Verifiable Labs

Capability audit report

Model under audit

Frontier Model A

2026-05-01

Generated by `vlabs-audit`

`aud_105178a1631e42f295f8f006586c8fee`

Contents

1. Executive Summary

This report presents an automated capability audit of Frontier Model A on 3 Verifiable Labs environments, totalling 90 episodes. The audit was run with conformal miscoverage $\alpha = 0.10$ (target empirical coverage 90%).

1.1. Headline finding

Across the 3 environments audited, the model achieved a mean reward of **0.420** with a 95% bootstrap CI of $[0.397,\ 0.446]$ and a parse-failure rate of 14.4%. Empirical (held-out) coverage averaged 92.0% across all environments versus a target of 90%.

1.2. Coverage compliance

Environment	Empirical coverage	Target	Status
Phase Retrieval	0.877	0.900	**below target**
Sparse Fourier Recovery	0.953	0.900	**within target**
Image Super-Resolution (DIV2K, ×4)	0.906	0.900	**within target**

1.3. Recommendation summary

5 actionable recommendations are listed in Section 5; the most consequential are summarised in the per-environment results that follow.

2. Methodology

2.1. Conformal calibration in one paragraph

Each Verifiable Labs environment exposes a scoring function with an attached *conformal* component: for every episode we compute a non-conformity score and turn the empirical quantile of those scores at level $1 - \alpha$ into a prediction interval that, under exchangeability, contains the held-out reference reward with probability at least $1 - \alpha$ (see [1]). This audit reports the *empirical* coverage achieved over a held-out fold of every audit run; deviations from $1 - \alpha = 0.90$ flag either undercoverage (miscalibration) or excessive interval width (over-conservative). The conformal layer is provided by the open-source `vlabs-calibrate v0.1.0a1` package (pinned for reproducibility).

2.2. Audit configuration

Setting	Value
Conformal α	0.10
Target coverage	0.90
Episodes per environment	30
Initial seed	0
Held-out fold	seed-sorted second half (50/50 split)

The held-out fold is the half of episodes with the larger seeds within each environment; coverage is averaged over those held-out traces. Mean reward is reported with a 95% percentile-bootstrap confidence interval ($n_{\text{resamples}} = 1000$, deterministic RNG seed for reproducibility).

2.3. Environments under audit

- **Phase Retrieval** — Recover the phase of a complex signal from magnitude-only measurements (a classic ill-posed inverse problem in optics and imaging). The environment scores spectral overlap with the ground-truth signal under the standard global-phase ambiguity.
- **Sparse Fourier Recovery** — Recover a k-sparse complex signal from a small number of compressive Fourier measurements. The model returns the support indices and complex amplitudes; the environment scores normalised mean-squared error against the ground-truth signal and reports a conformal interval over the reconstruction quality.
- **Image Super-Resolution (DIV2K, ×4)** — 4× single-image super-resolution on the DIV2K validation set. The environment compares structural similarity between the reconstructed and ground-truth high-resolution images.

2.4. What this audit does *not* measure

This is an offline, single-turn capability audit: the model receives a prompt, returns a structured prediction, and the environment scores it. The audit therefore says nothing about multi-turn tool use, planning, training-data leakage, or robustness to adversarial inputs. It also does not establish causal claims about how to *improve* the model — only how it currently performs on this fixed evaluation regime.

References

[1] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. *Distribution-free predictive inference for regression*. Journal of the American Statistical Association, 113(523), 1094–1111, 2018.

3. Results

This section reports per-environment results followed by three aggregate visualisations spanning every environment.

3.1. Phase Retrieval

Recover the phase of a complex signal from magnitude-only measurements (a classic ill-posed inverse problem in optics and imaging). The environment scores spectral overlap with the ground-truth signal under the standard global-phase ambiguity.

Metric	Value
Episodes scheduled	30
Successful episodes	21
Failed episodes	9
Mean reward	0.335
95% bootstrap CI	[0.319, 0.350]
Parse-failure rate	30.0%
Format-validity rate	100.0%
Held-out empirical coverage	0.877

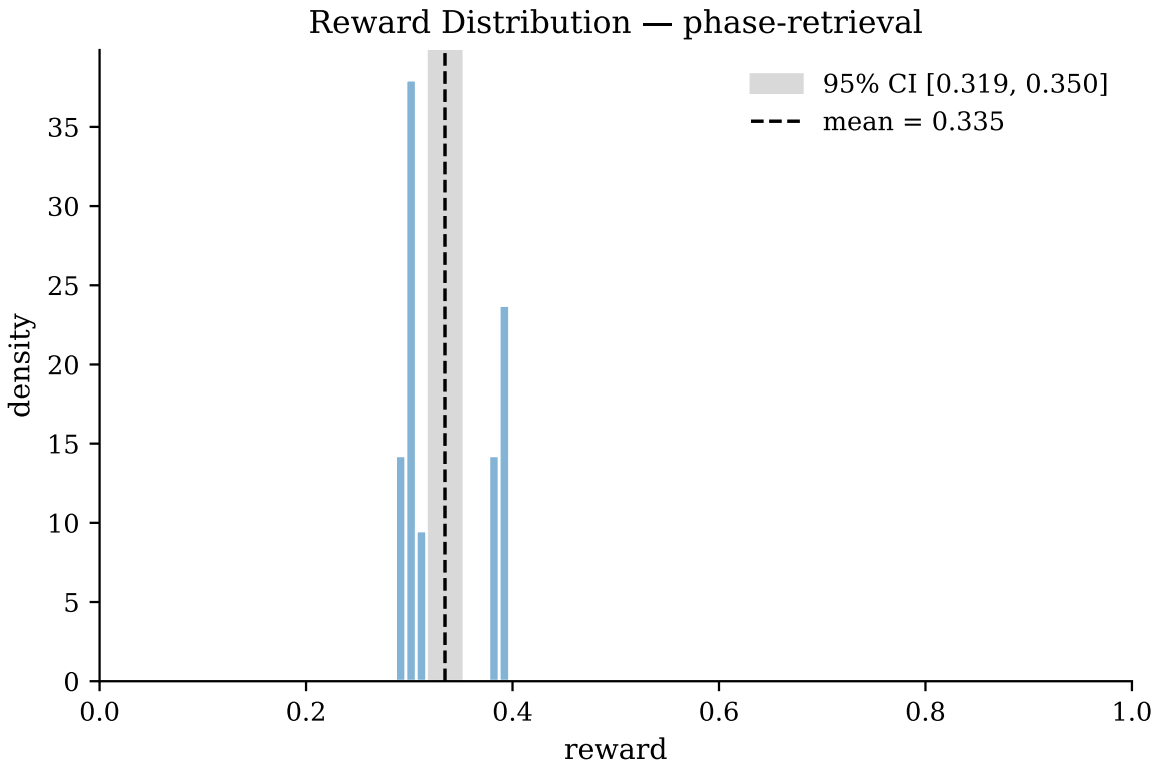

Figure 1: Per-episode reward distribution on `phase-retrieval`. The shaded band marks the 95% bootstrap confidence interval on the mean; the dashed vertical line is the sample mean.

3.2. Sparse Fourier Recovery

Recover a k-sparse complex signal from a small number of compressive Fourier measurements. The model returns the support indices and complex amplitudes; the environment scores normalised mean-squared error against the ground-truth signal and reports a conformal interval over the reconstruction quality.

Metric	Value
Episodes scheduled	30
Successful episodes	30
Failed episodes	0
Mean reward	0.348
95% bootstrap CI	[0.339, 0.357]
Parse-failure rate	0.0%
Format-validity rate	100.0%
Held-out empirical coverage	0.953

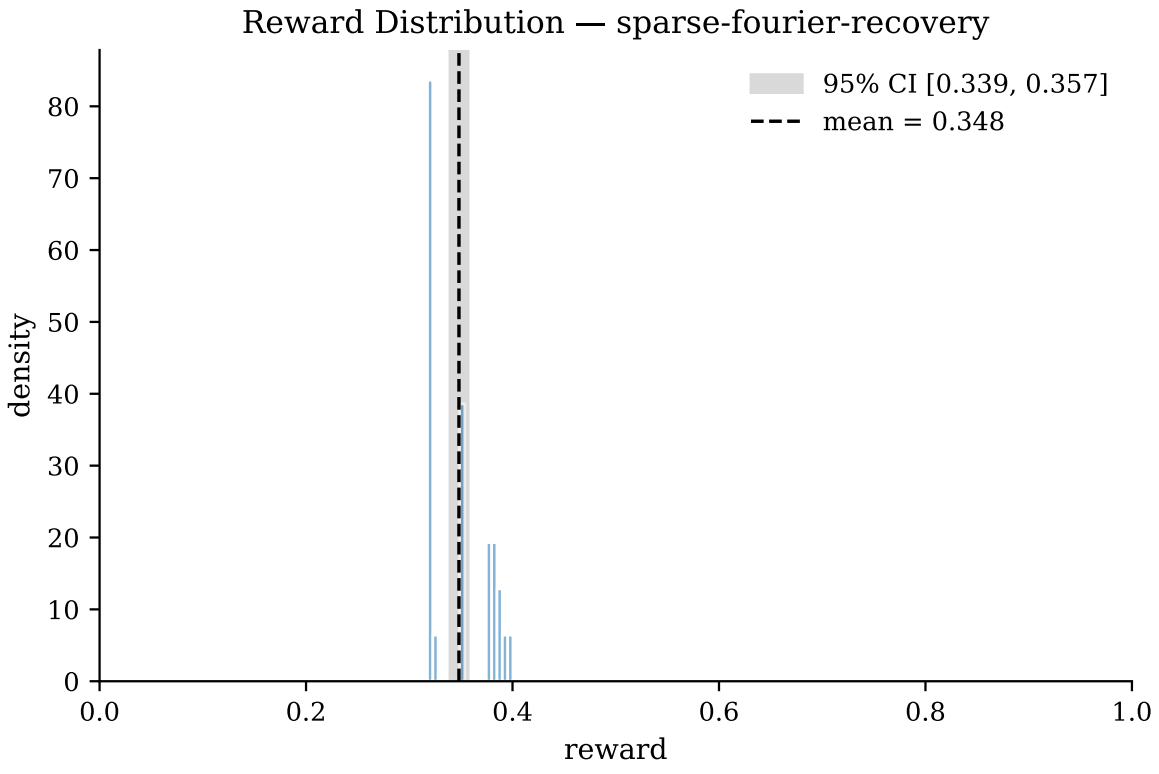

Figure 2: Per-episode reward distribution on `sparse-fourier-recovery`. The shaded band marks the 95% bootstrap confidence interval on the mean; the dashed vertical line is the sample mean.

3.3. Image Super-Resolution (DIV2K, ×4)

4× single-image super-resolution on the DIV2K validation set. The environment compares structural similarity between the reconstructed and ground-truth high-resolution images.

Metric	Value
Episodes scheduled	30
Successful episodes	26
Failed episodes	4
Mean reward	0.572
95% bootstrap CI	[0.537, 0.610]
Parse-failure rate	13.3%
Format-validity rate	100.0%
Held-out empirical coverage	0.906

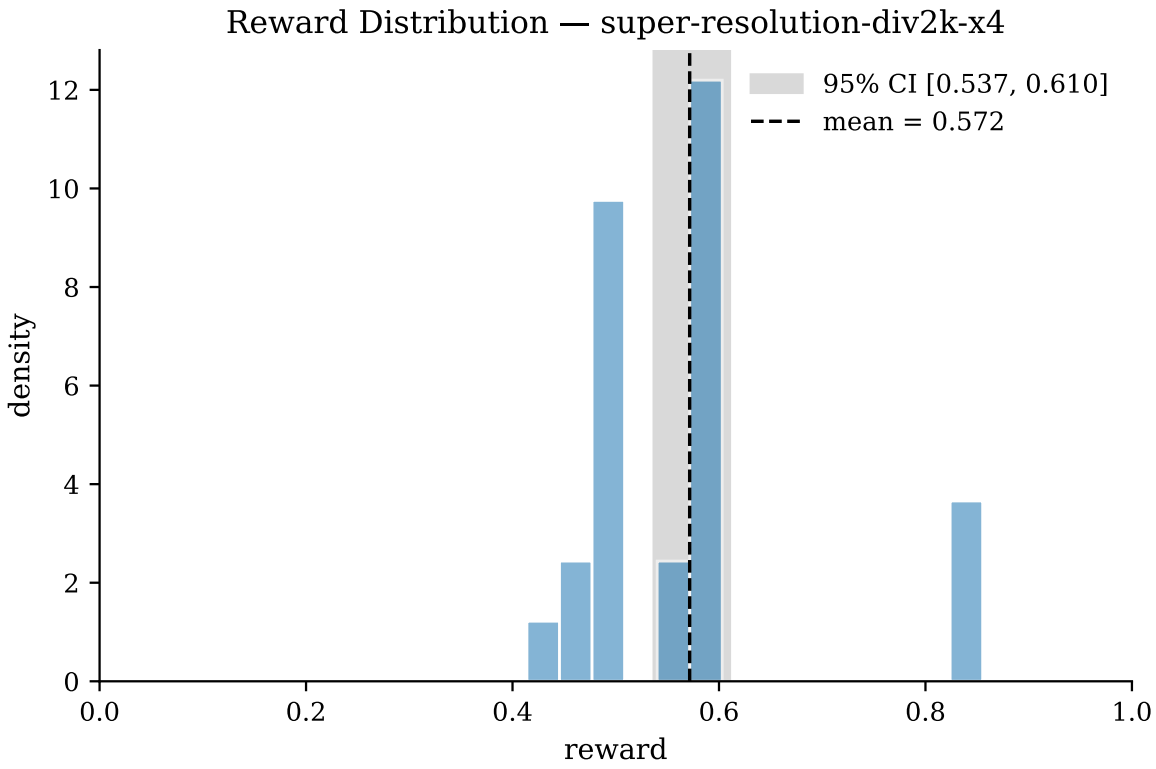

Figure 3: Per-episode reward distribution on `super-resolution-div2k-x4`. The shaded band marks the 95% bootstrap confidence interval on the mean; the dashed vertical line is the sample mean.

3.4. Coverage calibration across environments

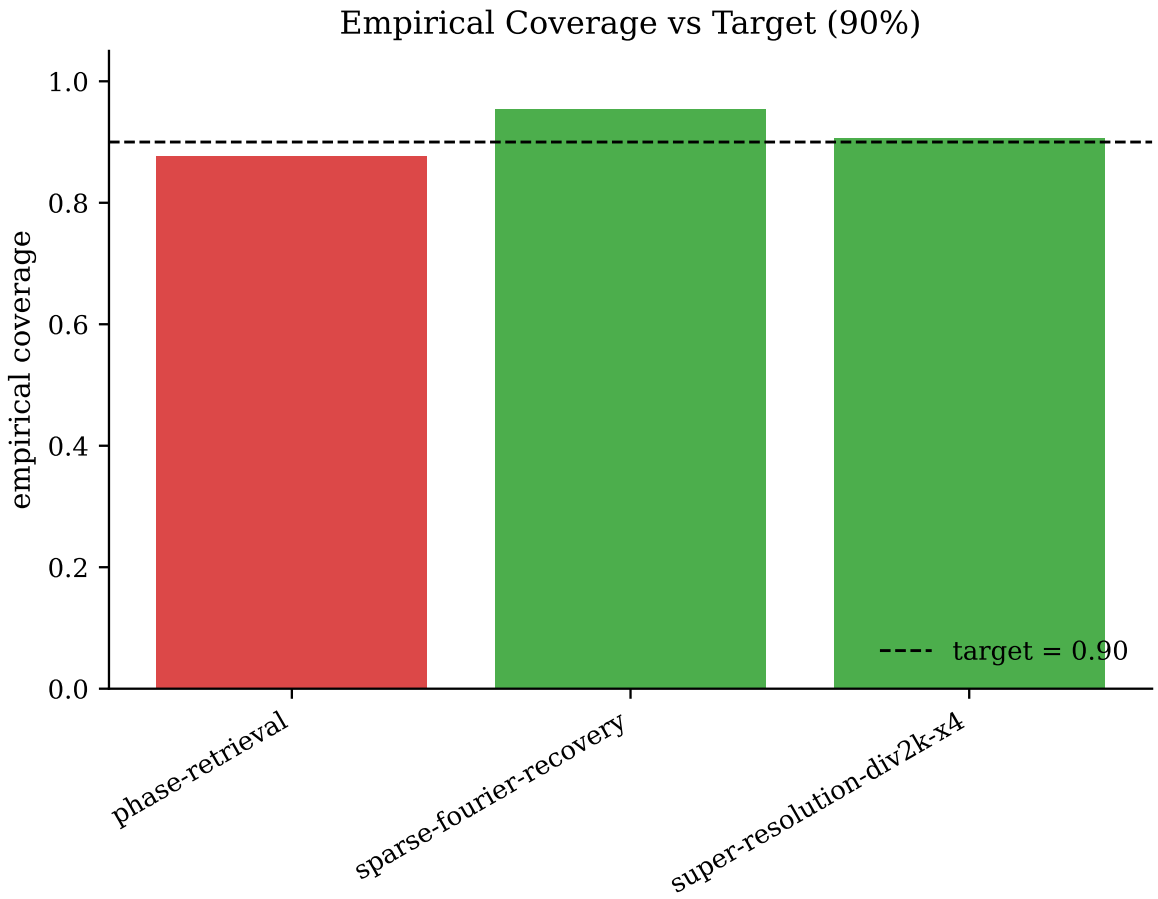

Figure 4: Empirical (held-out) coverage versus target $1 - \alpha = 0.90$. Bars at or above the dashed target line are calibrated; bars below indicate under-coverage.

3.5. Quality breakdown

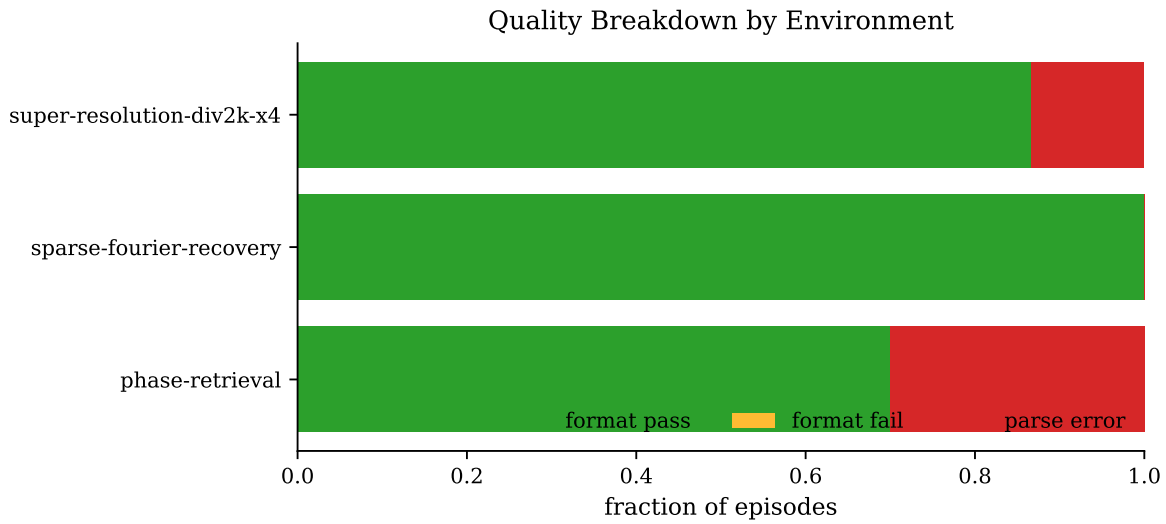

Figure 5: Stacked-bar decomposition of every scheduled episode by outcome class (format pass / format fail / parse error). Segments sum to one per environment.

3.6. Cost analysis

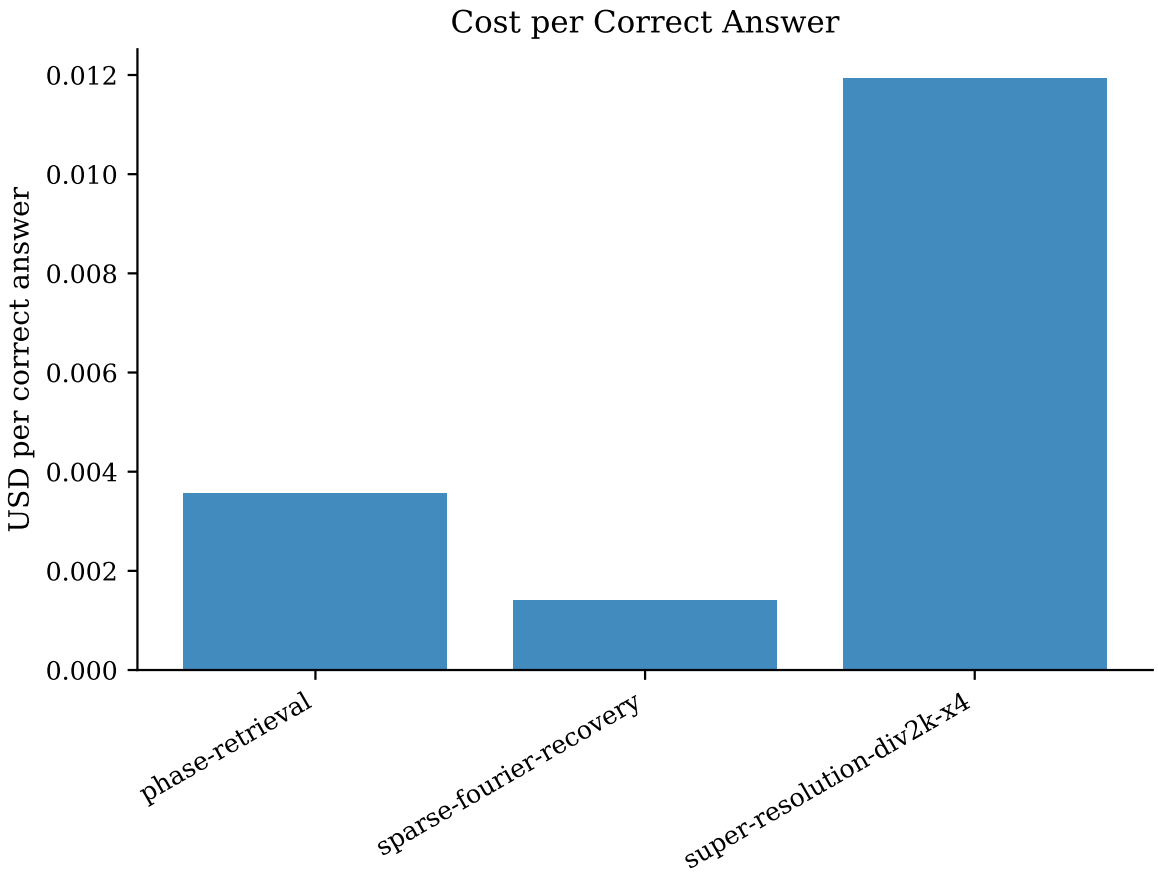

Figure 6: Estimated USD per successful episode, derived from per-trace `estimated_cost_usd` when reported by the SDK. When the underlying traces do not carry cost data the figure renders an informational placeholder rather than fabricated numbers.

4. Recommendations

The list below is generated mechanically from the per-environment statistics on the preceding pages. Recommendations are intentionally framed in conservative language: they describe what the numbers *warrant*, not what is hypothetically possible.

- Mean reward on `phase-retrieval` is 0.335, at or below the 0.4 threshold typical for non-trivial models on this task class. A larger or chain-of-thought-prompted model may be required before deployment in this environment.
- Empirical coverage on `phase-retrieval` (0.877) falls below the target (0.90). The most likely cause is calibration sample-size insufficiency; rerun with `--episodes` doubled before drawing calibration conclusions.
- Parse-failure rate on `phase-retrieval` (30.0%) exceeds 5%. Tightening the format-compliance instructions in the prompt will recover most of these episodes without changing the underlying model.
- Mean reward on `sparse-fourier-recovery` is 0.348, at or below the 0.4 threshold typical for non-trivial models on this task class. A larger or chain-of-thought-prompted model may be required before deployment in this environment.
- Parse-failure rate on `super-resolution-div2k-x4` (13.3%) exceeds 5%. Tightening the format-compliance instructions in the prompt will recover most of these episodes without changing the underlying model.

4.1. What this report cannot tell you

A single audit at fixed α and fixed sample size cannot establish:

- that fixing the issues above will improve downstream task performance (only a follow-up audit can verify);
- that the model is or is not safe to deploy in any specific production setting (deployment risk is a function of cost-of-error, which the audit does not see);
- that performance generalises to environments outside this audit (each Verifiable Labs environment has a narrow scientific domain).

Use the per-environment numbers as a calibrated starting point for follow-up engineering decisions, not as a final verdict.

5. Appendix

5.1. Reproducibility metadata

Field	Value
Audit id	`aud_105178a1631e42f295f8f006586c8fee`
Model	`Frontier Model A`
`vlabs-audit`	v0.0.1
`vlabs-calibrate`	v0.1.0a1
Python	3.12.3 (CPython)
Generated	2026-05-01T16:13:01+00:00
α	0.10
Episodes / env	30
Initial seed	0

5.2. Raw stats


```
{
  "audit_id": "aud_105178a1631e42f295f8f006586c8fee",
  "model": "Frontier Model A",
  "alpha": 0.1,
  "n_episodes_per_env": 30,
  "per_env": [
    {
      "env": "phase-retrieval",
      "n_episodes": 30,
      "n_success": 21,
      "n_failed": 9,
      "mean_reward": 0.33459474605477607,
      "ci_low": 0.3193077351888335,
      "ci_high": 0.35021948475161735,
      "parse_failure_rate": 0.3,
      "format_valid_rate": 1.0,
      "coverage_holdout": 0.8772727272727273,
      "rewards": [
        0.3806512793694042,
        0.3047496230430136,
        0.3828363137776952,
        0.3076021644859369,
        0.30913867694021413,
        0.3899334502522049,
        0.3047985551728337,
        0.3053015892397804,
        0.30385682292160054,
        0.29707738162292563,
```

```
      0.388403542061285,
      0.28877402783814715,
      0.39487994979488805,
      0.3851429970283733,
      0.29990255706476854,
      0.3971105838016676,
      0.2867086031407219,
      0.3067609145229889,
      0.39502779600883653,
      0.30364920816458335,
      0.29418363089842836
    ],
    "total_cost_usd": 0.0749
  },
  {
    "env": "sparse-fourier-recovery",
    "n_episodes": 30,
    "n_success": 30,
    "n_failed": 0,
    "mean_reward": 0.3481183271096276,
    "ci_low": 0.33943833439838555,
    "ci_high": 0.35695659477408204,
    "parse_failure_rate": 0.0,
    "format_valid_rate": 1.0,
    "coverage_holdout": 0.9533333333333334,
    "rewards": [
      0.3207729210553295,
      0.3202931940614722,
      0.3214672685301425,
      0.3880485339366536,
      0.37988801427630775,
      0.3175837115962359,
      0.37616983764698914,
      0.3521150885419581,
      0.40047077245654084,
      0.32194917096267106,
      0.38121275177955616,
      0.3523046606450457,
      0.3209505037111972,
      0.3187589136210489,
      0.3216142050247319,
      0.32045977996305586,
      0.3512466562507496,
      0.35180254535411964,
      0.32130646003052016,
      0.3219255922808313,
      0.3888744813649378,
      0.32247445759997756,
```

```
      0.35211776637427783,
      0.381002070405688,
      0.3760094644759443,
      0.3187302144702293,
      0.3232060912806499,
      0.3765959830054493,
      0.39272135994506585,
      0.35147734264145264
    ],
    "total_cost_usd": 0.041999999999999975
  },
  {
    "env": "super-resolution-div2k-x4",
    "n_episodes": 30,
    "n_success": 26,
    "n_failed": 4,
    "mean_reward": 0.5716060802452707,
    "ci_low": 0.5369725470856597,
    "ci_high": 0.610342891240227,
    "parse_failure_rate": 0.13333333333333333,
    "format_valid_rate": 1.0,
    "coverage_holdout": 0.9060321514423075,
    "rewards": [
      0.6015453453131042,
      0.4865524893418537,
      0.5558327469715052,
      0.8554153535723441,
      0.4881583308740081,
      0.5929752692976842,
      0.6002560980778936,
      0.46461443224614596,
      0.4139579595706607,
      0.8517069566911666,
      0.48639579172300174,
      0.5981863325781505,
      0.6001759809280167,
      0.4738649337197961,
      0.5412993713312046,
      0.4846474816950398,
      0.6019749633875007,
      0.47851464968407553,
      0.6027800767258931,
      0.4852420350022655,
      0.599471843075188,
      0.48726335884535965,
      0.5902225263518395,
      0.8422870312260897,
      0.4856687373374649,
```

```
        0.592747990809787
      ],
      "total_cost_usd": 0.3100000000000001
    }
  ],
  "aggregate_mean_reward": 0.4198934748937165,
  "aggregate_ci_low": 0.3965709297267318,
  "aggregate_ci_high": 0.44588850089550536,
  "aggregate_parse_failure_rate": 0.14444444444444443,
  "aggregate_format_valid_rate": 1.0,
  "aggregate_coverage_holdout": 0.9197246844951925
}
```


5.3. How to cite this audit

```
@misc{verifiable-labs-audit-105178a1,
  title  = {Capability audit of Frontier Model A},
  author = {Verifiable Labs},
  year   = {2026},
  note   = {Generated by vlabs-audit v0.0.1;
            audit id aud\_105178a1631e42f295f8f006586c8fee.},
}
```

5.4. License

The Verifiable Labs SDK and `vlabs-audit` are released under the Apache 2.0 license. The numeric results in this report are machine-derived from publicly described environments and may be reproduced under any license consistent with reasonable scientific disclosure.