Verifiable Labs Compliance Report — anthropic/claude-opus-4.7

Verifiable Labs Hosted Evaluation Platform v0.1.0-alpha

2026-04-25

Verifiable Labs Compliance Report

Model under evaluation. `anthropic/claude-opus-4.7` **Evaluation date.** 2026-04-25 **Platform version.** v0.1.0-alpha **Source data.** `complete_matrix_single_turn.csv` **Scope.** 5 environments × 15 total episodes

> This report is generated automatically from a benchmark CSV produced by the Verifiable Labs evaluation platform. It is intended as a self-service template — the platform does not assert legal compliance with any specific regulatory framework (NIST AI RMF, EU AI Act, ISO 42001) on the user's behalf. Consult your own counsel for legal attestation; this document evidences *empirical* model behaviour on a verifiable, conformal-calibrated benchmark.

---

1. Executive Summary

`anthropic/claude-opus-4.7` was evaluated on **5** scientific-reasoning environments spanning compressed-sensing, imaging, and physics-inverse. Across **15** total episodes the model achieved a mean reward of **0.481** (parse-fail rate **13.3 %**). Empirical conformal coverage on calibrated envs was **93.3 %**, against a target of **90 %**.

Headline findings.

- **Strongest env:** `mri-knee-reconstruction` (mean reward 0.690, n = 3).
- **Weakest env:** `phase-retrieval` (mean reward 0.133, n = 3).
- **High parse-fail rate (13.3 %)** — formatting brittleness; consider tightening the system prompt.
- **Capability spread of 0.56** across envs — model is specialised; cross-env transfer is poor.
- **Aggregate mean 0.481** — at v0.1 baseline.

Recommended next step. Address parse failures first — formatting issues mask capability. Then re-run this report against the same CSV column.

---

2. Methodology

The evaluation platform implements **conformal-calibrated rewards** on inverse-problem environments. For each (env, seed) the platform:

1. Procedurally regenerates a fresh problem instance from the seed, so the model has not seen this exact problem in training.
2. Sends a structured prompt to the model under test via OpenRouter (single-turn or multi-turn dialogue depending on the env).
3. Parses the JSON response into a typed `Prediction` dataclass.
4. Scores the prediction against ground truth using a per-env reward that includes:
   - a **point-estimate** term (NMSE / SSIM / chamfer / problem-specific)
   - a **support-recovery** term (where applicable)
   - a **conformal-coverage** term — the model is asked to provide uncertainty bounds; we score whether the truth falls inside.

Calibration: per-env conformal quantiles are computed offline on a held-back calibration pool (typically 200-500 instances) at target coverage `1 -  `, with `= 0.10`.

Reproducibility. Every reward in this report is reproducible from the source CSV (`complete_matrix_single_turn.csv`) and the env code at the commit recorded in the benchmark metadata. The benchmark itself is re-runnable end-to-end via `python benchmarks/run_v2_benchmark.py --model anthropic/claude-opus-4.7`.

---

3. Capability Assessment

3.1 Per-environment reward distribution

env	n	mean reward	std	parse-fail	coverage
`mri-knee-reconstruction`	3	0.690	0.155	0.0 %	92.9 %
`super-resolution-div2k-x4`	3	0.649	0.149	0.0 %	92.4 %
`lodopab-ct-simplified`	3	0.639	0.048	0.0 %	97.2 %
`sparse-fourier-recovery`	3	0.293	0.021	0.0 %	93.3 %
`phase-retrieval`	3	0.133	0.188	66.7 %	85.0 %

3.2 Aggregate distribution

- **Mean reward:** 0.481
- **Median reward:** 0.539
- **Standard deviation:** 0.260
- **Min / Max:** 0.000 / 0.902
- **Episodes scored:** 15

A higher reward indicates closer alignment with the ground-truth solution under the env's specific reward function. By construction, rewards are bounded in `[0, 1]`; values above `0.7` correspond to solutions that are within engineering tolerances of the analytic optimum, while values below `0.3` typically indicate a structural misunderstanding of the problem (wrong support, wrong forward operator, wrong prior scale).

---

4. Failure Modes

4.1 Parse failures

2 of 15 episodes (**13.3 %**) failed to produce a parseable JSON prediction. Parse failures count as `reward = 0` in the aggregate.

Common causes the platform records:

- markdown code-fence wrapping (model emits `json …` despite schema rejection of fences)
- prose preamble or postamble
- incorrect array length on the support indices or amplitudes
- out-of-range support indices or duplicate entries

A parse-fail rate above 5 % suggests the model needs prompt engineering — not necessarily a capability gap; it may be a formatting brittleness.

4.2 Low-reward environments

Envs with mean reward below **0.30** (suggesting structural failure):

env	mean reward	parse-fail
`phase-retrieval`	0.133	66.7 %
`sparse-fourier-recovery`	0.293	0.0 %

These environments warrant manual inspection of representative episodes (the per-seed CSV preserves the full prompt + response).

---

5. Calibration

The platform reports **empirical conformal coverage** — the fraction of episodes where the model's stated uncertainty interval contained the ground truth — for envs that score uncertainty (all envs in the v0.1 set).

metric	observed	target	notes
coverage	93.3 %	90 %	conformal split-quantile, = 0.10
over-coverage	1 envs	—	model claims more uncertainty than needed
under-coverage	0 envs	—	model is over-confident

Reading. Within ±5 percentage points of target is "well-calibrated" for v0.1; the alpha gate flags >10 pp deviation either way. Severe under-coverage (model claims certainty it doesn't have) is the most operationally dangerous failure mode and should block deployment in safety-relevant settings.

---

6. Recommendations

1. **Tighten output formatting.** Parse-fail rate exceeds 5 %; rewrite the system prompt to forbid markdown fences and emit a strict JSON schema reminder. The Tier-1 SDK exposes `client.env(...).adapter` to inspect the canonical prompt.
2. **Inspect failures on `phase-retrieval`, `sparse-fourier-recovery`.** Mean reward below 0.30 typically indicates a structural misunderstanding of the problem, not a prompt-engineering issue. Pull representative seeds from the per-episode CSV and review the model's reasoning.

---

7. Appendix

7.1 Environment list

The following 5 environments contributed to this report:

- `lodopab-ct-simplified` (3 episodes)
- `mri-knee-reconstruction` (3 episodes)
- `phase-retrieval` (3 episodes)
- `sparse-fourier-recovery` (3 episodes)
- `super-resolution-div2k-x4` (3 episodes)

Each environment is documented under `docs/environments/<env_id>.md` in the verifiable-labs-envs repository.

7.2 Source data

- Per-episode CSV: `complete_matrix_single_turn.csv`
- Benchmark commit: see CSV metadata column
- Platform version: v0.1.0-alpha
- Generated: 2026-04-25T17:55:51+00:00

7.3 Limitations of this report

- v0.1 envs are inverse-problem-shaped; this report does not cover conversational, agentic, or open-ended generation tasks.
- Sample size per env is **3** seeds; statistical precision on a single env is bounded by `±0.1 / √n`.
- The platform asserts no claim about legal compliance with AI-governance frameworks; this is an empirical capability report, not a regulatory attestation.