Supplementary Information for

Auditable Release Assurance for Quantized LLM Deployment Configurations: Anytime-Valid Non-Inferiority and Executable Certificates

Software Quality Journal

Myunglae Jo[a], Donghyun Kim[b], Changhan Son[b],
Minseong Park[b], Il-Hyuk Yu[b,c], and Hyungkwang Lim[b,d,*]

[a] Department of Physics, Kyungpook National University, Daegu 41566, Republic of Korea

[b] Palusomni Co., Ltd., Seoul 06628, Republic of Korea

[c] Xivilization Co., Ltd., Fukuoka 810-0001, Japan

[d] Advanced Institute of Science and Technology (AIST), Kyungpook National University, Daegu 41566, Republic of Korea

* Corresponding author: Hyungkwang Lim; lim@palusomni.com

Online Resource 1

Artifact and Reproduction Details

This resource records the command-level implementation, protocol compatibility, fixture catalogue, and reproduction contract that are deliberately kept out of the main research narrative. It introduces no new experiment or scientific claim. Unless stated otherwise, every path is relative to the extracted root of `Online_Resource_2_Code_Data_and_Certificate_Artifact.zip`.

1 Artifact scope and evidence levels

The archive contains the Protocol 3.1 schemas, decision engine, issuer and verifier, synthetic positive and negative fixtures, deterministic analysis drivers, retained correctness records, authoritative derived arrays, and figure source data. Full-scope Protocol 3.1 fixtures are synthetic mechanics tests; the measured certificate example is a legacy retrospective replay. Neither one proves prospective physical execution.

The public artifact distinguishes exactly three evidence levels:

1. **certificate-decision replay**: the supplied verifier recomputes the reference implementation's contract and decision checks;
2. **paired-correctness reconstruction**: complete retained pairs support reconstruction for eight populations, while authoritative derived arrays support downstream replay for four populations; and
3. **original-generation re-parsing/re-scoring**: not supported by the sanitized public artifact.

The public artifact supports independent reconstruction of paired degradation from retained correctness records, but not independent re-parsing or re-scoring of the original generations. Free-text generations are not included. They should not be restored unless the authors separately confirm both redistribution rights and scientific necessity.

The supplied software is a verifier that can be run by an independent third party to replay the reference implementation's checks. The verifier is not a second, independently developed implementation of the statistical procedure.

2 One default reviewer command

After extracting the archive into an empty directory, create an isolated environment, install the pinned direct review dependencies, and run:

```
python3 -m venv .venv
.venv/bin/python -m pip install -r requirements.txt
PATH="$PWD/.venv/bin:$PATH" make review-checks
```

The default target executes, in order: the legacy protocol regression suite; the focused Protocol 3.1 regression suite; the assurance-layer responsibility cases; schema validation; the factorial, without-replacement, and fixed-baseline drivers;

the affected-output rebuild check; output-consistency checks; and retained-figure hash checks. It finishes with `V33 REVIEW CHECKS COMPLETE` only when every component returns zero.

The exact root-level schema command is:

```
python3 cert/validate_schemas.py
```

Schema paths are resolved from `cert/validate_schemas.py`, not from the caller's working directory. Thus the command above is the documented artifact-root route, and the script also remains path-independent if invoked by absolute or relative script path from another working directory.

For the command-line interface:

```
python3 cert/certify.py --help
```

3 Protocol and schema mapping

Object or action	File	Purpose
Canonical CLI	`cert/certify.py`	Protocol 3.1 commit, select, finalize, issue, and verify entry point
Protocol implementation	`cert/certify_v31.py`	Closed schema/semantic validation and issuer/verifier orchestration
Decision engine	`code/decision_engine_v31.py`	Exact rational risk values and outward-rounded Decimal intervals
Stage 1 commitment	`cert/protocol_commitment_v31.schema.json`	Configuration, evidence, candidate, chronology, order mode, and exact risk commitments
Selection record	`cert/selection_record_v31.schema.json`	Selected candidate fixed before fresh Stage-2 randomness
Stage 2 object	`cert/stage2_v31.schema.json`	Public-randomness claim and finalized order binding
Certificate	`cert/certificate_v31.schema.json`	Evidence bindings, assurance result, decision basis, and signature
Deployment profile	`cert/deployment_configuration_profile_v1.schema.json`	Closed deployment-configuration facts
Legacy replay	`cert/certify_v29_legacy.py`, `cert/certify_v30.py`	Explicitly separated historical compatibility routes

The validator also checks the legacy `cert/certificate.schema.json`, the v30 commitment/selection/Stage-2/certificate schemas, all v31 schemas above, and every retained deployment-profile fixture. Schema success checks JSON Schema conformance; semantic and cryptographic conditions remain the responsibility of the issuer/verifier tests.

4 Assurance-layer responsibility analysis

The deterministic driver `code/assurance_layer_responsibility.py` constructs a fixed Ed25519 test key from a declared seed, uses the shipped synthetic fixtures, and executes the actual Protocol 3.1 command line. Its output is `data/populations/assurance_layer_responsibility.json`.

Case	Statistics alone	Signature/file integrity alone	Full Protocol 3.1	Named reason
Valid committed workflow	Pass — executed	Pass — executed	Pass — executed	Commit, issuance, and verifier replay all succeed
Evaluator-source mutation	Pass — executed	Detected — executed	Reject — executed	Evaluator bytes no longer match the signed Stage-1 hash
Post-order candidate selection	Not checked by this layer	Pass — executed	Reject — executed	Adaptive R2 holdout with $K \geq 2$ prohibits local-order mode fixed before selection
Inexact risk allocation	Not checked by this layer	Pass — executed	Reject — executed	Rounded `0.016667` is not exactly $0.05/3 = 1/60$; Stage-1 semantics fail

This is a responsibility classification, not a throughput, accuracy, or product-performance ablation. For the post-order-selection and inexact-allocation cases, "not checked" indicates that the isolated layer contains no corresponding validity check; the full-protocol rejection was executed. Outside all three layers are the truth of physical execution, the truth of external timestamps or public-randomness claims, and the absence of hidden candidate pre-screening.

5 Deterministic auxiliary-analysis drivers

Each driver embeds its authoritative relative input path and SHA-256, random seed, replicate count, calibration/null grid, exact RNG consumption order, confidence- interval convention, output schema, and output path. Without `-write`, a driver recomputes its result and fails if the retained JSON differs byte-for-byte. The retained results use these declared reproduction rules.

5.1 Factorial intervals

`code/derive_factorial_cis.py` reads these four jointly keyed 500-item records:

Input	SHA-256
`data/runs/run_q3-1.7b-bf16_jetson_mmlu.jsonl`	`88c7c726375ae2c5b3db9cb6264a3ce77a61152de8a6882970db4ce608d67bcb`
`data/runs/run_q3-1.7b-q4km_jetson_mmlu.jsonl`	`afabb0553cd46579b8f0212fed9dfe86d93f40a1bd89d7e3fbdc4bc3292432f1`
`data/runs/run_q3-1.7b-bf16_jetsoncpu_mmlu.jsonl`	`b65a960195b979ea46690fcee7490198a9c9863a4786cc1b875ba026eba8e3ad`
`data/runs/run_q3-1.7b-q4km_jetsoncpu_mmlu.jsonl`	`5bf2d1b7e082a22313610c823259d1f9fa93624636fda9baf79385912b1f5437`

The seed is 20260906 and the replicate count is 20,000. Replicates are consumed in index order from one NumPy `default_rng`; each draws 500 indices with replacement, and the same index vector is used for all four cells. Equal-tail 95% percentiles use `numpy.quantile(method='linear')`; manuscript endpoints are rounded to 0.1 percentage point. Agreement uses two-sided 95% Wilson intervals. There is no calibration/null grid. Output `data/populations/factorial_cis.json` follows `FactorialCIsV32`.

The quantization effect is +2.9 percentage points with 95% interval $[-0.1, +5.8]$; backend and interaction summaries and all affected prose use the same JSON.

5.2 Without-replacement instrument comparison

`code/derive_wor_baseline_comparison.py` reads `data/populations/populations.json`, SHA-256 `f5759949a6c58ea1a5ee5958a2e5b1ba96446a4840f92d0a7cf285a01e14089e`. The seed is 20260906 and each cell

uses 400 permutations. The declared grid uses $\varepsilon = 0.03$, $\delta = 0.05$, degradation/margin ratios 0.33, 0.67, and 1.0, with 1.0 the null boundary, four named populations in source order, and the betting- mixture, Hoeffding, and empirical-Bernstein-style methods in source order. One RNG stream performs any controlled relabeling before the 400 permutations; all three methods share each permutation and consume no RNG. Certification intervals are two-sided 95% Wilson intervals; stopping quantiles use linear interpolation and censor non-certification at $N + 1$. Output `data/populations/wor_baseline_comparison.json` follows `WoRBaselineComparisonV32` with 36 rows.

5.3 Fixed and net-count comparison

`code/derive_fixed_baseline_comparison.py` uses the same population file and hash. Its seed is 20260906. The null calibration grid uses $\varepsilon = 0.03$, $\delta = 0.05$, sample sizes 300 and 500, and beneficial-flip probabilities $0, 0.01, 0.025, 0.05, 0.10, 0.20, 0.30, 0.40, 0.48$, with harmful probability equal to beneficial probability plus ε. Calibration uses 50,000 multinomial replicates per grid cell; each of four analysis cases uses 2,000 with-replacement samples. One RNG stream consumes the $n = 300$ grid, then $n = 500$, then the named analysis cases and ratios. All four statistics share each sample. Analysis-rate intervals are two-sided 95% Wilson intervals. The finite-grid net-count threshold is explicitly exploratory, not a certified level test. Output `data/populations/fixed_baseline_comparison.json` follows `FixedBaselineComparisonV32`.

Run the drivers separately with:

```
python3 code/derive_factorial_cis.py
python3 code/derive_wor_baseline_comparison.py
python3 code/derive_fixed_baseline_comparison.py
python3 code/assurance_layer_responsibility.py
python3 code/regenerate_readable_figures.py
```

The final command is presentation-only: it verifies the retained CSV hashes and regenerates manuscript Figures 2 and 5 at approximately the final text width. It does not rerun an experiment or rewrite source data.

6 Protocol versions and compatibility

The version labels identify separately replayable artifact contracts.

Legacy baseline. Historical schemas, certificates, fixtures, and replay logic remain separately available under the legacy route; legacy replay does not establish real-world execution provenance.

Legacy compatibility implementation. The v30 implementation enforces selection-before-fresh-randomness chronology, mutually exclusive local/public order modes, closed deployment-profile schema and semantics, and bindings to the actual task, evaluator, frames, selection evidence, and ordered records. Certificate validation rejects malformed serialization, records, signatures, parent/ledger relations, and order claims. The retained regression is `tests/test_v30_protocol.py`.

Protocol 3.1 implementation. Risk allocations use reduced numerator/denominator pairs. Issuance and replay share one outward-rounded Decimal interval engine and abstain when the boundary cannot be separated. Deterministic, wealth, and census R2 states are evaluated at the end of every observed prefix. The focused regression is `tests/test_v31_protocol.py`.

Reproducibility utilities. Schema validation is path-independent. The responsibility driver, three declared auxiliary drivers, and bounded checks are available through `make review-checks`.

6.1 R2 prefix-boundary and numerical decision behavior

Protocol 3.1 recognizes deterministic null impossibility after consuming the 499th record without requiring look-ahead. Median stopping times are 167, 202, and 423; the near-boundary stopping time is 499, with conservative upper bound 0.028008 at that prefix. The theorem uses ideal arithmetic; simulation and plotting use binary64 only for non-authoritative diagnostics. Exact rationals and outward-rounded Decimal intervals govern certificate decisions, with numerical abstention if strict separation is unresolved.

7 Fixture-migration and exhaustive negative catalogue

The v29 and v30 objects are retained with their original schemas and explicit legacy entry points. Protocol 3.1 fixtures live in `cert/fixtures/v31/`; they add exact risk objects, complete deployment profiles, boundary sequences, and synthetic inputs for local, Bonferroni, and holdout flows. Legacy certificates retain their original protocol identities.

Category	Representative negative condition	Required result
Risk representation	non-reduced fraction, zero denominator, rounded 0.05/3 surrogate	structured refusal
Decision boundary	Decimal interval touches the reciprocal-risk threshold at maximum precision	numerical abstention
Configuration	missing, unknown, or ill-typed deployment-profile fact	invalid/refusal
Evidence binding	omitted or mutated task, evaluator, frame, selection, profile, or record bytes	invalid/refusal
Candidate selection	relabeled candidate, wrong winner, reused holdout, or post-order selection	invalid/refusal
Order mode	mixed local/public fields, forged Stage 2, false parent, retry-like chronology	invalid/refusal
Serialization	duplicate JSON member, non-finite number, malformed top level, noncanonical signed bytes	invalid/refusal
Record stream	Boolean degradation, empty stream, wrong item/order, invalid prefix length	invalid/refusal
Signature/key	changed payload, forged signature, or untrusted key	invalid
Ledger/family	overspent allocation, sibling collision, future parent, or release-family mismatch	invalid/refusal

`cert/run_negative_fixtures.py`, `tests/test_v30_protocol.py`, and `tests/test_v31_protocol.py` enumerate the concrete retained fixture names. A negative test passes only when it returns the expected named invalid/refusal or abstention state; a crash is a failure.

8 Data provenance

Population	Evidence class	Available replay
Qwen3-1.7B Q8 / MMLU	retained full pair	paired-correctness reconstruction and downstream replay
Qwen3-1.7B Q4 / MMLU	retained full pair	same
Qwen3-1.7B Q8 / GSM8K	retained full pair	same; stop metadata retained
Qwen3-1.7B Q4 / GSM8K	retained full pair	same; stop metadata retained
Llama-3.2-1B Q8 / MMLU	retained full pair	paired-correctness reconstruction and downstream replay
Llama-3.2-1B Q4 / MMLU	retained full pair	same
Llama-3.2-1B Q8 / GSM8K	retained full pair	same; stop metadata retained
Llama-3.2-1B Q4 / GSM8K	retained full pair	same; stop metadata retained
Llama-3.2-3B Q8 / MMLU	derived array	downstream replay; historical reference absent
Llama-3.2-3B Q4 / MMLU	derived array	downstream replay; historical reference absent
Qwen3-8B Q4 / MMLU	derived array	downstream replay; historical reference absent
Qwen3-8B Q8 / MMLU	derived array	downstream replay; both raw sides absent

The package does not reconstruct every historical environment. The 11.8% parsed-prediction disagreement is an archived outcome-record comparison using a rebuilt Mac configuration; the exact historical Mac-producing environment was not preserved. A signed supplied candidate manifest does not prove the absence of hidden candidate pre-screening.

9 Environment and dependency record

The recorded reproduction checks passed on macOS 15.6.1 (build 24G90), arm64, with CPython 3.13.5. Direct pins were NumPy 2.4.6, SciPy 1.17.1, Matplotlib 3.11.1, cryptography 50.0.0, jsonschema 4.26.0, and rfc8785 0.1.4. The complete resolved `pip freeze` and commands are retained in `environment.txt`. These are pinned direct review dependencies, not a platform-resolved hash-locked, hermetic, or fully cross-platform environment.

The manuscript build used pdfTeX 1.40.29 and BibTeX 0.99e from TeX Live 2026. The flat manuscript source archive includes the official `sn-jnl.cls` and selected author–year `sn-mathphys-ay.bst`; the package manifest records their exact SHA-256 values.

10 Third-party material and trust boundary

The supplementary artifact excludes MMLU/GSM8K benchmark question and answer files, removes `out_head` and `out_full` generation fields, and contains no model weights, copied publication figures/tables, screenshots, or logos. It retains item identifiers and hashes, correctness and parsed outcomes, derived arrays, protocol code, tests, and drivers. Dependencies are named for separate installation; their source trees are not bundled. Original source code is copyright Palusomni Co., Ltd. and Xivilization Co., Ltd. and is licensed under the MIT License. Non-code data and documentation carry no separate open-content licence unless a file expressly states otherwise.

Signatures and bound hashes establish integrity of supplied bytes, not truth of physical execution, external timestamp or public-randomness claims, key ownership, or completeness of the unseen candidate universe. These are external trust inputs, not verifier conclusions.