

PrismManifest - MASTER SPECIFICATION

PrismManifest AI Architecture Master Executive Summary

Field Value
Project PrismManifest / FinancePackBench
Status Approved Master Executive Specification
Version 2.2.0
Date 2026-08-08 (addendum 2026-08-09)
Author Senior QA / Architect & System Design Au

1. Executive Narrative: Resolving The Tool-Argument Trust Gap

As AI agentic systems increasingly invoke deterministic software tools (tax calculators, loan underwriting models, financial ledgers, database APIs), a fundamental architectural flaw emerges: The Tool-Argument Trust Gap.

The Fundamental Equation of Tool-Argument Security

$$\text{Deterministic Engine} + \text{Unverified Probabilistic Input} = \text{Deterministic Wrong Answer}$$

PrismManifest Zero-Trust Pipeline

Probabilistic Extraction >>> Evidence Binding >>> Independent Verification >>> Parameter Manifest >>> Deterministic DAG

2. Quorum Tolerance Thresholds & Aligned Status Decision Rules

PASS:

- * Disagreement = \$0 (exact candidate consensus).
 - * Dual-OCR confidence >= 0.98, form anchors verified.
- PASS_WITH_HUMAN:
- * Immaterial quorum disagreement (> \$0 but <= \$1,000).
 - * OR dual-OCR confidence < 0.98.
 - * OR missing layout anchor, but span grounding passes.
 - * Action: Generates signed manifest, but routes to human review queue.

REFUSE:

- * Material quorum disagreement (> \$1,000).
- * OR 0% span grounding match in corpus.
- * OR Plausibility Sub-DAG order-of-magnitude failure.
- * Action: Hard rejection. Group 3 execution is blocked.

3. CUDA Parity Policy

CUDA Parity Invariant: The experimental CUDA GPU acceleration path (Phase 3) must produce bit-identical Parameter Manifests and 100% identical gate decisions (PASS, PASS_WITH_HUMAN, REFUSE) as the primary

host-side C++ verification engine across the entire FinancePackBench benchmark suite.

4. Production SLAs on FinancePackBench

Metric Target
False-Accept Rate (Critical Money Fields $\leq 0.01\%$
False-Refusal / Escalation Rate $\leq 5.0\%$
CUDA Parity Rate 100% Gate Decision Match
Audit Replay Reproducibility 100% Bit-Identical on PASS manifests
Gate Verification Latency (Host Path) ≤ 15 ms (p99)

5. Master Specification Reference Links

- * PrismManifest System Design Specification v2.2.0
- * PrismManifest Implementation Plan v2.2.0
- * Handoff / pilot evidence (living; does not supersede Design/Plan)
- * Pilot pack report (generated evidence)

6. Implementation Status (2026-08-09) - Executive

Normative SLAs in §4 are unchanged. Engineering proof to date:

Claim Evidence
Python / C++ / CUDA agree on gate decisi Pytest + CUDA parity PASS; C++ binaries
Host gate latency ? LLM time Python in-process ~0.08 ms p50; C++ CLI
CUDA useful when batched / resident Warm ~1.2 ms; ~718k fields/s @ batch 102
Synthetic critical FA FinancePackBench / G4 / ugly+LLM pilot -
Pilot packaging prismmanifest-gate pilot-pack

Not yet claimed: live customer fax corpus SLA; multi-model 1000-doc LLM critical-FA paper.

Document Control

Item Value
Spec ID PRISMMANIFEST-MAS-v2.2.0
Classification Approved Master Executive Specification
Supersedes PRISMMANIFEST-MAS-v2.1.0
Role Executive narrative + index only; Design