ORION Speculative Decoding Lab: An Evidence-Aware Architecture, Prototype Evaluation, and Integration Roadmap

Aaron Singh

***Abstract*—This paper presents ORION Speculative Decoding Lab, a project in a twenty-project AI infrastructure portfolio. The current repository is an active prototype with implemented behavior, automated correctness tests, and explicit boundaries around unverified hardware or production claims. We describe the system model, current implementation, goals, and evaluation method, then propose five integrations that advance the project toward reproducible system-level validation. A local audit executed 5 project tests successfully. No accelerator, deployment, or performance conclusion is inferred unless a corresponding committed benchmark or profiler artifact exists.**

***Index Terms*—ORION, AI infrastructure, reproducibility, prototype validation, systems evaluation, hardware–software co-design**

I. Introduction

ORION Speculative Decoding Lab (P06) addresses a bounded problem within the portfolio's track-02-llm-serving track. Its declared status is *Active Prototype* [1]. The project validates its central mechanism before adding device-specific acceleration, production orchestration, or real-world hardware. This ordering matters because optimization without a trusted reference can make incorrect behavior appear successful.

The project goals are to establish deterministic behavior, encode correctness as tests, create machine-readable evidence, identify limiting resources through measurement, and integrate results with adjacent portfolio systems without losing provenance. The repository contains one paper named exactly after its singular folder: `P06-orion-speculative-decoding-lab.tex`.

II. Technical Context

Language-model serving couples execution with cache management, request scheduling, and latency objectives. A simulator can validate policy logic, but only a real model runtime and measured workload can support performance claims. The portfolio benchmark standard requires environment, workload, method, metric, artifact, reproduction, and limitation fields; unknown values remain explicitly unmeasured [3].

III. System Model and Architecture

The prototype is organized around the following domain model:

$$\mathbb{E}[A] = \sum_{j=1}^{K} \alpha^j, \quad \text{tokens per verifier round} = A + 1 \tag{1}$$

The equation is a design and test abstraction rather than a claimed empirical law. It supports invariants and expected-value checks while later implementations replace synthetic inputs with representative workloads or devices.

The software architecture contains an input/configuration layer, a deterministic core, validation and evidence output, and a local visualization. The principal inspected source artifacts are `python/orion_simulator.py`. Unsupported real-world conditions are surfaced as limitations rather than silently simulated.

IV. Detailed Script Operation and Rationale

The simulator generates a repeatable target token stream, creates draft blocks at a controlled acceptance probability, accepts matching tokens until the first mismatch, and lets the target supply the correction. It records verifier rounds and acceptance statistics without claiming wall-clock model speedup.

The execution path is:

1) Validate output length, draft width, probability, vocabulary, and seed.
2) Generate a deterministic target sequence.
3) Produce draft blocks with controlled agreement.
4) Accept the matching prefix and correct the first mismatch with the target.
5) Verify exact output equality and record acceptance and verifier-round metrics.

V. Implemented Prototype

The metadata and source audit found these completed features [2]:

- Deterministic Python speculative-decoding simulator
- Draft acceptance and first-mismatch verification behavior
- Correctness tests for full, zero, and partial acceptance
- Machine-readable JSONL simulation evidence

`python3 -m unittest discover -s tests -p 'test_*.py'` completed successfully with 5 tests. The inspected test artifacts are `tests/unit/test_orion_simulator.py`. This is evidence of local correctness for encoded cases, not production scale or hardware performance.

TABLE I
IMPLEMENTATION ARTIFACTS AND WHY THEY EXIST

Artifact	Observed responsibility	Engineering rationale
python/orion_simulator.py	SimulationResult, validate_settings, make_target_tokens, wrong_token, make_draft, simulate, git_value, build_record, main	Implements the inspectable, unit-tested project core.
scripts/reproduce.sh	Fixed test and demonstration entry point	Gives another developer one command for local reproduction.
PROJECT.yaml and ANALYSIS.md	Status, completed work, planned work, and claim boundaries	Separates declared intent from evidence-backed implementation.
streamlit_app.py	Local evidence and status visualization	Makes outputs inspectable without upgrading simulation into a hardware claim.

VI. Testing Methodology and Observed Results

Testing uses Python's standard `unittest` discovery and exercises the public behavior of the reference implementation. The audit reran the suite from the project folder with `python3 -m unittest discover -s tests -p 'test_*.py'`. All 5 discovered tests passed. The result establishes correctness only for the encoded local cases; it does not establish accelerator correctness, real-device behavior, production reliability, or benchmark completion.

No numerical performance result is promoted by this test run. Where scripts emit JSON or JSONL, those outputs remain raw or simulation-specific until a reviewed summary includes hardware, software, workload, warm-up, repetition, correctness threshold, Git revision, and limitations.

VII. Claim Boundaries and Risks

The project records these unverified or excluded claims:

- Simulator target-call reduction is not a wall-clock model speedup

The main risk is confusing synthetic or modeled behavior with deployed-system behavior. Other risks include incomplete workloads, platform-dependent timing, missing failure injection, and interfaces not yet exercised across device boundaries. Performance claims require a reviewed record meeting the portfolio standard.

VIII. Evaluation Plan

Evaluation proceeds through correctness tests, deterministic reproduction with Git and environment metadata, repeated benchmarks reporting latency/throughput/memory/error metrics, and a named profiler capture tied to exact hardware and source revision. Success requires reference equivalence within a documented tolerance, preservation of safety and resource invariants, clear failures, and evidence reproducible from a clean environment.

IX. Goals, Milestones, and Success Criteria

The project goals are staged so that correctness precedes performance and integration. A goal is complete only when its proof artifact is committed or otherwise reviewable; prose or a simulated number alone is insufficient.

X. Future Work and Integrations

The five project-specific next steps are:

1) Acceptance probability sweep and chart
2) Toy probabilistic model adapter
3) Real draft and target model experiment
4) Measure quality and latency across real draft/target model pairs.
5) Integrate decoding traces into ARGUS and LYRA.

The early items complete declared evidence; the later items connect downstream portfolio consumers. Each integration should add tests and a reviewable artifact such as JSONL evidence, a report, profiler capture, deployment manifest, trace, or labeled data set.

XI. Conclusion

ORION Speculative Decoding Lab is an evidence-aware active prototype: its implemented behavior and tests are real, while unbuilt hardware, deployment, and performance goals remain labeled. Completing the five integrations in dependency order will advance it from a learning artifact toward a credible portfolio component.

References

[1] Aaron Singh, "ORION Speculative Decoding Lab PROJECT.yaml," local portfolio repository.

[2] Aaron Singh, "ORION Speculative Decoding Lab: README, ANALYSIS, source, and test artifacts," local portfolio repository.

[3] Aaron Singh, "AI Infrastructure Portfolio Benchmark Standard," local portfolio repository.

TABLE II
AUDITED TEST MATRIX

Test artifact and case	Behavior being checked	Observed result
tests/unit/test_orion_simulator.py:test_output_always_matches_target	Output always matches target.	Pass (local)
tests/unit/test_orion_simulator.py:test_full_acceptance_uses_block_rounds	Full acceptance uses block rounds.	Pass (local)
tests/unit/test_orion_simulator.py:test_zero_acceptance_uses_one_round_per_token	Zero acceptance uses one round per token.	Pass (local)
tests/unit/test_orion_simulator.py:test_same_seed_repeats_result	Same seed repeats result.	Pass (local)
tests/unit/test_orion_simulator.py:test_invalid_settings_are_rejected	Invalid settings are rejected.	Pass (local)

TABLE III
DECLARED STATUS VERSUS AUDITED EVIDENCE

Audit field	Finding	Evidence source
Declared status	Active Prototype	PROJECT.yaml
Evidence-backed status	Active local prototype; 5 tests passed	Source plus local unittest run
Accepted measured results	None recorded in measured_results	PROJECT.yaml
Mismatch / claim boundary	Simulator target-call reduction is not a wall-clock model speedup	PROJECT.yaml and ANALYSIS.md
Next proof required	Acceptance probability sweep and chart; Toy probabilistic model adapter	Planned features

TABLE IV
PROJECT GOALS AND REQUIRED PROOF

ID	Goal	Completion evidence
G1	Acceptance probability sweep and chart	Passing tests and a reviewed source artifact
G2	Toy probabilistic model adapter	Machine-readable result with reproduction metadata
G3	Real draft and target model experiment	Reference-equivalence or domain-correctness report
G4	Run real draft and target models	Named profiler, deployment, or integration artifact
G5	Separate quality, acceptance, and wall-clock effects	Dashboard/report link preserving provenance and limitations