an algorithmic and statistical audit of the public Google Willow dataset under the author's published analytical pipeline.

2 Implementation Provenance and Architecture

To ensure complete transparency regarding software provenance, our replication codebase [3] is organized into three distinct operational layers, separating independent execution infrastructure from author-derived conventions and analytical operations:

> *"The primary detector decoding path is independently implemented; the statistical estimator on the analytical path deliberately transcribes the author's published operations under the preregistered design."*

The primary detector reading pipeline in our audit (`read_detection_counts_fast`) uses NumPy's vectorized byte unpacking (`np.unpackbits(..., bitorder="little")`), achieving a $> 30\times$ speedup over byte-by-byte bit-shifting loops. Under the validation flag `-verify-popcount`, Layer C runs the reference implementation across 10 distinct experiments, confirming 10/10 exact integer identity ($\Delta = 0$).

3 Preregistered Verification Pipeline

The replication was conducted according to a preregistration protocol (`PREREGISTRATION.md`, commit `ff84608`) committed prior to reading archive data.

3.1 Premise and Invalidation Gates (P_1–P_5)

Before computing statistical summaries, the harness evaluates five preregistered premise conditions defined at Step 0:

- **P_1 (Experiment Count and Distance Span, Halting):** The archive at DOI `10.5281/zenodo.13273331` contains exactly 420 surface-code experiments spanning $d \in \{3, 5, 7\}$.

Provenance Layer	Component Description	Implementation Origin
Layer A	Archive integrity pinning (MD5, size) and `PREREGISTRATION.md` SHA-256 pin	Independent Implementation
(Independent	Cryptographic verification of `data_manifest.json` (1,260 member digests)	Independent Implementation
Infrastructure)	Step 0 premise checking harness and halting invalidation gates (P_1–P_5)	Independent Implementation
	Fast vectorized popcount decoder (`read_detection_counts_fast`)	Independent Implementation
	Mechanical decision logic and tolerance rules (R_1–R_{10})	Independent Implementation
	CI harness and deterministic byte-identical JSON reproduction engine	Independent Implementation
Layer B	Archive layout traversal and path convention (`enumerate_experiments`)	Author Convention (`willow_fano_analysis.py`)
(Author-Derived	Detector count parsing from Stim circuit string (`count_detectors`)	Author Convention (`willow_fano_analysis.py`)
on Analytical	Sample variance with Bessel correction (`ddof=1`)	Transcribed from Author Script (§2)
Path)	Per-experiment Fano factor computation ($F_i = s_i^2/\bar{x}_i$)	Transcribed from Author Script (§2)
	Aggregate mean ($\bar{F}$) and population standard deviation of F_i (σ_F)	Transcribed from Author Script (§2)
	One-sample t-statistic against $F = 1$ ($t = (\bar{F} - 1)/(\sigma_F/\sqrt{N})$)	Transcribed from Author Script (§2)
	Distance-stratified means and one-way ANOVA (F-statistic, p-value)	Transcribed from Author Script (§2)
Layer C	Reference bit-shift popcount loop (`read_detection_counts_author`)	Transcribed from Author Script
(Off-Path Only)	Executed exclusively under `–verify-popcount` for bitwise equivalence	Reference / Invariant Check

Table 1. Three-layer implementation provenance architecture of the replication audit.

- P_2 **(Unprocessed Count Recoverability, Halting):** Per-shot detection-event counts are recoverable from the archive with no preprocessing beyond decoding the packed detector records (`detection_events.b8`).
- P_3 **(Underlying Cohort Identity, Halting):** The same 420 experiments underlie W_1, W_2, W_3, and W_4.
- P_4 **(Missing Author Output Artifact, Non-Halting):** The author's repository contains no committed output file from `willow/willow_fano_analysis.py` (condition measured as **Falsified-as-Predicted** at preregistration freeze on 2026-09-03).
- P_5 **(Archive Byte Identity, Non-Halting):** The archive bytes are identical to those used in `willow-decoder-audit-s1`.

All three halting premises (P_1, P_2, P_3) evaluated to **PASS**, allowing execution of the analytical path without triggering the Premise Invalidation Gate.

3.2 Decision Rules and Pre-Registered Tolerances

The primary claim W_1 is evaluated under rules R_1–R_4 from `PREREGISTRATION.md @ ff84608`:

- R_1 **(Verified):** $|\bar{F}_{\text{recomputed}} - 2.42| < 0.005$ (0.005 is the half-width of the last printed digit).
- R_2 **(Verified with Limitations):** $0.005 \leq |\bar{F}_{\text{recomputed}} - 2.42| \leq 0.36$ (0.36 is the reported spread used as the outer admissible band).
- R_3 **(Not Verified):** $|\bar{F}_{\text{recomputed}} - 2.42| > 0.36$.
- R_4 **(Halting Invalidation):** P_1, P_2, or P_3 falsified $\implies$ halt execution, no verdict issued.

Secondary claims W_2–W_4 carry independent decision rules:

- R_5 **(W_2 Distance Stratification):** All three per-distance values reproduce within $|\bar{F}_d - F_{d,\text{reported}}| < 0.005 \implies W_2$: **Verified** (failing ± 0.05 yields R_6 : **Not Verified**).
- R_7 **(W_3 ANOVA F-Statistic):** Recomputed ANOVA F-statistic within ± 0.5 of 59.1 $\implies$ W_3 : **Verified**.
- R_8 **(W_4 t-Statistic):** Recomputed t-statistic within ± 2.0 of $+80.0 \implies W_4$: **Verified**.

4 Replication Results

All numerical results were computed deterministically from raw detector bitstreams (`detection_events.b8`) using `reproduce.py` and archived in `results/results.json`. Cryptographic integrity of all 1,260 consumed member files was verified against the committed `data_manifest.json`.
As shown in Table 2, every targeted statistic falls within the pre-registered tolerance bounds. The minor differences ($|\Delta| \leq 0.0043$ for Fano factor means) are consistent with the precision at which the source values were reported in the original manuscript's text and tables.

5 Epistemic Findings and Limitations

Beyond verifying arithmetic correctness, our audit identified two methodological and epistemic nuances:

Claim	Quantity	Reported	Recomputed	Absolute Delta $\|\Delta\|$	Verdict / Status
W_1	Aggregate Mean F	2.42	2.4157	0.0043	Verified (Rule R_1)
W_1	Population Std. Dev. of F_i (σ_F)	0.36	0.3620	0.0020	(Descriptive / Part of W_1)
W_2	Distance $d = 3$ Mean	2.29	2.2942 ($N = 270$)	0.0042	Verified
W_2	Distance $d = 5$ Mean	2.59	2.5935 ($N = 120$)	0.0035	Verified
W_2	Distance $d = 7$ Mean	2.80	2.7978 ($N = 30$)	0.0022	Verified (Rule R_5)
W_3	One-way ANOVA F	59.1	59.1337	0.0337	Verified (Rule R_7)
W_3	One-way ANOVA p-val	≈ 0	2.4624×10^{-23}	—	(Supporting / Part of W_3)
W_4	One-sample t-stat	+80.0	+80.1510	0.1510	Verified (Rule R_8)

Table 2. Summary of numerical replication results against reported claims in Stenberg (2026). Claim verdicts are governed strictly by the mechanical decision rules R_1–R_8 from `PREREGISTRATION.md @ ff84608`.

5.1 Structural Zero-Power in Round Sensitivity Test

Rules R_9 and R_{10} were designed to test sensitivity to QEC round duration by comparing native weighting against a round-uniform weighting $F_{\text{uniform, rounds}}$. The observed discrepancy was exactly zero:

$$|F_{\text{uniform, rounds}} - F_{\text{native}}| = 0.000000. \tag{2}$$

Analysis of the raw archive revealed that Google's experimental design is perfectly balanced: exactly $N_r = 28$ experimental configurations exist for every round count $r \in \{1, 10, 13, 30, 50, 70, 90, 110, 130, 150, 170, 190, 210, 230, 250\}$. Consequently, the two weighting schemes are algebraically identical over this dataset. The test passed formally due to experimental design symmetry rather than demonstrating empirical round invariance.

5.2 Distance Composition Sensitivity

In the Google Willow public repository, distance $d = 3$ accounts for 64.29% of all experiments (270/420), $d = 5$ for 28.57% (120/420), and $d = 7$ for only 7.14% (30/420). If equal weighting is assigned to each code distance (1/3 per distance group), the aggregate mean shifts:

$$\bar{F}_{\text{distance-uniform}} = \frac{2.2942 + 2.5935 + 2.7978}{3} = 2.5618, \tag{3}$$

a shift of $|\Delta| = 0.1461$ (+6.0%). Thus, the headline aggregate value $F \approx 2.42$ strongly reflects the abundance of $d = 3$ runs in the public archive rather than an equal-distance hardware average.

6 Conclusion and Code Availability

We have replicated Claims W_1–W_4 from Section II F and Table III of Stenberg (2026) [1]. The super-Poissonian Fano factors, distance monotonicity, ANOVA significance, and t-test statistics were recomputed within the pre-registered numerical tolerances from the public archive.

The complete audit code, verification harness, and results are open source under the MIT License and permanently archived:

- **Software Repository:** https://github.com/VolMax-Studio/rotation-gap-fano-s1
- **Software Archive DOI:** 10.5281/zenodo.22309778 [3]
- **Underlying Dataset DOI:** 10.5281/zenodo.13273331 [2]

References

1. S. Stenberg. "The Rotation Gap Is Not An Error — Ternary Structure in IBM Quantum Hardware." In: **arXiv preprint arXiv:2604.11963v1** (2026). URL: https://arxiv.org/abs/2604.11963.
2. Google Quantum AI and Collaborators. **Quantum Error Correction on the 105-Qubit Willow Architecture: Surface-Code Telemetry and Detector Event Datasets**. Zenodo, 2024. DOI: 10.5281/zenodo.13273331. URL: https://doi.org/10.5281/zenodo.13273331.
3. I. Nestorov. **rotation-gap-fano-s1: Pre-registered Reproduction and Deterministic Verification of Google Willow Fano Factor Claims**. Version v1.0.0. 2026. DOI: 10.5281/zenodo.22309778. URL: https://doi.org/10.5281/zenodo.22309778.